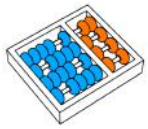
Universidade Estadual de Campinas
Instituto de Computação



Caroline Mazini Rodrigues

Análise de Representatividade de Imagens para Descrição de Eventos

CAMPINAS
2020

Caroline Mazini Rodrigues

Análise de Representatividade de Imagens para Descrição de Eventos

Dissertação apresentada ao Instituto de Computação da Universidade Estadual de Campinas como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Orientador: Prof. Dr. Zanoni Dias
Coorientador: Prof. Dr. Anderson de Rezende Rocha

Este exemplar corresponde à versão final da Dissertação defendida por Caroline Mazini Rodrigues e orientada pelo Prof. Dr. Zanoni Dias.

CAMPINAS
2020

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

Rodrigues, Caroline Mazini, 1995-
R618a Análise de representatividade de imagens para descrição de eventos / Caroline Mazini Rodrigues. – Campinas, SP : [s.n.], 2020.

Orientador: Zanoni Dias.
Coorientador: Anderson de Rezende Rocha.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Computação.

1. Representação de imagens. 2. Computação forense. 3. Aprendizado manifold. 4. Descrição de eventos (Computação). I. Dias, Zanoni, 1975-. II. Rocha, Anderson de Rezende, 1980-. III. Universidade Estadual de Campinas. Instituto de Computação. IV. Título.

Informações para Biblioteca Digital

Título em outro idioma: Image representativeness analysis for event description
Palavras-chave em inglês:
Image representation
Forensic computing
Manifold learning
Events description (Computer science)
Área de concentração: Ciência da Computação
Titulação: Mestra em Ciência da Computação
Banca examinadora:
Zanoni Dias [Orientador]
Moacir Antonelli Ponti
Hélio Pedrini
Data de defesa: 04-06-2020
Programa de Pós-Graduação: Ciência da Computação

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0001-7838-3038
- Currículo Lattes do autor: http://lattes.cnpq.br/1282737238188635

Universidade Estadual de Campinas
Instituto de Computação

Caroline Mazini Rodrigues

Análise de Representatividade de Imagens para Descrição de Eventos

Banca Examinadora:

- Prof. Dr. Zanoni Dias
  Instituto de Computação – Unicamp
- Prof. Dr. Moacir Antonelli Ponti
  Instituto de Ciências Matemáticas e de Computação – USP
- Prof. Dr. Hélio Pedrini
  Instituto de Computação – Unicamp

A ata da defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria do Programa da Unidade.

Campinas, 04 de junho de 2020

Dedicatória

Ao meu pai, que me acompanha por todo o caminho, fisicamente ou não.

Rien n'est plus puissant qu'une idée dont le temps est venu.
Nada é mais poderoso que uma ideia cuja hora chegou.
(Victor Hugo)

Agradecimentos

Tenho muito a agradecer a todos que um dia contribuíram, direta ou indiretamente, para que eu pudesse chegar até aqui. Algumas pessoas e instituições em especial.

Agradeço ao apoio financeiro do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pela bolsa de pesquisa concedida (processo nº 130849/2018-9), entre março e novembro de 2018. Agradeço também à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), pela bolsa de pesquisa concedida por meio do processo nº 2018/16214-3, Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) (vinculado ao processo nº 2017/12646-3), entre dezembro de 2018 e fevereiro de 2020.

Agradeço a todos os professores que me incentivaram desde o início. Àqueles que são minha inspiração para continuar na pesquisa e almejar seguir seus passos no ensino. Àqueles que mesmo com tanto trabalho e pouco reconhecimento mostraram sua paixão pela profissão. Agradeço principalmente aos meus orientadores Professor Zanoni Dias e Professor Anderson Rocha pela acolhida, confiança e parceria que demonstraram, mesmo nos momentos mais difíceis da pesquisa. A eles que me mostraram, cada um a sua maneira, a importância do profissionalismo e responsabilidade para ser um transmissor de conhecimento. Aos melhores orientadores que eu poderia ter escolhido. Agradeço ainda ao Professor Luis Pereira, que desempenhou diversos papeis durante esses dois anos. A ele que foi meu colega de laboratório, meu orientador não oficial e meu amigo. Por me mostrar que discordar e questionar é parte do aprendizado e por compartilhar comigo um pouco do amor pela escrita.

Agradeço ao meu laboratório RECOD, minha mais nova família, pelo apoio e companherismo que me ajudaram a atravessar esses dois anos. Pelas risadas, conversas altas e pela amizade que cresceu além do trabalho. Por me ensinarem que a vida é mais do que ser a melhor, é estar feliz ao lado de pessoas especiais.

Agradeço a todos os amigos que fizeram parte da minha trajetória inicial e acreditaram em mim mesmo quando eu não acreditava, em especial Ana Luísa e Nicolas. Porque mesmo longe, eles sempre farão parte da minha história. Agradeço também aos novos amigos que encontrei pelo caminho, em especial minha amiga Sarah, que em pouco tempo se tornou uma amiga para a vida, me apoiando durante todo o mestrado e todas as crises que surgiram no processo.

Agradeço à minha irmã Karine, que mesmo sendo mais nova, por vezes assumiu o papel de mais velha, para me aconselhar ou incentivar. À minha irmãzinha, que foi capaz de me suportar com o carinho de pequenos gestos. Agradeço especialmente aos meus pais que nunca me pediram nada em troca, apesar de terem me dado tudo. À minha mãe Maria, que me mostra a cada dia como ser forte, mesmo quando nada parece ter saída. Ao meu pai Antonio, para quem não pude mostrar o meu trabalho finalizado, mas que sempre se orgulhou de mim mesmo com ele incompleto. Aos meus pais, à quem devo o que sou.

Agradeço a Deus, por todas essas grandes pessoas que colocou em meu caminho.

Resumo

Uma diversidade de eventos, como atos terroristas e catástrofes naturais, ocorrem frequentemente em todo o mundo. A disponibilidade de imagens na internet pode ajudar a entender esses eventos. Ao lidar com imagens de eventos, filtrar as imagens é um dos principais desafios. Os dados cruciais, que realmente poderiam representar o evento, podem estar misturados com quantidades ainda maiores de dados sem importância. No entanto, a seleção manual de imagens representativas (úteis) em uma grande quantidade de dados pode ser inviável. Assim, uma pergunta surge: *como separar automaticamente as imagens representativas das não-representativas?* Nesta pesquisa, propomos técnicas para lidar com essa questão, considerando a falta de imagens rotuladas para indicar representatividade. Lidamos com a recuperação de imagens por representatividade usando métodos de recuperação de imagem baseados em conteúdo (*Content-Based Image Retrieval* – CBIR), posteriormente aprimorados por métricas de avaliação de qualidade de ranking. No entanto, um dos maiores problemas ao recuperar imagens é representá-las de maneira correta semanticamente. Para propor representações capazes de capturar a semântica de eventos, apresentamos duas abordagens. Nossas abordagens são baseadas em representações de componentes que podem codificar as informações necessárias para descrever os eventos, como pessoas que fizeram parte do evento (por exemplo, suspeitos ou vítimas); objetos que aparecem na cena (por exemplo, carros ou armas); e o local onde o evento se desenrolou (por exemplo, parques, estádios ou prédios). A primeira abordagem proposta, chamada Espaço Semântico de Evento, pretende descrever imagens como uma representação de baixa dimensionalidade usando uma pequena quantidade de imagens representativas conhecidas. A segunda abordagem, chamada Espaço Combinado de Evento, visa melhorar os resultados de precisão da primeira abordagem aprendendo uma maneira de combinar os componentes representativos. Os resultados em três conjuntos de dados de eventos do mundo real atestam a capacidade de nossos métodos para representar eventos com base na combinação de componentes representativos.

Abstract

Different events, such as terrorist acts and natural catastrophes, frequently occur across the world. The availability of images on the internet can help to understand events. When dealing with images from events, filter the images is one of the major challenges. The crucial data, which could indeed represent the event, might be mixed with even more massive amounts of non-important data. However, manually selecting representative (helpful) images from a massive amount of data can be infeasible. Hence, the question becomes: *How to automatically separate the representative images from the non-representative ones?* In this research, we propose techniques to deal with this question considering the lack of labeled images to indicate representativeness. We cope with the representativeness image retrieval using methods of Content-Based Image Retrieval – CBIR, posteriorly improved by quality ranking metrics. Nevertheless, one of the biggest problems when retrieving images is to correctly represent these images semantically. In order to propose representations which could capture the event semantics we present two approaches. Our approaches are based on representations of components which could encode the information necessary to describe the events, such as people attending it (e.g. suspects or victims); objects that appear in the event scene (e.g., cars,gun, backpack); and the place where the event unfolded (e.g.,park, stadium, building). The first approach proposed, called Event Semantic Space, intend to describe images as a low-dimensional representation using a small quantity of known representative images. The second approach, called Event Combined Space, aims to overcome the precision results of the first one by learning a manner to combine the representative components. Results on three real-world event datasets attest the capability of our methods to represent events based on representative components combination.

Lista de Figuras

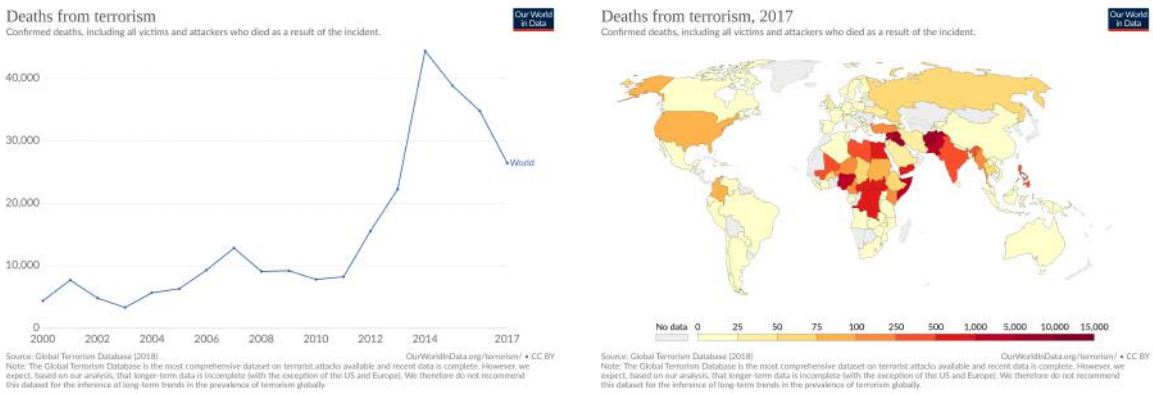
Lista de Tabelas

Sumário

Capítulo 1

Introdução

Atos terroristas, catástrofes naturais, e outros tipos de eventos frequentemente acometem sociedades ao redor do mundo e entendê-los tem se tornado um desafio global [34]. Para que tenhamos uma ideia, entre os anos de 2000 e 2017, as estatísticas mostram mais de 271.460 mortes por atos terroristas ocorridos em todo o mundo e a concentração dessas ocorrências é representada pela Figura 1.1. Catástrofes naturais como inundações, incêndios e terremotos que ocorreram pelo mundo nesse período causaram mais de 1.208.360 mortes, com estatísticas e áreas de concentração dessas catástrofes mostradas na Figura 1.2.

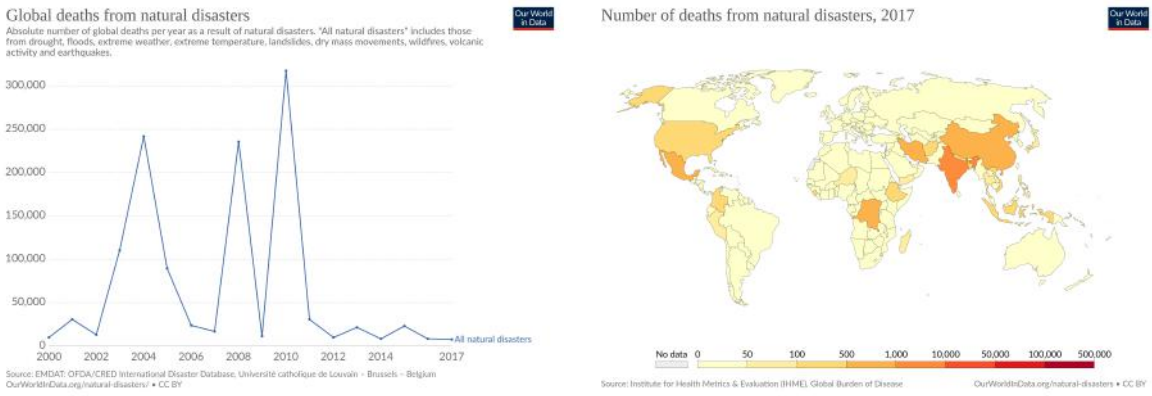


(a) Número de mortes

(b) Áreas de concentração das mortes

Figura 1.1: Número de mortes e áreas de concentração de atos terroristas ocorridos ao redor do mundo entre 2000 e 2017 [16].

Entender esses eventos é crucial, seja para prevenção de incidentes ou para identificação de culpados. Com o aumento do uso de dispositivos móveis, dados tornaram-se disponíveis rapidamente na internet e em grandes quantidades [40]. Esses dados, compartilhados em plataformas como *Twitter*, *Facebook* e *Instagram* [41], e acessíveis a partir de dispositivos móveis, proporcionam oportunidades para obter informações do evento.



(a) Número de mortes (b) Áreas de concentração das mortes

Figura 1.2: Número de mortes e áreas de concentração de catástrofes naturais ocorridas ao redor do mundo entre 2000 e 2017 [48].

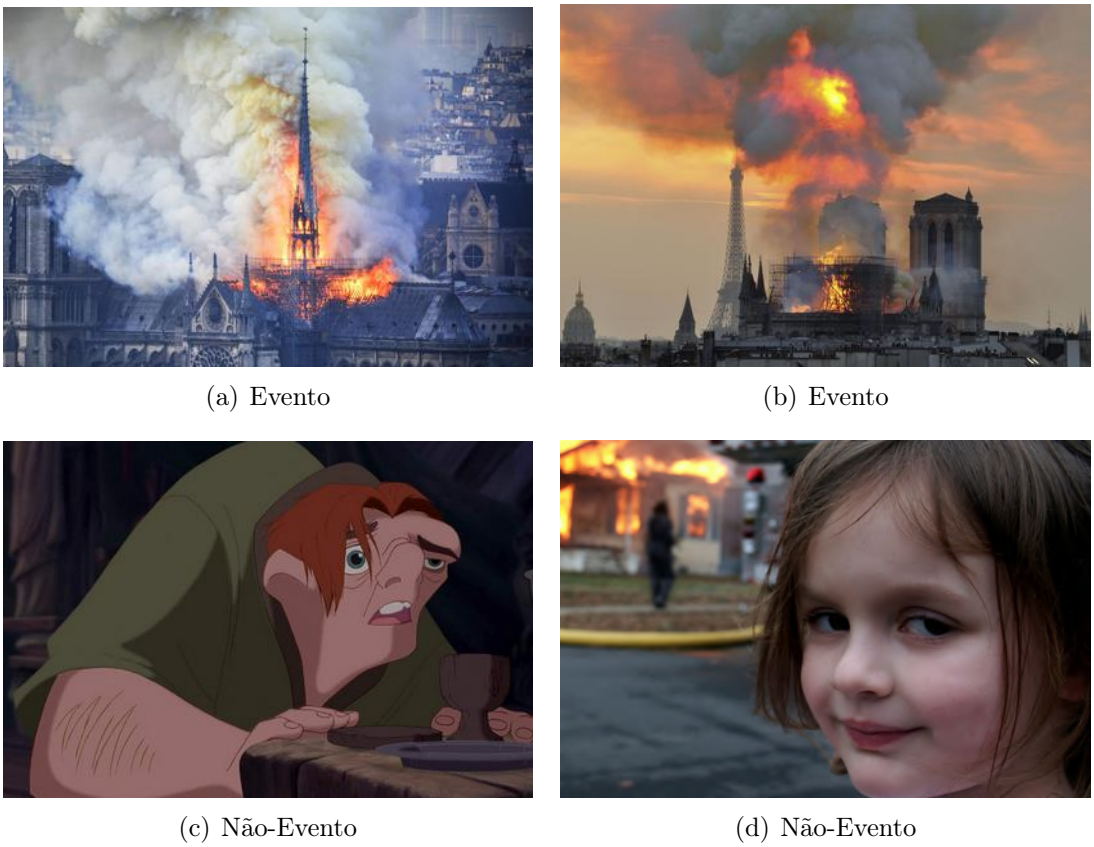
Total de *tweets*	7.888.374
Total de usuários	3.677.531
Retweets	4.464.201
Respostas	260.627
Hora das explosões	Seg. 15 Abril 18:50:00 2013
Hora do primeiro *tweet*	Seg. 15 Abril 18:53:47 2013
Hora da primeira imagem de explosão	Seg. 15 Abril 18:54:06 2013

Tabela 1.1: Estatísticas de número de *tweets* relacionados ao evento Maratona de Boston de 2013. Dados apresentados por *Gupta et al.* [14].

1.1 Descrição de Eventos Utilizando Imagens

Considere um evento — como um ataque terrorista — que ocorreu em um lugar com muitas pessoas com celulares, como estádios ou teatros. Em minutos, centenas ou até milhares de mensagens de texto, imagens e vídeos podem ser compartilhados nas mídias sociais. Apresentamos um exemplo dessa quantidade de dados na Tabela 1.1. Foram coletados 7.888.374 *tweets* relacionados ao evento Maratona de Boston de 2013, sendo o primeiro deles obtido menos de quatro minutos depois da primeira explosão.

Se corretamente analisados, esses dados podem auxiliar no entendimento de eventos. Esse entendimento envolve a capacidade de descrever uma sequência de acontecimentos que compõem o evento para que posteriormente seja possível inferir informações. Alguns estudos relacionados com a disseminação de conteúdo textual já foram realizados, como, por exemplo, a análise de *tweets* durante situações emergenciais [18, 24] como o estudo realizado por *Starbird at al.* [64] com a inundação do *Red River Valley* nos Estados Unidos e Canadá, em março e abril de 2009. Entender o comportamento desses *tweets* poderia auxiliar na resposta governamental em situações de crise. Nesta linha, alguns trabalhos também exploram informações textuais [19] ou a correlação de textos e imagens [7, 44] para descrever uma sequência de acontecimentos em desastres, geralmente explorando a semântica textual ao invés da visual.



(a) Evento (b) Evento

(c) Não-Evento (d) Não-Evento

Figura 1.3: Quando consideramos o evento Incêndio da Catedral de Notre Dame que ocorreu em 15 de abril de 2019 as figuras (a) e (b) são do evento, mas as figuras (c) e (d), também recuperadas em uma busca pelo evento, não pertencem ao evento.

Um problema encontrado para a realização dessa análise é que, conforme o número de dados aumenta, também aumentam os dados não pertencentes ao evento, dados replicados ou manipulados. De acordo com *Gupta et al.* [14], dos 7.888.374 *tweets* coletados sobre a Maratona de Boston de 2013 (Tabela 1.1), 29% são informações falsas ou rumores, 51% são comentários e opiniões genéricas e, somente 20% dos *tweets* contêm informações úteis.

Outros exemplos de dados não pertencentes ao evento de interesse são apresentados na Figura 1.3, onde uma busca pelo evento Incêndio da Catedral de Notre Dame retorna também *cartoons* e memes [1] (figuras 1.3(c) e 1.3(d)) que foram mencionados na internet no contexto do evento.

Se analisarmos dados não pertencentes ao evento ou alterados, as consequências para o entendimento podem ser negativas. Por exemplo, se um inocente for considerado suspeito, as ações tomadas em decorrência a uma análise errada impactam na vida do inocente e podem estender-se a outras pessoas (como familiares). Por essa razão, é importante assegurar a corretude da análise.

Dada essa criticidade em entender eventos, devemos ser capazes de filtrar os dados disponíveis. Essa filtragem pode ser realizada por meio da separação dos dados principalmente em dois grupos: Representativos e Não-Representativos. Dentro do grupo

[1] Uma imagem ou vídeo divulgado na internet, geralmente com conteúdo alterado para efeitos de humor.

Representativo estão os dados que representam o momento e o lugar do evento, ou são muito próximos espacialmente e/ou temporalmente (nas redondezas e/ou alguns momentos antes ou depois do evento). Dentro do grupo Não-Representativo estão dados não pertencentes ao evento — que podem ou não estar relacionados — e dados manipulados e falsos.

Dentro do conjunto de dados recuperados de diferentes modalidades — textos, imagens ou vídeos — nós acreditamos que, para descrever eventos, imagens são fontes muito valiosas de informação [54,70]. Por essa razão, optamos por trabalhar na tentativa de descrever imagens semanticamente. Quanto mais imagens representando o evento, maiores as chances de entendê-lo. No entanto, considerando a quantidade massiva de imagens disponibilizadas todos os dias na internet (inclusive durante os eventos), empregar recursos humanos (e.g., investigadores) na tarefa de separar manualmente imagens Representativas de imagens Não-Representativas pode ser tedioso, levar muito tempo e/ou requerer treinamento de especialista, sendo inviável de realizar. Neste cenário, surge a questão principal deste trabalho: *dado um evento, como separar automaticamente imagens Representativas de imagens Não-representativas?*

1.2 Recuperação de Imagens Representativas

Nesta pesquisa de mestrado, trabalhamos com abordagens de recuperação de imagem. Inicialmente, buscamos uma maneira de representar eventos para que a recuperação por representatividade seja otimizada, em outras palavras, a maioria das imagens representativas (preferencialmente todas) seja recuperada. Como nosso foco são imagens, a recuperação é feita por conteúdo de imagem (*Content-Based Image Retrieval* – CBIR), o que torna-a altamente dependente de como esse conteúdo é representado. Uma forma de representar imagens é por meio de descritores que extraem características específicas dessas imagens e dos objetos presentes nelas. Dentre os descritores utilizados, os mais clássicos são os que descrevem cor, textura e forma, ainda utilizados em algumas aplicações como, por exemplo, a apresentada por *Wu et al.* [76] na recuperação de marcas comerciais. Alguns trabalhos realizam a comparação entre esses descritores [42, 43], no entanto, atualmente, muitas aplicações utilizam as características extraídas de redes profundas. Essas redes, treinadas para um contexto específico, buscam descrever semanticamente as imagens de maneira global [21,43,45] ou local, através de pontos de interesse [83], obtendo bons resultados na tarefa de recuperação.

Podemos abordar esse problema de representatividade otimizando duas abordagens diferentes: a representação das imagens, e/ou a maneira como elas são recuperadas. Ambas as abordagens envolvem certo conhecimento do domínio da aplicação. Sem um conhecimento prévio, duas imagens como as apresentadas na Figura 1.4 podem ser consideradas representativas, já que são de maratonas e com pessoas perto da linha de chegada. Porém, se determinarmos que queremos apenas a maratona de Boston de 2013 na qual ocorreram duas explosões, apenas a Figura 1.4(a) é representativa.

Como queremos representar eventos, podemos introduzir conhecimentos gerais de eventos na representação das imagens, como por exemplo, informações de localização

(a) Maratona de Boston de 2013 (b) Maratona de Boston de 2018

Figura 1.4: As figuras (a) e (b) representam maratonas de Boston. O que determina representatividade para um evento é o conhecimento prévio do que queremos recuperar. Se objetivo é representar maratonas de Boston, ambas são representativas. No entanto, se queremos recuperar imagens da maratona de Boston de 2013, onde ocorreram explosões, apenas a figura (a) é representativa.

geral (hemisfério e continente), estação do ano, ou conhecimentos mais específicos de acordo com o tipo do evento como: se foi um ataque terrorista ou catástrofe natural, se ocorreu em uma maratona ou em um concerto musical.

Apesar da grande quantidade de dados disponíveis, a anotação de todos os possíveis eventos (explosões, tiroteios, inundações, incêndios, etc.) que podem vir a ser de interesse é impraticável. Isso faz com que nosso problema tenha um cenário aberto [57], onde um evento de interesse pode ser desconhecido. Além disso, se considerarmos dois eventos, mesmo que sejam do mesmo tipo (como duas explosões, por exemplo) podem haver aspectos muito diferentes (como localização, clima ou quantidade de pessoas reunidas), necessitando de uma grande quantidade e variedade de amostras de treinamento para a generalização de um modelo que separe imagens Representativas de Não-Representativas. Por isso, decidimos aproveitar o conhecimento aprendido por outras redes de propósitos mais gerais (e treinadas com muitas imagens) e informações fornecidas por especialistas forenses.

Com esse conhecimento disponível, optamos inicialmente por tentar representar eventos de acordo com as características que parecem ser comumente discriminativas. Se observarmos a Figura 1.5 e considerarmos a Figura 1.5(a) nossa imagem de referência (a que sabemos ser representativa para o evento), podemos tentar distinguir outras pertencentes ao evento, buscando, por exemplo, por similaridades com relação ao local representado nas imagens (e.g., parques, estádios, prédios), as pessoas presentes (e.g., convidados, vítimas, suspeitos), a disposição dos objetos (e.g., carros, armas, acessórios), dentre outras coisas. Essas características podem nos proporcionar noções de localização e temporalidade. Podemos dessa maneira, definir nossos eventos de interesse como acontecimentos que ocorreram em uma localização e intervalo de tempo específicos e o desafio torna-se determinar qual a localização e tempo representadas em cada uma das imagens analisadas.

Para determinar a pertinência de uma imagem a um evento, decidimos abordar de forma indireta o fator temporal, em outras palavras, utilizamos elementos como pessoas, objetos e até mesmo algumas características do lugar para tentar traduzir temporalidade. Optamos por decompor os eventos nesses componentes representativos já mencionados —

(a) Casamento Real William e Catherine

(b) Casamento Real Elizabeth e Phillip

(c) Maratona de Boston de 2018

(d) Casamento Real William e Catherine

Figura 1.5: As figuras (a) e (d) representam o evento de interesse, o casamento real entre o príncipe William e Catherine Middleton. Podemos utilizar características visuais para separar eventos, como o local apresentado na imagem, as pessoas presentes e os objetos que aparecem. Algumas vezes, como é o caso da figura (b) que representa o casamento real entre a rainha Elizabeth e o príncipe Phillip, somente um tipo de característica visual não é suficiente (ambos os casamentos ocorreram no mesmo lugar). A figura (c) é da maratona de Boston de 2018 e não tem qualquer relação com as demais.

lugares, objetos, pessoas — com o objetivo de capturar essas características de eventos no processo de representação. Em linhas gerais, seguimos um processo de quatro etapas principais representadas na Figura 1.6: (**a**) decomposição do evento; (**b**) representação das imagens de acordo com os componentes da decomposição; (**c**) combinação da representação dos componentes; e (**d**) recuperação de imagens representativas para o evento.

1.3 Desafios na Análise de Representatividade

Um dos primeiros desafios durante os experimentos envolvendo a recuperação de imagens representativas foi a dificuldade em encontrar conjuntos de dados específicos para a avaliação dos modelos nessa tarefa. Nosso objetivo era testar nossas abordagens em cenários próximos do real. Por exemplo, em um cenário real, um conjunto grande de imagens retornadas pela busca por *Explosão Maratona de Boston* poderia conter imagens de maratonas de Boston de outros anos que não o da explosão. Nós queríamos ser capazes de verificar a capacidade dos métodos em separar imagens de eventos muito similares que pudessem ser retornados por essa busca. Para isso, anotamos três conjuntos de dados de eventos reais, dois deles no contexto forense. O primeiro conjunto de dados apresenta como representativas as imagens do casamento real entre o príncipe Willian e Catherine

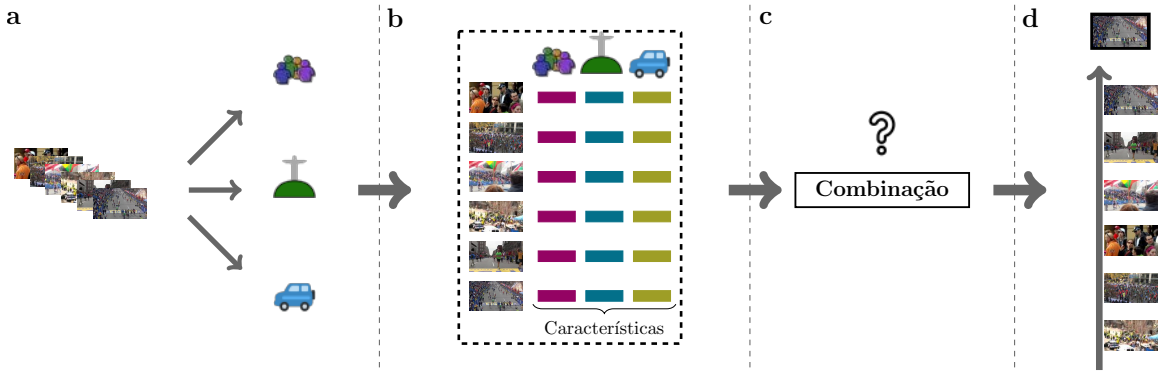


Figura 1.6: Processo de quatro etapas principais seguidas neste trabalho. Inicialmente, (**a**) decompomos o evento nos componentes lugares, objetos e pessoas; na sequência, (**b**) representamos as imagens analisadas de acordo com cada componente; com as imagens representadas, (**c**) propomos abordagens de combinação das características dos componentes; por fim, (**d**) utilizamos as características finais para representação de cada imagem e recuperação das imagens mais representativas de acordo com uma consulta.

Middleton; o segundo apresenta o incêndio da Catedral de Notre Dame, Paris; e o último contém imagens das explosões ocorridas na Maratona de Boston de 2013. No Capítulo 3 descrevemos com detalhes a forma como os conjuntos de dados mencionados foram montados.

O próximo desafio encontrado foi o fato de, em um cenário real, não termos dados anotados em grande quantidade. Considerando que um especialista forense poderia fornecer um número limitado de anotações em imagens, buscamos compreender padrões presentes nas imagens e utilizar métodos para a separação por representatividade. Para isso foi necessário compreender como imagens representativas e não-representativas relacionavam-se entre si e com os componentes representativos. Por essa razão, nós propomos nossa primeira representação de imagens de eventos, o *Espaço Semântico do Evento* (Capítulo 4). Nosso objetivo principal com essa representação foi representar cada imagem explicitamente em um espaço de baixa dimensionalidade baseado-se em Imagens Representativas do Evento (ERI) indicadas por um especialista.

Percebemos que essa representação, apesar da baixa dimensionalidade, conseguiu obter precisão na recuperação equiparável aos métodos que utilizam características obtidas por redes profundas. No entanto, percebemos também que cada evento apresenta desempenhos diferentes para os componentes representativos (aspectos discriminativos para eventos), em outras palavras, cada componente contribui em quantidades diferentes dependendo do evento de interesse. Por essa razão, a precisão na recuperação do método proposto é limitada, já que os componentes são tratados com igual contribuição.

Aprender a combinar esses componentes pareceu ser a solução natural para esse problema, no entanto, ainda teríamos que lidar com a falta de dados para treinamento (anotados). Nosso próximo método proposto é chamado de *Espaço Combinado de Evento* e tem como objetivo aprender um espaço de representação combinando as características obtidas para os componentes representativos (Capítulo 5). Ele é um método supervisionado que utiliza uma rede rasa apenas para combinação de características previamente extraídas. Essas características são obtidas com o uso de redes profundas especializadas

em um dos componentes representativos, cujo treinamento foi realizado em conjuntos de dados maiores (mesmo que não para um evento). O fato de trabalharmos com uma rede rasa para combinação de características de redes profundas possibilitou o aprendizado com um número inferior de amostras de treinamento, tornando-se viável para a aplicação em eventos reais.

Por fim, destacamos as principais contribuições deste trabalho de mestrado: a coleta e anotação de três conjuntos de dados contendo imagens representativas e não-representativas de três eventos do mundo real; a proposta de um método semi-supervisionado para representação de imagens de evento em baixa dimensionalidade; e, a proposta de um método supervisionado e uma rede de combinação de componentes representativos de evento que apresenta bons resultados mesmo com poucas imagens de treinamento.

1.4 Organização da Dissertação

No que se refere a organização desta dissertação: no Capítulo 2 descrevemos conceitos utilizados nesta pesquisa, direta ou indiretamente; no Capítulo 3 apresentamos os conjuntos de dados utilizados; no Capítulo 4 descrevemos o primeiro método de representação de imagens proposto, Espaço Semântico de Evento, apresentando o conjunto de experimentos e resultados obtidos; no Capítulo 5 descrevemos o segundo método de representação de imagens proposto, Espaço Combinado de Evento, onde também apresentamos os experimentos realizados e resultados obtidos. Por fim, concluímos esta dissertação no Capítulo 6 apresentando as contribuições deste mestrado e possíveis caminhos ainda a serem explorados relacionados a esta pesquisa.

Capítulo 2

Conceitos

Neste capítulo, apresentamos alguns conceitos que serão utilizados ao longo do trabalho. Incluímos a definição de Evento utilizada (Seção 2.1), assim como de dados *Representativos* e *Não-Representativos* (Seção 2.2). Além de descrever alguns conceitos importantes, também apresentamos as escolhas feitas durante o trabalho, includindo: métodos de comparação e descritores de imagens (Seção 2.4), técnicas utilizadas para classificação e combinação de características (Seção 2.6), o método utilizado para redução de dimensionalidade (Seção 2.7) e, por fim, as técnicas escolhidas para construção e agregação de *rankings* (Seção 2.8).

2.1 Evento

Definimos um evento $\mathcal{E}$ como um acontecimento ocorrido em um tempo e espaço específicos. O acontecimento de um evento pode ser cotidiano como acordar, ir ao mercado ou receber uma correspondência; mas também pode ser único, como um casamento; ou inesperado, como uma explosão ou tiroteio. Nosso foco é a análise de eventos não-cotidianos, em especial, eventos onde houveram muitas testemunhas e com quantidades massivas de dados compartilhados pela internet.

Podemos dizer que nosso foco são eventos com destaque midiático. Ainda dentro desse conjunto de eventos em destaque, podemos separar os que são de contexto forense ou não. Quando falamos de eventos no contexto forense nos referimos a eventos nos quais surgem questões com necessidade de resposta como: *"quando"*, *"quem"*, *"como"*, *"onde"* e o *"porquê"* [3]. Como apresentamos no Capítulo 3, vamos analisar três conjuntos de dados que descrevem: um evento de contexto não-forense (Casamento Real) e dois eventos de contexto forense (Incêndio na Catedral de Notre-Dame e Explosão na Maratona de Boston), com o objetivo de verificar o desempenho dos métodos em ambos os contextos.

2.2 Dados *Representativos* e *Não-Representativos*

Podemos dividir os dados obtidos para um evento $\mathcal{E}$ em dois grupos principais: *Representativos* e *Não-Representativos*. Dentre os dados do grupo *Representativo* estão aqueles que pertencem ao evento $\mathcal{E}$ e, de alguma maneira, podem ajudar no entendimento deste. Já

dentre os dados do grupo *Não-Representativo* estão aqueles que não pertencem ao evento $\mathcal{E}$, mas que podem ou não apresentar alguma similaridade com ele.

Nós ainda dividimos os dados *Não-Representativos* em três sub-grupos de acordo com o nível de similaridade para $\mathcal{E}$. Estes sub-grupos são *Muito Próximo* (MP), *Próximo* (P) e *Distante* (D) e, cada um desses grupos possui as seguintes características:

1. *Muito Próximo* (MP): são dados que possuem similaridades semânticas com o evento $\mathcal{E}$ como o mesmo lugar onde este aconteceu e/ou mesmas pessoas de $\mathcal{E}$;

2. *Próximo* (P): dados que apresentam lugares ou situações similares, que podem confundir-se com o evento $\mathcal{E}$;

3. *Distante* (D): dados sobre tópicos gerais, sem relação com o evento $\mathcal{E}$.

Como nosso foco é a análise de imagens, a partir de agora falaremos sobre imagens *Representativas* e *Não-Representativas*, seguindo os critérios e categorias definidas para dados no geral.

2.3 Aprendizado

Para realizar essa separação em dois grupos de imagens, *Representativas* e *Não-Representativas*, o problema pode ser modelado como aprendizado supervisionado, não-supervisionado ou semi-supervisionado.

No aprendizado supervisionado partimos do pressuposto de que é possível obter amostras rotuladas em quantidade suficiente de forma a treinar modelos que alcancem a generalização. Assim, a tarefa de predição da classe de uma nova amostra é realizada com base em modelos que aprenderam a distribuição das classes vistas no treinamento [27]. No entanto, nem sempre contamos com dados anotados ou em quantidade suficiente para modelar um problema supervisionado. Para superar essa limitação, podemos utilizar técnicas de aprendizado não-supervisionado que lidam com dados não anotados ou, técnicas de aprendizado semi-supervisionado que utilizam pequenas quantidades de dados anotados.

Com o uso de técnicas não-supervisionadas, nós poderíamos obter agrupamentos de acordo com similaridades (dissimilaridades) entre amostras, sem o uso de rótulos explícitos. Esses agrupamentos podem ser realizados pela comparação das representações dos dados usando distâncias (ou métricas de similaridade) para o estabelecimento de relações de proximidade [81]. Já com as técnicas semi-supervisionadas, podemos utilizar alguns dados como referenciais de cada grupo, inspirados pelo aprendizado supervisionado, e estabelecer esses relacionamentos entre amostras, como no aprendizado não-supervisionado [72]. Apesar da associação do termo aprendizado semi-supervisionado às técnicas de propagação de rótulos, neste trabalho utilizamos o conceito para descrever técnicas que utilizam uma pequena quantidade de imagens anotadas como referências.

2.4 Comparação de Imagens

Nosso problema lida com o desafio da falta de imagens rotuladas como *Representativas* ou *Não-Representativas*. Por essa razão, exploramos técnicas não-supervisionadas ou semi-supervisionadas. Experimentamos métodos de comparação de imagens, buscando encontrar dissimilaridades/similaridades que pudessem separá-las em *Representativas* e *Não-Representativas*. Uma das abordagens — e a que utilizamos nesse trabalho — que realiza a comparação entre imagens é a *Recuperação de Imagem por Conteúdo*.

2.4.1 Recuperação de Imagem por Conteúdo

A Recuperação de Imagem por Conteúdo, ou *Content-Based Image Retrieval* (CBIR), envolve um conjunto de métodos que buscam comparar imagens com base em seu conteúdo representado. Técnicas de CBIR buscam determinar quais imagens são consideradas mais próximas de uma dada imagem de consulta (vide Seção 2.8). Espera-se que o conteúdo representado consiga capturar a semântica da imagem, através de representações de conceitos de alto nível como pessoas e objetos presentes nela, locais representados e/ou ações sendo realizadas. No entanto, esse é um dos maiores desafios de CBIR.

Esse desafio é chamado de *gap semântico* [62, 75] que consiste na dificuldade em representar conceitos de alto nível presentes na imagem — como por exemplo, um pai e um filho jogando bola em um gramado ao pôr-do-sol — utilizando características de baixo nível [2, 31, 32], extraídas da imagem por meio de descritores que geram valores computacionalmente comparáveis. Descritores de imagem tentam superar esse *gap* de diversas maneiras, algumas delas descritas na Seção 2.4.2.

2.4.2 Descritores de Imagem

Os descritores utilizados para extração de características de imagem podem ser divididos em locais ou globais [60,62,69], além de baseados em engenharia de características (*hand-crafted*) ou orientados a dados [86]. Quando falamos de descritores locais, incluímos os descritores que utilizam pontos de interesse da imagem como características para comparação. Em geral, esses descritores apresentam invariâncias a escala, rotação e translação de objetos [86], o que auxilia na comparação sob diferentes perspectivas. Um dos métodos que utiliza esse conceito de pontos de interesse é o *Scale-Invariant Features* (SIFT) [36,37].

Já no grupo de descritores globais — principalmente baseados em técnicas de processamento de imagem — temos diversos descritores que utilizam características de cor, forma e textura [2, 31, 32]. Esse tipo de característica já foi extensivamente explorado, principalmente em tarefas de recuperação de imagem [9]. No entanto, seu uso pode não dispor de todas as invariâncias como descritores locais e, pode requerer algumas técnicas de pré-processamento e/ou segmentação de imagem que tornariam o problema ainda mais complexo, considerando nosso cenário de descrição de eventos.

Visando superar o *gap semântico* entre o que representamos e o que a imagem significa, atualmente, as técnicas mais utilizadas passaram das baseadas em engenharia de características — as técnicas mencionadas anteriormente — para orientadas a dados. O objetivo das técnicas orientadas a dados é aprender como representar o que queremos comparar

nas imagens. Como exemplo de método de extração de características locais orientado a dados temos o *Learned Invariant Feature Transform* (LIFT) [77], baseado no SIFT, mas que inclui no processo o aprendizado dos pontos de interesse. Já como exemplo de método de extração de características globais orientado a dados, podemos destacar o uso de Redes Neurais Convolucionais Profundas (vide Seção 2.4.3) — que obtém ótimos resultados na tarefa de recuperação de imagens [45] — utilizando filtros nas camadas iniciais para extração de características, baseados nas técnicas de processamento de imagem.

Optamos por abordar as técnicas de extração de características orientadas a dados. Descritores locais parecem uma boa solução considerando sua invariância a escala, rotação e translação. No entanto, quando consideramos nossa definição de evento, percebemos que os acontecimentos não-cotidianos — que podem inclusive ser de contexto forense — abrem espaço para variações muito grandes na aparência do local (do evento) em um período de tempo curto.

Considere como exemplo um evento (imaginário) onde uma bomba explode ao lado de um prédio, podemos ter diferentes tipos de imagens: instantes antes da explosão com a fachada visível, no momento da explosão com fumaça obstruindo a fachada, depois da explosão com a fachada destruída. Os três tipos de imagens podem ser classificados como *Representativos*, no entanto, pode não ser possível encontrar pontos de interesse com correspondência entre essa imagens, já que o ambiente sofre muitas modificações.

Outro exemplo (real) é o incêndio da *Grenfell Tower* — bloco de apartamentos em North Kensington, West London — ocorrido em 14 de Junho de 2017. Durante as horas de duração do incêndio, o fogo consumiu de cima para baixo as quatro fachadas de maneiras diferentes. As partes já consumidas pelo fogo dificilmente apresentariam pontos de interesse para a comparação. Algumas imagens do prédio são apresentadas na Figura 2.1, com o prédio intacto, e com diferentes estágios do incêndio.

Dada a imprevisibilidade dos eventos de interesse que iremos abordar, além de optarmos por trabalhar com a extração orientada a dados, decidimos extrair características globais de imagem, evitando a necessidade de encontrar pontos de interesse – que podem não estar mais presentes. Para isso utilizamos Redes Neurais Convolucionais Profundas.

2.4.3 Redes Neurais Convolucionais Profundas

Antes de descrevermos redes neurais convolucionais profundas, precisamos apresentar o conceito que deu origem a elas: as Redes Neurais. As Redes Neurais surgiram inspiradas nas redes neurais biológicas compostas por neurônios interconectados. Um dos primeiros trabalhos apresentando unidades baseadas nos neurônios biológicos foi o trabalho propondo o *Perceptron* por *Rosenblatt* [50]. Basicamente, as redes neurais possuem a seguinte estrutura: uma camada de entrada, seguida por camadas escondidas, finalizando com uma camada de saída, conforme a Figura 2.2.

Dado um conjunto de dados na camada de entrada, os valores de cada neurônio são multiplicados por pesos das conexões que os ligam aos neurônios da próxima camada. O processo de aprendizagem propõe o ajuste desses pesos, de forma que a saída atinja o resultado esperado. Dessa maneira, redes neurais são capazes de aproximar funções não-lineares e auxiliar na solução de inúmeros problemas [25].

Figura 2.1: Exemplos de diferentes imagens da *Grenfell Tower*, onde pode haver dificuldade em encontrar pontos de interesse correspondentes. Imagens da esquerda para a direita apresentam o prédio: antes do incêndio; em estágio avançado no incêndio; com o fogo sendo extinto; com quase todo o fogo extinto. Fonte: *BBC News*.

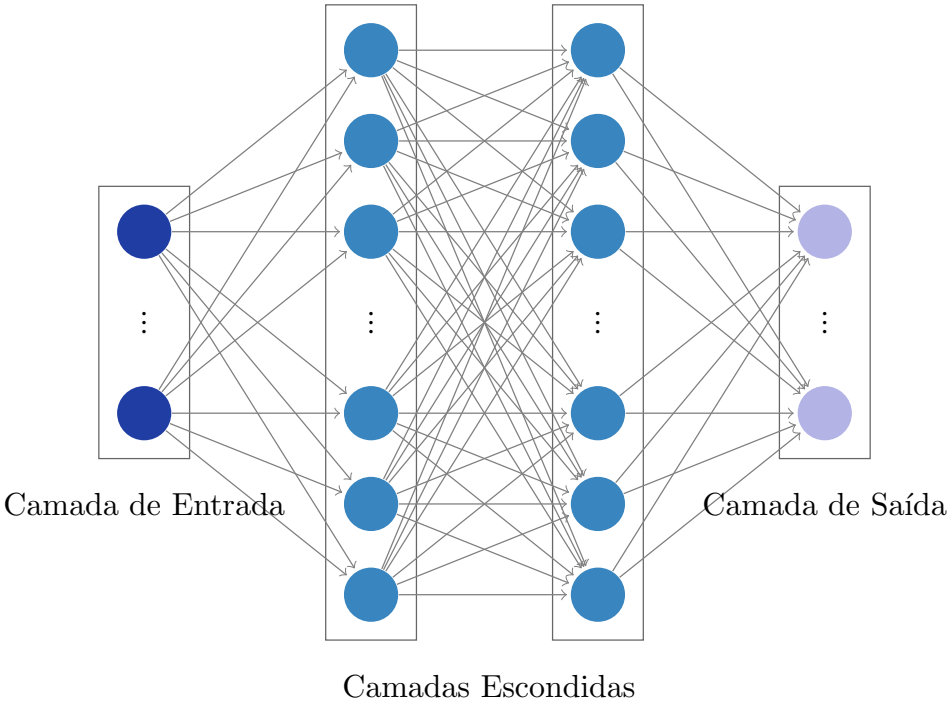


Figura 2.2: A arquitetura Genérica de Redes Neurais é composta por uma camada de entrada, camadas escondidas e uma camada de saída. Cada neurônio pertencente a uma camada está conectado a todos os neurônios da camada subsequente.

Além desse processo de alimentação para a frente (*feed-forward*) [66], é necessário definir como o erro da saída pode auxiliar no ajuste dos pesos. Para isso, redes neurais utilizam o algoritmo *Back-Propagation* [51] para a propagação do erro para as camadas

anteriores. Esse algoritmo utiliza uma função de perda que avalia o quanto as decisões estão erradas e qual o erro a ser propagado para correção dos pesos.

Com o passar do tempo, as Redes Neurais tornaram-se mais complexas, aumentando sua profundidade e número de neurônios. Quando tratamos de imagens, redes muito utilizas são as Redes Neurais Convolucionais Profundas, ou *Convolutional Neural Networks* (CNNs). Elas foram inicialmente propostas (em formato similar ao atualmente utilizado) por *LeCun et al.* [30] com a rede chamada *LeNet*, em um trabalho de reconhecimento de caracteres em documentos. Nesse primeiro trabalho, a proposta era atingir a invariância à escala, distorção e translação, com campos receptivos — filtros convolucionais com parâmetros compartilhados — e amostragem temporal e espacial [30]. No entanto, somente mais tarde as CNNs se popularizaram, especialmente, devido aos resultados apresentados por *Krizhevsky et al.* [28] no conjunto de dados *Imagenet* [11].

A ideia dessas redes é aprender representações diretamente das imagens, utilizando arquiteturas que contém: filtros convolucionais em suas primeiras camadas, de onde são extraídas características globais de baixo nível da imagem; e camadas completamente conectadas (similares as Redes Neurais), para aprender características específicas da aplicação para qual a rede está sendo treinada – baseadas nas características de baixo nível da imagem [12, 67].

Uma das principais características das CNNs é a capacidade de generalização — especialmente, nas camadas iniciais — fazendo com que essas redes possam ser adaptadas para outros conjuntos de dados e/ou aplicações [12, 79]. Para muitos problemas não é possível treinar uma CNN completa, já que essas redes convolucionais exigem muitos dados rotulados para a otimização da grande quantidade de parâmetros existentes [41]. Por isso, diversos trabalhos fazem uso da capacidade de generalização e extraem características por meio de CNNs pré-treinadas — com outros conjuntos de dados, para outras aplicações — e adaptadas para o problema específico [12, 55, 73, 74, 80, 83].

Essa adaptação pode ser realizada obtendo as saídas (para cada imagem a ser representada) de uma ou mais camadas combinadas da CNN pré-treinada, utilizando-as como características para representação das imagens. Essas características são dadas como entrada para outra Rede Neural e/ou classificador (menor) que será treinado (vide Seção 2.6), adaptando-o para a aplicação em questão [73]. Ou ainda, pode ser realizado um processo de *fine-tuning* que consiste no treinamento somente de algumas camadas da CNN [12, 74, 83], reduzindo assim o número de parâmetros a serem otimizados.

Como a falta de imagens rotuladas por representatividade é um dos nossos desafios, nós também utilizamos aqui características extraídas por CNNs pré-treinadas. Entraremos em maiores detalhes de como essas características são extraídas e utilizadas nos capítulos 4 e 5.

Uma metodologia também utilizada para aumentar o número de imagens anotadas é descrita na Seção 2.5. Além disso, também exploramos formas de treinamento que exigem menos dados anotados e podem explorar as representações extraídas por CNNs pré-treinadas como dados de entrada (Seção 2.6).

2.5 Aumentação de Imagens

Diversos métodos de aumentação de dados podem ser empregados quando existe a falta de dados anotados. Quando consideramos aumentação de imagens podemos citar transformações geométricas [61], redes adversariais generativas [56], transferência de estilo [22], dentre outras [61].

A aumentação de dados é uma das técnicas utilizadas para reduzir o sobreajuste/subajuste (*overfitting/underfitting*) da rede no conjunto de treinamento. Esse sobreajuste/subajuste pode ser causado pelo número reduzido de amostras de treinamento e/ou pelo desbalanceamento de classes, dificultando a generalização do modelo [61]. Nesse trabalho utilizamos a técnica de Aumentação de Imagens.

Algumas das transformações que podem ser aplicadas em imagens são espelhamento, modificações no espaço de cor, recorte, rotação, translação, ampliação, redução e inserção de ruído.

No entanto, um dos pontos críticos a ser observado no processo de aumentação é a segurança na preservação dos rótulos. Podemos prejudicar o aprendizado da rede ao aplicarmos uma técnica de aumentação que modifica os rótulos das imagens. Isso porque, ao transmitirmos o rótulo das imagens originais para as aumentadas estaremos anotando imagens erroneamente. Um exemplo apresentado por *Shorten et al.* [61] é a aplicação de transformações de rotação e espelhamento nas imagens, que é segura no conjunto de dados Imagenet [11], ou seja, não modifica o rótulo original da imagem, mas não é segura em conjuntos de dados de dígitos, podendo causar confusões entre 6 e 9, por exemplo.

2.6 Técnicas Rasas para Classificação e Comparação

O aprendizado de máquina é utilizado para inúmeras tarefas envolvendo imagens como: classificação [28], recuperação (como mencionado anteriormente) [45], detecção e reconhecimento de objetos e faces [4, 63], dentre outras. Já falamos da capacidade de CNNs para extração de características de imagens. Podemos ainda ressaltar que essas redes podem ser utilizadas inclusive como uma sequência completa para a solução dos problemas (*end-to-end*) caso existam dados suficientes para treinamento [41].

No entanto, apesar das CNNs estarem entre os métodos estado da arte, não são as únicas maneiras de realizar o aprendizado. De fato, quando não temos dados suficientes para treinamento, podemos utilizar informações já aprendidas por CNNs pré-treinadas como maneira de extrair características de imagem (conforme mencionado na Seção 2.4.3). Assim que essas características são extraídas, outros métodos — que exijam menos dados rotulados — de classificação ou de aprendizado de distâncias entre imagens podem ser utilizados. Alguns desses métodos, que abordaremos ao longo desse trabalho, são as Florestas Aleatórias e as já mencionadas Redes Neurais.

2.6.1 Florestas Aleatórias

Florestas Aleatórias [17] é um método de classificação que utiliza um conjunto de Árvores Binárias de Decisão, construídas com conjuntos de atributos aleatoriamente selecionados.

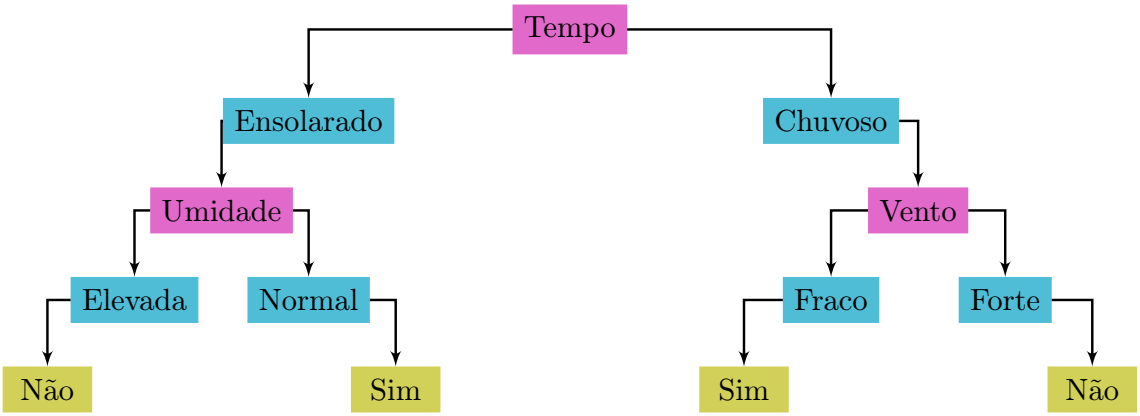


Figura 2.3: Um exemplo comum para ilustrar a classificação por Árvore Binária de Decisão é a decisão de jogar ou não Tênis. Primeiro podemos analisar se o Tempo está Ensolarado ou Chuvoso, de acordo com a decisão analisamos atributos diferentes. Se está Ensolarado, podemos analisar a Umidade. Se está Chuvoso, podemos analisar o Vento.

O princípio de uma Árvore Binária de Decisão é quebrar uma decisão complexa em diversas decisões menores [53]. Um exemplo clássico é a decisão de jogar ou não Tênis, que poderia ser dividida na análise dos atributos Tempo, Umidade e Vento, conforme Figura 2.3.

No entanto, a seleção dos atributos importantes para a decisão é um desafio. Assim, as Florestas Aleatórias surgem como maneira de superar esse desafio. Como são construídas árvores de decisão com atributos aleatórios, mesmo as árvores muito ruins não impactam tanto na decisão final, já que ela é tomada com base em todo o conjunto de árvores, como no exemplo da Figura 2.4.

2.6.2 Redes Neurais

Conforme apresentado na Seção 2.4.3, as Redes Neurais (assim como CNNs) possuem diferentes aplicações. Nosso foco são as aplicações que incluem classificação e comparação entre imagens. Sabemos que a função de perda é a responsável por determinar o quão errada está a Rede Neural durante o treinamento. Assim, ela deve ser definida de acordo com a resposta que desejamos obter.

Quando falamos de classificação, uma das funções de perda mais utilizadas é a Entropia Cruzada (*Cross-Entropy Loss*) [39]. Essa função é dada pela Equação 2.1, onde M é o número de classes, $y_{o,c}$ é o valor esperado (0 ou 1) para a classe c da imagem o e $p_{o,c}$ é o valor predito para a classe c da imagem o. Essa função é utilizada para modelos de classificação onde esperamos uma probabilidade entre 0 e 1 de uma amostra pertencer a cada uma das classes do problema.

$$CE = -\sum_{c=1}^{M} y_{o,c} log(p_{o,c}) \tag{2.1}$$

Além da classificação, as redes neurais podem também ser utilizadas para aprendizado de métricas de comparação. Durante o aprendizado de métricas (*Metric Learning* [29]), o objetivo é aprender a distribuição dos dados e encontrar *manifolds* [84] onde amostras da mesma classe estejam próximas e de classes diferentes distantes.

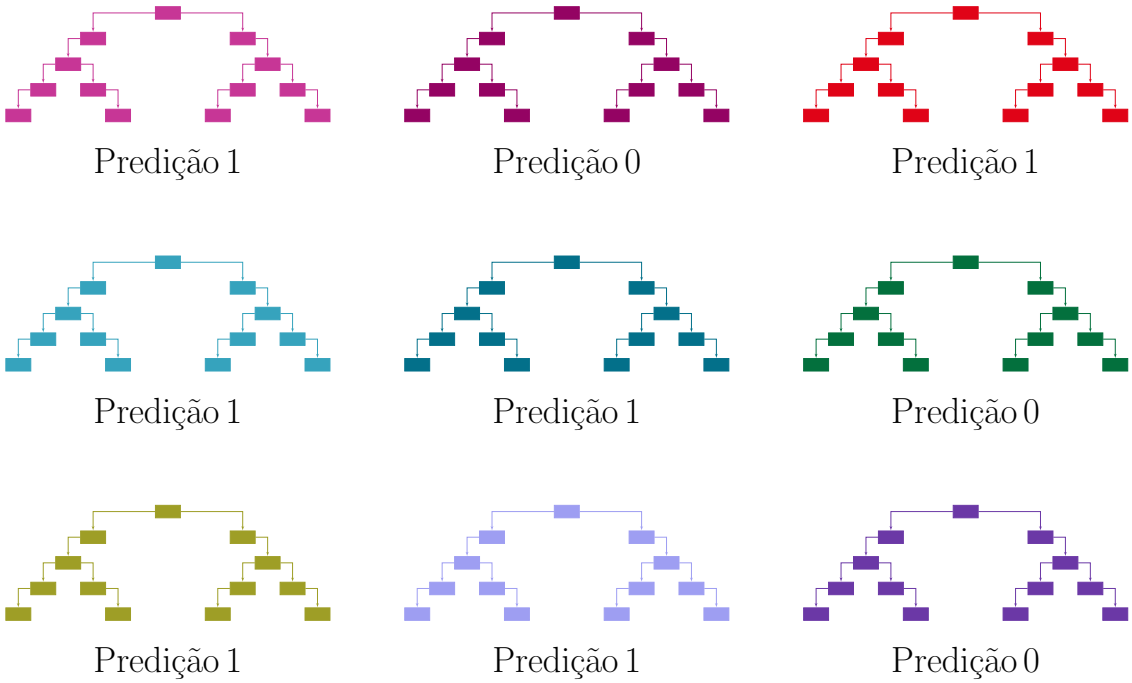


Figura 2.4: Exemplo de uma Floresta Aleatória que deve predizer entre classes 0 ou 1. Dentre as nove árvores da floresta, seis predisseram 1, portanto, essa foi a classe final predita.

Manifolds são espaços topológicos localmente Euclidianos. A hipótese de *manifold* propõe que o espaço de entrada é composto de múltiplos *manifolds* de baixa dimensionalidade onde estão os pontos. Os pontos pertencentes à mesma classe deveriam estar no mesmo *manifold* [72]. Muitas vezes a representação dos dados não é capaz de obter essa separação. Nesses casos, pode-se utilizar um aprendizado de *manifolds*, aproximando *manifolds* com pontos pertencentes à mesma classe.

Uma das maneiras de aprender *manifolds* é utilizando redes que sejam capazes de comparar duas ou mais amostras. Essas redes procuram encontrar uma representação para cada amostra em um espaço vetorial que leva em consideração relacionamentos entre as amostras [72]. Uma estrutura possível de ser usada para realizarmos comparações e obtenção de representações é a Rede Siamesa [6], onde múltiplas sub-redes compartilham os mesmos pesos.

Em sua estrutura mais simples, podemos utilizar a Perda Contrastiva durante o treinamento. Nesse caso, a Rede Siamesa é composta por dois ramos que compartilham pesos. Cada ramo recebe uma amostra ($\mathbf{I}_a$ e $\mathbf{I}_b$), conforme apresentado na Figura 2.5. A forma de representar as amostras é aprendida de forma a aproximar amostras da mesma classe no *manifold* e distanciar de classes diferentes.

A função de Perda Contrastiva (*Contrastive Loss*) apresentada por *Hadsell et al.* [15] é descrita pela Equação 2.2, onde y representa o valor esperado (1 se as amostras forem da mesma classe e 0 caso contrário), p representa o valor de distância obtido e N é o número de pares analisados pela função de perda. O parâmetro de margem m indica a distância necessária entre as amostras de pares negativos para que a função de perda não

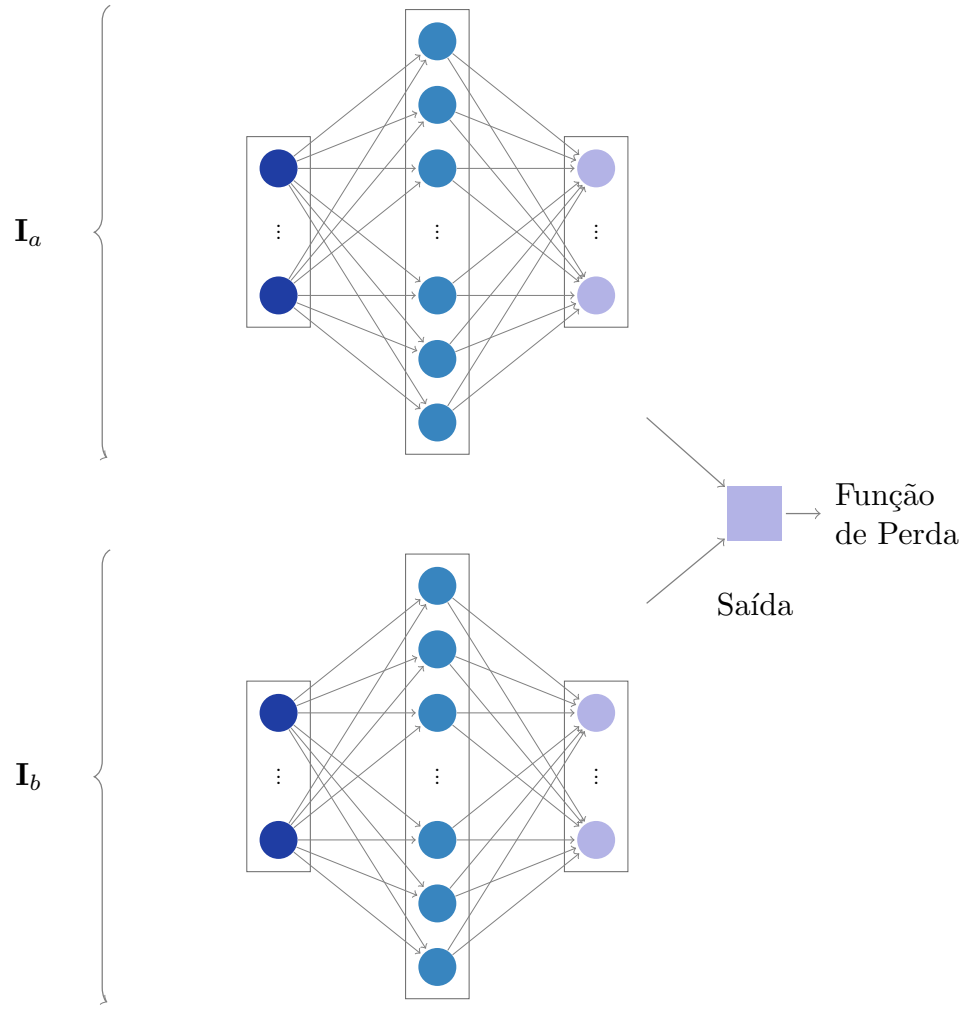


Figura 2.5: Exemplo de uma arquitetura simples siamesa. Cada ramo da rede é constituído por redes que compartilham os pesos. Cada ramo tem como entrada uma amostra ($\mathbf{I}_a$ e $\mathbf{I}_b$) que desejamos saber se pertencem à mesma classe ou não. Ao final desses ramos, as características são utilizadas para cálculo da Função de Perda. O ajuste dos pesos é também realizado pela propagação do erro.

aumente significativamente.

$$C = \sum_{i}^{N} [y \times p^2 + (1 - y) \times (max(m - p, 0))^2] \tag{2.2}$$

Com o intuito de aproximar amostras da mesma classe e de distanciar amostras de classes diferentes, podemos utilizar uma estrutura siamesa de três ramos com a função de Perda Tripla (*Triplet Loss*). Essa função proposta por *Scroff et al.* [58] aumenta o valor da função de perda quando amostras da mesma classe estão distantes e/ou quando amostras de classes diferentes estão próximas, de acordo com a Equação 2.3, onde N é o número de triplas analisadas pela função de perda, $f(\mathbf{I}_i^a)$ é considerada a representação da amostra de interesse (âncora), $f(\mathbf{I}_i^p)$ é a representação de uma amostra da mesma classe que a âncora, e $f(\mathbf{I}_i^n)$ é a representação de uma amostra de classe diferente da âncora. Como as representações são vetores, para o cálculo das distâncias entre as representações

é utilizada a Norma Euclidiana denotada por $||f(\mathbf{y}) - f(\mathbf{z})||$. Assim como a função de Perda Contrastiva (Equação 2.2), a função de Perda Tripla também utiliza o parâmetro de margem m, dessa vez indicando a distância necessária entre os exemplos positivos e negativos para que a função de perda não aumente significativamente.

$$T = \sum_{i}^{N} [max(||f(\mathbf{I}_i^a) - f(\mathbf{I}_i^p)||^2 - ||f(\mathbf{I}_i^a) - f(\mathbf{I}_i^n)||^2 + m, 0)] \tag{2.3}$$

2.7 Redução de Dimensionalidade para Visualização

Além de representar as imagens, buscamos avaliar a qualidade das representações. Uma das maneiras utilizadas é através de técnicas de visualização. Nós utilizamos a projeção das imagens em gráficos de dispersão bidimensionais [13], de forma a ser possível observar se as representações são capazes de aproximar imagens *Representativas* e distanciá-las das *Não-Representativas*.

Para realizar a projeção nos gráficos bidimensionais, no entanto, é necessário definir apenas duas coordenadas x e y para representar cada imagem como um ponto (x, y) no plano cartesiano. As representações utilizadas contam com centenas ou até milhares de atributos. Portanto, é necessário utilizar técnicas de redução de dimensionalidade antes de projetar as representações para visualização.

Uma das técnicas clássicas para redução de dimensionalidade é conhecida como Análise de Componente Principal (*Principal Component Analysis – PCA*), onde o objetivo é projetar os dados em um espaço de dimensionalidade reduzida de forma a maximizar a variância, minimizando a perda de informação [23]. O PCA é uma técnica linear e portanto não é capaz de capturar relação não-lineares que podem estar presentes em dados com dimensionalidades muito altas. Para capturar essas relações, algumas técnicas não-lineares para redução de dimensionalidade podem ser destacadas: *Uniform Manifold Approximation and Projection* (UMAP) [38] e *t-distributed Stochastic Neighbor Embedding* (t-SNE) [71].

2.7.1 Projeção e Aproximação de *Manifold* Uniforme

O UMAP, em português Projeção e Aproximação de *Manifold* Uniforme, utilizado neste trabalho, é uma das técnicas mais atuais de redução de dimensionalidade. Essa técnica procura estabelecer estruturas locais mantendo a estrutura global dos dados. É uma técnica similar ao t-SNE, mas que busca uma aplicação mais geral (não somente para visualização).

Os parâmetros principais desse método induzem a criação de um grafo de relacionamento entre as amostras. Esses parâmetros são: o número de vizinhos e a distância entre eles. Para obtenção da distância, também é necessário definir a métrica de distância utilizada.

2.8 Construção de *Rankings*

Durante o trabalho desenvolvido, a abordagem que utilizamos para separar imagens *Representativas* de *Não-Representativas* foi a CBIR (vide Seção 2.4.1). Para isso, precisamos saber como construir um *ranking* baseado em uma consulta *Representativa*. Aqui vamos definir o conceito utilizado de *ranking*.

Considere $\mathcal{X} = \{\mathbf{I}_i\}_{i=1}^{n}$ o conjunto de n imagens que queremos analisar, e $\mathcal{Q} = \{\mathbf{q}_j\}_{j=1}^{\ell}$ o conjunto de ℓ imagens que já sabemos ser *Representativas*, de acordo com especialistas, que serão utilizadas como consultas. Para recuperar imagens *Representativas* de $\mathcal{X}$ procuramos pelas imagens mais próximas a uma consulta específica $\mathbf{q}_j$. Para isso, precisamos definir a função f que descreverá a imagem tornando-a comparável, e a função δ que pode ser uma métrica de distância ou similaridade (e.g., distância euclidiana) que realizará a comparação, de forma que, $\delta(f(\mathbf{q}_j), f(\mathbf{I}_i))$ defina a distância d_{ij} (ou similaridade s_{ij}) entre as representações de $\mathbf{q}_j$ e $\mathbf{I}_i$.

Definidos os conjuntos $\mathcal{X}$ e $\mathcal{Q}$, assim como as funções f e δ, podemos definir um *ranking* $\mathcal{R}_{\mathbf{q}_j}$ como a permutação das imagens de $\mathcal{X}$ de forma que a primeira posição contenha a imagem mais próxima da consulta $\mathbf{q}_j$, e assim por diante, até que todas as imagens sejam ordenadas de acordo com sua proximidade de $\mathbf{q}_j$.

2.8.1 Função f

Utilizamos inicialmente CNNs pré-treinadas para a extração das características, como função f de descrição (ou representação) das imagens. A maneira como essas características foram combinadas — que também estamos considerando como parte da função f — será discutida nos Capítulos 4 e 5.

2.8.2 Função δ

A função δ é a responsável por quantificar a proximidade entre a representação de cada imagem de $\mathcal{X}$ e a representação da consulta $\mathbf{q}_j$. Essa proximidade pode ser calculada como distância (d_{ij}) ou como similaridade (s_{ij}). Quando falamos de distância, quanto mais próximo d_{ij} está de 0 melhor, indicando pouca distância entre as imagens. Já quando falamos de similaridade, quando maior o valor s_{ij} melhor, indicando muita similaridade entre as imagens.

Uma das métricas de distância mais utilizadas nas tarefas de recuperação de imagem é a Distância Euclidiana apresentada na Equação 2.4, onde $\mathbf{a} = f(\mathbf{q}_j)$ e $\mathbf{b} = f(\mathbf{I}_i)$ são vetores $\mathcal{D}-$dimensionais e a_i, b_i representam os valores de $\mathbf{a}$ e $\mathbf{b}$ na dimensão i. Isso porque o cálculo da Distância Euclidiana é mais simples do que a maioria das outras distâncias (como por exemplo a Mahalanobis ou a Minkwoski [33] de ordem maior que 2) — gastando um tempo bem inferior de processamento e criação dos *rankings* — e consegue manter desempenho aceitável na tafera de recuperação.

$$d_{ij} = \sqrt{\sum_{i}^{\mathcal{D}} (a_i - b_i)^2} \tag{2.4}$$

Existem métricas específicas para o cálculo da similaridade como, por exemplo, a Similaridade de Cossenos [8,33]. Podemos considerar ainda a similaridade como o inverso da distância, ou seja, quando menor a distância entre imagens, maior a similaridade entre elas. Dessa maneira, podemos determinar similaridade de acordo com a Equação 2.5, onde a similaridade s_{ij} é dada pelo máximo valor de d_j (que é o conjunto de todas as distâncias das imagens em $\mathcal{X}$ para a consulta $\mathbf{q}_j$) menos o valor da distância entre $\mathbf{I}_i$ e $\mathbf{q}_j$ (d_{ij}).

$$s_{ij} = max(d_j) - d_{ij} \tag{2.5}$$

Como nosso foco nesse trabalho é a função de representação f utilizada para descrever as imagens, nós optados por utilizar a distância Euclidiana (Equação 2.4) como função δ. Realizamos a conversão apresentada na Equação 2.5 quando existe a necessidade de utilizar similaridade ao invés de distância.

2.8.3 Agregação de *Rankings*

Quando possuímos mais de uma consulta ($\ell > 1$) o resultado da construção de *rankings* é um conjunto de ℓ *rankings* que podem ser diferentes. O objetivo é obter um *ranking* final $\mathcal{R}_F$ que possua as informações de todas os *rankings*, incluindo assim imagens similares a uma ou mais consultas.

Como nosso foco nesse trabalho é a representação das imagens, optamos por utilizar um método simples de agregação de *rankings* para obter nosso $\mathcal{R}_F$. Considerando que a escolha do conjunto de consultas $\mathcal{Q} = \{\mathbf{q}_j\}_{j=1}^{\ell}$ é realizada por um perito e que cada $\mathbf{q}_j$ é *Representativa* para o evento, decidimos recuperar qualquer imagem que seja muito próxima a pelo menos uma consulta. Portanto, o método de agregação utilizado é o *CombMin* [59] dado pela Equação 2.6, onde d_{iF} é o valor mínimo entre as distâncias (d_{ij}) de $\mathbf{I}_i$ para todas as consultas $\mathbf{q}_j$.

$$d_{iF} = min(\{d_{ij}\}_{j=1}^{\ell}) \tag{2.6}$$

Além do *CombMin*, uma das abordagens apresentadas no Capítulo 4 utiliza outro método de agregação chamado *CombSum* [59], como forma de utilizar a contribuição das distâncias de todos os *rankings* somadas, de acordo com a Equação2.7.

$$d_{iF} = \sum_{j=1}^{\ell} d_{ij} \tag{2.7}$$

Com as novas distâncias obtidas, o *ranking* $\mathcal{R}_F$ é gerado pela permutação ordenada dos novos valores de forma crescente. Assim, a imagem no topo do *ranking* final é a imagem que apresentou a menor distância dentre todos os $\mathcal{R}_{\mathbf{q}_j}$.

2.9 Métricas de Avaliação

Seja $\mathcal{R}p = \{\mathbf{rp}_s\}_{s=1}^{T}$ o conjunto de imagens Representativas no conjunto de dados, $\mathcal{NR}p = \{\mathbf{np}_a\}_{s=1}^{B}$ o conjunto de imagens Não-Representativas.

Quando trabalhamos com aprendizado de classificação existem métricas que avaliam a qualidade da separação de classes obtida, uma delas é a Acurácia Balanceada [5]. Considerando nosso problema de duas classes, Representativa ($\mathcal{R}p$) e Não-Representativa ($\mathcal{NR}p$), temos que as imagens Representativas corretamente classificadas são $V\mathcal{R}p$ e as Não-Representativas corretamente classificadas são $V\mathcal{NR}p$. Assim, podemos calcular a Acurácia Balanceada utilizando a Equação 2.8.

$$BAcc = \frac{\frac{|V\mathcal{R}p|}{|\mathcal{R}p|} + \frac{|V\mathcal{NR}p|}{|\mathcal{NR}p|}}{2} \tag{2.8}$$

Quando construímos *rankings*, existem métricas específicas para avaliação em recuperação de informação, dentre elas, Precisão (*Precision*), Revocação (*Recall*), Precisão Média (*Average Precision – AP*) e Média das Precisões Média (Mean Average Precision – MAP) [47]. Considere que, no problema que tentamos resolver, as imagens relevantes são as Representativas. Além disso, considere $\mathcal{R}e_t = \{\mathbf{re}_u\}_{u=1}^{t}$ o conjunto de imagens recuperadas no topo t do *ranking* $\mathcal{R}$.

A Precisão dos *rankings* determina o percentual de imagens Representativas dentre as recuperadas. O valor de Precisão para o topo t pode ser obtido pela Equação 2.9.

$$Precisão_t = \frac{|\mathcal{R}p \cap \mathcal{R}e_t|}{|\mathcal{R}e_t|} \tag{2.9}$$

A Revocação dos *rankings* determina o percentual de imagens representativas recuperadas dentre as representativas. O valor de Revocação para o topo t pode ser obtido pela Equação 2.10.

$$Revocação_t = \frac{|\mathcal{R}p \cap \mathcal{R}e_t|}{|\mathcal{R}p|} \tag{2.10}$$

A métrica AP obtêm a Precisão Média dos valores de Precisão calculados quando a Revocação aumenta. O valor de AP para um *ranking* $\mathcal{R}$ pode ser obtido pela Equação 2.11.

$$AP(\mathcal{R}) = \sum_{u=1}^{|\mathcal{R}|}(Revocação_{u+1} - Revocação_u) \times Precisão_{u+1} \tag{2.11}$$

A métrica MAP é a média dos valores de AP, dados pela Equação 2.11, obtidos considerando todos os *rankings* construídos pelas consultas q_j em $\mathcal{Q}$ (Equação 2.12).

$$MAP = \frac{\sum_j^{\ell} AP(\mathcal{R}_{q_j})}{\ell} \tag{2.12}$$

Capítulo 3

Conjuntos de Dados

Como mencionamos, uma das dificuldades deste trabalho foi a falta de conjuntos de dados que refletissem o problema que queremos resolver. Um dos primeiros conjuntos de dados avaliados era composto por oito eventos [1]: Maratona de Austin, Show Aéreo de Berlin, Maratona de Boston, Furacão Matthew, Oupação de Baltimore, Ocupação de Portland e Oshkosh. Cada um dos eventos contava com 200 imagens relacionadas, totalizando 1.600 imagens.

O problema aqui foi que, como nosso problema aborda cenários reais de recuperação, com quantidades massivas de imagens, esse conjunto de dados bem condicionado foi insuficiente para refletir o mundo real. Por isso, procuramos por outros conjuntos mais adequados.

Encontramos um conjunto de dados disponível para o desafio *Lifelog Moment Retrieval (LMRT)* do ImageCLEFlifelog 2018 [10, 20]. Esse conjunto de dados contém dados de múltiplas modalidades, coletados durante 50 dias no cotidiano de uma pessoa, totalizando de 1.500 a 2.500 imagens por dia, contendo anotações das atividades executadas. Apesar de ser um conjunto muito maior de imagens, os eventos contidos eram fatos cotidianos. Acreditamos que seria interessante abordar eventos específicos e com algum momento em destaque como: explosões, incêndios, casamentos, tiroteios, concertos, dentre outros. Assim, poderíamos simular cenários em que gostaríamos de aplicar os métodos desenvolvidos.

Outro conjunto de dados, dessa vez com eventos de destaque, é o EVVE (EVent VidEo) proposto por *Revaud et al.* [46]. Este é um conjunto de dados de vídeos para aplicações de *Event Repurposing* contendo 13 eventos com seus vídeos positivos e negativos, dentre os quais: concertos, casamentos e discursos. No entanto, como iríamos trabalhar com imagens, precisaríamos extrair os quadros e rotulá-los por representatividade para o evento.

Por essa razão, decidimos construir e rotular nossos próprios conjuntos de dados, que foram três: *Wedding*, *Fire* e *Bombing*. O conjunto de dados *Wedding* foi organizado com base em um conjunto previamente existente no EVVE [46], selecionando alguns vídeos de interesse e extraindo seus quadros. Os demais foram organizados considerando os critérios descritos na Seção 3.4, também com o uso de vídeos dos quais os quadros são extraídos e utilizados como imagens dos conjuntos de dados. As características de cada conjunto de dados são descritas a seguir.

3.1 Wedding

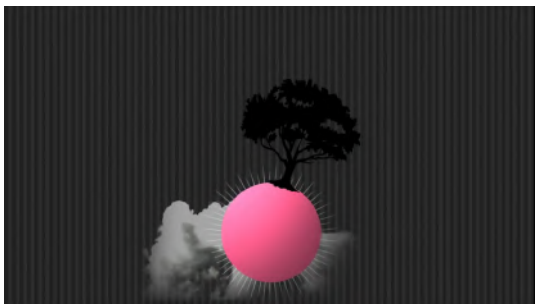
O conjunto de dados *Wedding* é formado por imagens do casamento real do príncipe William e Catherine Middleton. O evento aconteceu no dia 29 de abril de 2011 na Abadia de Westminster em Londres, Reino Unido.

A Figura 3.1 apresenta imagens representativas e não-representativas para o evento contidas no conjunto *Wedding*. A Figura 3.1(a) representa o casamento real do príncipe William e Catherine Middleton, já as demais (figuras 3.1(b), 3.1(c) e 3.1(d)) não representam o evento.

(a) Evento

(b) Rainha Elizabeth II e Príncipe Phillip

(c) Paródia do Evento



(d) Não Relacionada

Figura 3.1: Exemplos de imagens contidas no conjunto de dados *Wedding*. A figura (a) representa o evento, já as demais figuras são consideradas não-representativas, apresentando: (b) o casamento da Rainha Elizabeth II e do Príncipe Phillip que ocorreu no mesmo local; (c) uma paródia do casamento real do príncipe William e Catherine Middleton, e; (d) algo totalmente não relacionado ao evento.

3.2 Fire

O conjunto de dados *Fire* contém imagens do incêndio da Catedral de Notre Dame em Paris. O incêndio começou no dia 15 de abril de 2019 e a Catedral teve o pináculo e a maior parte do telhado destruídos. Os vídeos para este conjunto de dados foram selecionados manualmente para posterior extração dos quadros (vide Seção 3.4).

A Figura 3.2 apresenta imagens representativas e não-representativas para o evento contidas no conjunto *Fire*. A Figura 3.2(a) representa o incêndio da catedral de Notre Dame, já as demais (figuras 3.2(b), 3.2(c) e 3.2(d)) não representam o evento.

(a) Evento

(b) Catedral antes do incêndio

(c) Outra catedral gótica

(d) Não relacionada

Figura 3.2: Exemplos de imagens contidas no conjunto de dados *Fire*. A figura (a) representa o evento, já as demais figuras são consideradas não-representativas, apresentando: (b) a catedral antes do incêndio; (c) a catedral de York Minster que é uma construção gótica, e; (d) algo totalmente não relacionado ao evento.

3.3 Bombing

O conjunto de dados *Bombing* apresenta imagens da detonação de duas bombas de pressão caseiras na Maratona de Boston que ocorreu em 15 de abril de 2013. As explosões resultaram em três mortes e aproximadamente 264 feridos. Os vídeos para este conjunto de dados foram selecionados manualmente para posterior extração dos quadros (vide Seção 3.4).

A Figura 3.3 apresenta imagens representativas e não-representativas para o evento contidas no conjunto *Bombing*. A Figura 3.3(a) representa a Maratona de Boston de 2013 onde ocorreram as detonações, já as demais (figuras 3.3(b), 3.3(c) e 3.3(d)) não representam o evento.

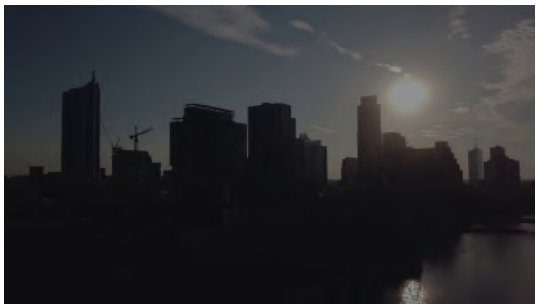
3.4 Escolha dos Vídeos e Rotulação

Os vídeos escolhidos para gerar o conjunto de imagens foram selecionados considerando diferentes níveis de representatividade, i.e., vídeos (a) pertencentes ao Evento ($\mathcal{E}$); (b) mostrando o lugar do evento e/ou mesmas pessoas, ou seja, vídeos Muito Próximos (MP); (c) apresentando atividades similares e/ou lugares similares, ou seja, vídeos Próximos (P); e (d) nada relacionados, ou seja, vídeos Distantes (D). Os vídeos mostrando o lugar do evento ou, lugares ou atividades similares foram escolhidos para aumentar a similaridade entre imagens de diferentes eventos, e assim, testar a robustez dos métodos. Exemplos

(a) Evento

(b) Maratona de Boston 2018

(c) Maratona de Tokyo 2019



(d) Não Relacionada

Figura 3.3: Exemplos de imagens contidas no conjunto de dados *Bombing*. A figura (a) representa o evento, já as demais figuras são consideradas não-representativas, apresentando: (b) a Maratona de Boston de 2018 (mesmo local do evento); (c) a Maratona de Tokyo de 2019; e (d) algo totalmente não relacionado ao evento.

de cada uma das categorias de vídeos (*(a)* E, *(b)* MP, *(c)* P, *(d)* D) são apresentadas nas figuras 3.1, 3.2 e 3.3 para os conjuntos de dados *Wedding, Fire* e *Bombing* respectivamente.

Os vídeos para o conjunto *Wedding*, como já mencionado, foram escolhidos dentre os dados de uma aplicação de repropósito de evento (*Event Repurposing*) [46]. Dentre esse conjunto de vídeos, utilizamos 130 vídeos sendo 54 deles pertencentes ao evento. Desses vídeos extraímos dois quadros por segundo e excluímos parte dos quadros com cenas repetidas representando o evento, gerando um conjunto de imagens reduzido (do evento).

Os conjuntos de dados *Fire* e *Bombing* foram montados a partir de vídeos do YouTube relacionados aos eventos de acordo com os diferentes níveis de representatividade. Para o conjunto *Fire*, foram selecionados um total de 40 vídeos onde 10 deles pertencem ao evento. Para o conjunto *Bombing*, foram selecionados 30 vídeos onde 10 pertencem ao evento. Para criar ambos os conjuntos de dados, extraímos um quadro por segundo dos vídeos selecionados (não excluímos quadros).

O número total de imagens nos três conjuntos de dados é apresentado na Tabela 3.1. Anotamos cada quadro de acordo com a representatividade para o evento, gerando duas categorias de imagens: *Representativa* e *Não-Representativa*. Uma imagem foi considerada Representativa se ela apresenta aspectos importantes do evento, como por exemplo, *i*) se ela pertence ao evento, *ii*) se ela apresenta momentos imadiatamente antes ou depois do evento, e *iii*) se ela ajuda a descrever pontos chave do evento. Os números de images rotuladas como Representativa ou Não-Representativa em cada conjunto de dados também são apresentados na Tabela 3.1.

	Categoria da imagem		
Conjunto de dados	Representativa	Não-representativa	Total
Wedding	439	31.594	32.033
Fire	973	44.107	45.080
Bombing	3.182	45.224	48.406

Tabela 3.1: Quantidade de imagens dos conjuntos de dados.

Contamos também o número de quadros dos diferentes níveis de não-representatividade das imagens (Muito Próxima (MP), Próxima (P) e Distante (D) contidas em nossos conjuntos de dados, ou seja, as quantidades dos tipos de quadro MP, P e D. Essas quantidades são apresentadas na Tabela 3.2.

	Categoria da imagem			
Conjunto de dados	MP	P	D	Total
Wedding	2.272	16.165	13.157	31.594
Fire	182	20.619	23.306	44.107
Bombing	1.669	14.430	29.125	45.224

Tabela 3.2: Quantidade de imagens dos conjuntos de imagens Não-Representativas.

Para as técnicas supervisionadas que requerem a separação de subconjuntos de treinamento, validação e teste (vide Capítulo 5), incluímos quadros do mesmo vídeo no mesmo subconjunto, de forma a evitar contaminação no treinamento.

Capítulo 4

Espaço Semântico de Evento

Sistemas de Recuperação de Imagem Baseada em Conteúdo (*Content-Based Image Retrieval* – CBIR) [78] podem ser alternativas para recuperar imagens relevantes utilizando imagens de consulta dadas. No entanto, realizar essa tarefa com representações convencionais, i.e., as que utilizam características globais e locais de baixo nível (e.g., textura, cor, forma) [2, 31, 32], podem não apresententar resultados satisfatórios.

Para ilustrar esse cenário, considere as imagens apresentadas na Figura 4.1. Usando a Figura 4.1(a) como consulta para um método de CBIR que usa uma representação de imagem convencional, a Figura 4.1(c) é considerada a mais similar, enquanto que usando a Figura 4.1(b) como consulta, a Figura 4.1(d) é a mais similar. Entretanto, as figuras 4.1(a) e 4.1(b) são do mesmo evento, o que significa que a representação deu mais importância às similaridades dos padrões visuais do que aos aspectos semânticos. Outra solução possível para entender e separar imagens baseada em aspectos semânticos pode ser o treinamento de classificadores para a tarefa. No entanto, a necessidade de dados rotulados pode ser um obstáculo dada a frequência dos eventos e a necessidade de respostas rápidas a eles.

Hipótese 1: *É possível realizar uma melhor separação de imagens representativas e não-representativas para um evento quando as decompomos em componentes representativos.*

Nós propomos aqui um método de representação semântica de imagem para entender eventos em uma dimensionalidade reduzida. Nosso método, chamado Espaço Semântico de Evento (*Event Semantic Space* – ESS), usa um pequeno conjunto de Imagens Representativas para o Evento (*Event Representative Images* – ERIs) e um conjunto de Componentes Representativos (*Representative Components* – RCs), associados ao evento, para obter uma representação de características das imagens. As ERIs são imagens previamente selecionadas com alta representatividade para o evento. Os RCs têm como objetivo codificar informações necessárias para descrever a semântica do evento como pessoas que participaram do evento (e.g., suspeitos ou vítimas); objetos que apareceram no evento (e.g., carros, armas ou mochilas); e o lugar onde o evento ocorreu (e.g., parque, estádio ou prédio).

As características relacionadas com cada representação de imagem são dadas pela distância computada entre as imagens e as ERIs com relação a cada RCs. Portanto, cada representação de imagem é um decomposição levando em consideração sua similaridade

(a) Maratona de Boston de 2013

(b) Maratona de Boston de 2013

(c) Maratona de Austin de 2019

(d) Maratona de Boston de 2018

Figura 4.1: Representações de imagem convencionais consideram as similaridades de padrões visuais mais que os aspectos semânticos do evento. Usando a figura (a) como consulta, a figura (c) é a mais similar, enquanto que utilizando a figura (b) como consulta, a figura (d) é a mais similar. As figuras (a) e (b) deveriam ser as mais similares entre si já que pertencem ao mesmo evento, a Maratona de Boston de 2013.

com as ERIs e seu nível de pertinência de acordo com os RCs. ESS fornece uma meta-representação que utiliza representações convencionais de RCs para descrever imagens. A ideia é explicitamente explorar essas representações usando os RCs como guias para o entendimento do evento.

A criação da representação ESS tem como objetivo ajudar a entender a complexidade de eventos. Enquanto diferentes abordagens podem ser utilizadas para determinar imagens representativas incluindo métodos de recuperação e classificação, nenhuma delas é ideal para retornar a confiança necessária nos resultados (como podemos observar na Seção 4.3). O ESS pode melhorar o entendimento de eventos complexos utilizando características que correspondem a distâncias em cada dimensão de RC.

Os casos em que o ESS pode ser útil para o entendimento de evento incluem seu uso como método guiado para agregação de RCs conhecidos dos eventos para recuperação de imagens representativas, e, como uma possível maneira de melhorar o treinamento de métodos supervisionados quando possuímos um conjunto de treinamento pequeno, dada a reduzida dimensionalidade da representação ESS.

4.1 Construção do Espaço Semântico de Evento

O raciocínio por trás do ESS é o seguinte: a combinação de decompor o evento em RCs e usar o conhecimento proporcionado pelas ERIs resulta em uma representação semân-

tica de imagem capaz de separar imagens representativas das não-representativas. Uma exemplificação das etapas do ESS para extrair uma representação 3-dimensional é apresentada na Figura 4.2. Na sequência apresentamos como construir uma representação de características de maior ordem usando o ESS.

4.1.1 Componentes Representativos (RCs)

Primeiro, um conjunto $\mathcal{C} = \{c_k\}_{k=1}^{m}$ de RCs é definido. Esses componentes devem ser escolhidos baseando-se na natureza do evento de interesse. Por exemplo, lugares onde um evento ocorreu (e.g., parque, estádio ou prédio), objetos que compoem a cena do evento (e.g., arma, carro ou mochila), e atores envolvidos no evento (e.g., vítimas ou possíveis suspeitos).

4.1.2 Extração de Características

Depois de escolhidos os RCs que melhor descrevem o evento, as imagens do conjunto de dados precisam ser representadas como vetores de características (Figura 4.2(b)). A partir das imagens representadas, algumas imagens determinadas como altamente representativas para o evento são selecionadas para tornarem-se ERIs. As ERIs devem ser escolhidas por alguém com conhecimento prévio do evento, um especialista.

Seja $\mathcal{X} = \{\mathbf{I}_i\}_{i=1}^{n}$ um conjunto de n imagens; seja $\mathcal{Q} = \{\mathbf{q}_j\}_{j=1}^{\ell}$ um conjunto de ERIs, onde $\mathcal{Q} \subseteq \mathcal{X}$; e $f_k : \mathcal{X} \to \mathbb{R}^{\mathcal{D}}$ o *k-ésimo* extrator de características que mapeia as imagens em um espaço de características $\mathcal{D}$-dimensional. Note que $\mathcal{D}$ — a dimensão do vetor de características — não precisa ser igual para todos os m RCs. Cada imagem $\mathbf{I}_i$ é representada em termos dos m RCs. Portanto, haverá m vetores de características por imagem usando o extrator de características f_k. Uma propriedade importante do nosso método é o fato de que qualquer extrator de características pode ser utilizado contanto que este descreva apropriadamente o RC de interesse.

4.1.3 Distâncias

Usando os vetores de características, as distâncias são computadas entre cada ERI $\mathbf{q}_j$ e as imagens $\mathbf{I}_i$ usando uma métrica $\delta(f_k(\mathbf{q}_j), f_k(\mathbf{I}_i)) = d_{ij}^k$, onde δ pode ser qualquer métrica válida (e.g., distância Euclidiana).

Dessa maneira, d_{ij}^k é a distância entre $\mathbf{q}_j$ (a *j-ésima* ERI) e $\mathbf{I}_i$ (a *i-ésima* imagem) usando o extrator de características f_k considerando o *k-ésimo* RC. O resultado desse proceesso é a obtenção de m conjuntos de distâncias para cada ERI, como ilustrado na Figura 4.2(c). Note que, como as ERIs pertencem ao conjunto de dados ($\mathcal{Q} \subseteq \mathcal{X}$), depois de computados os conjuntos de distâncias, elas são as imagens mais similares em seus respectivos conjuntos.

4.1.4 Representação de Caraterísticas ESS

Para a construção do ESS, cada distância d_{ij}^k associada a uma imagem $\mathbf{I}_i$ é mapeada para uma coordenada no eixo e_j^k do espaço, onde j representa a *j-ésima* ERI, e k o *k-ésimo* RC.

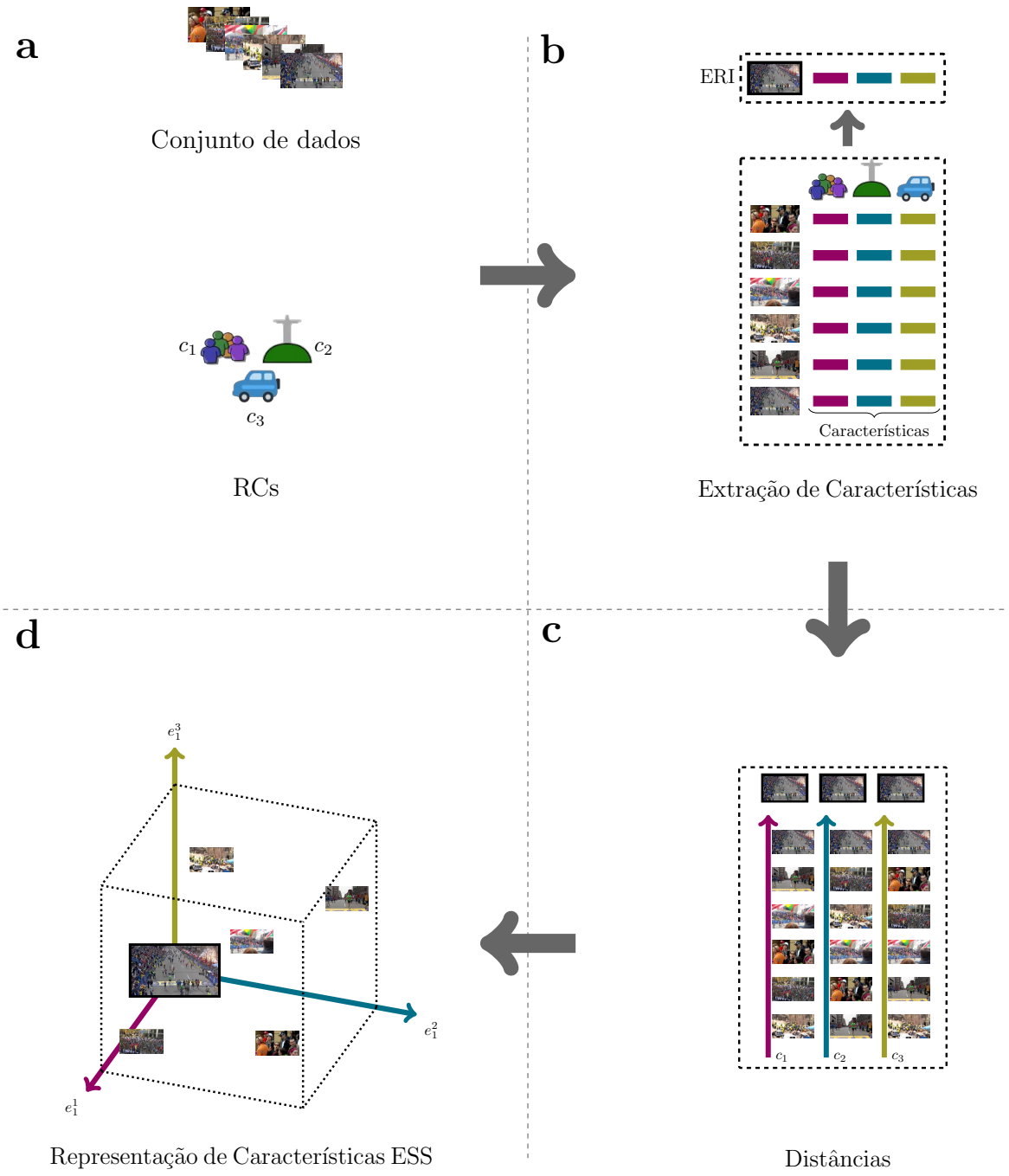


Figura 4.2: Um exemplo da construção de uma representação 3D de características usando ESS: (a) Componentes Representativos (RCs) c_1, c_2, and c_3 são escolhidos de acordo com o evento de interesse; (b) extração de características é realizada para cada imagem do conjunto de dados e para a Imagem Representativa do Evento (ERI) de acordo com c_1, c_2, and c_3; (c) as distâncias entre a ERI e as imagens são calculadas, para c_1, c_2, and c_3; e (d) essas distâncias determinam as coordenadas de cada representação de imagem no ESS. Um espaço de maior ordem pode ser obtido utilizando mais ERIs e/ou RCs.

Codificando todas as distâncias de acordo com a imagem $\mathbf{I}_i$ resulta no vetor de coordenadas $(d_{i1}^1, d_{i2}^1, \ldots, d_{i\ell}^1, d_{i1}^2, d_{i2}^2, \ldots, d_{i\ell}^2, \ldots, d_{i1}^m, d_{i2}^m, \ldots, d_{i\ell}^m)$. Este vetor é uma representação de características da imagem $\mathbf{I}_i$ no Espaço Semântico do Evento. A dimensão do espaço é

dada pelo número de RCs e ERIs. Logo, para m RCs e ℓ ERIs, a representação ESS é um vetor de características $m \times \ell$-dimensional.

Como as coordenadas são obtidas pelo cálculo das distâncias entre as ERIs e as imagens, a representação ESS faz com que imagens mais similares a uma ERI estejam situadas próximas a ela, o que significa que todas as imagens próximas a uma ERI têm uma quantidade similar de representatividade do evento. Se uma imagem nunca vista $\mathbf{I}_p$ precisar ser testada, sua representatividade é calculada computando a distância d_{pj}^k entre ela e cada $\mathbf{q}_j$ usando um vetor de características f_k, sem qualquer necessidade de recalcular todo o processo para o conjunto de dados inteiro. Logo, a representação ESS de $\mathbf{I}_p$ será $(d_{p1}^1, d_{p2}^1, \ldots, d_{p\ell}^1, d_{p1}^2, d_{p2}^2, \ldots, d_{p\ell}^2, \ldots, d_{p1}^m, d_{p2}^m, \ldots, d_{p\ell}^m)$.

4.1.5 Exemplo de Representação ESS

Vamos apresentar um exemplo prático da construção de uma representação ESS obtida com três RCs e duas ERIs, para seis imagens.

a) RCs: inicialmente, três RCs (c_1 = pessoas, c_2 = lugares e c_3 = objetos) são escolhidos para representar o conjunto de dados que possuímos para o evento (Figura 4.3). Nesse caso, nós queremos representar seis imagens do conjunto de dados.

Conjunto de dados RCs

Figura 4.3: Os RCs (c_1 = pessoas, c_2 = lugares e c_3 = objetos) são escolhidos com base no evento e em quais características queremos representar.

b) Extração de Características: após a escolha dos RCs, precisamos encontrar descritores f_k (com k sendo o k-ésimo RC) para representar os dados, de acordo com cada RC individualmente. Cada descritor de RC obterá um conjunto de características para cada imagem. A Figura 4.4 apresenta vetores de características 5-dimensionais para cada RC. Também precisamos representar as ERIs. Em nosso exemplo, as duas ERIs utilizadas também fazem parte do conjunto de dados (o que não é um requisito). Na Figura 4.4 cada linha dos quadros contém as representações (para a imagem indicada mais a esquerda) correspondentes aos três RCs. As cores — rosa, azul e verde — indicam características por RC.

c) Distâncias: com as características obtidas, calculamos as distâncias δ de cada imagem para cada ERI, considerando as características de cada RC separadamente. A distância utilizada neste exemplo foi a Euclidiana. Os valores das distâncias são apresentados na Figura 4.5.

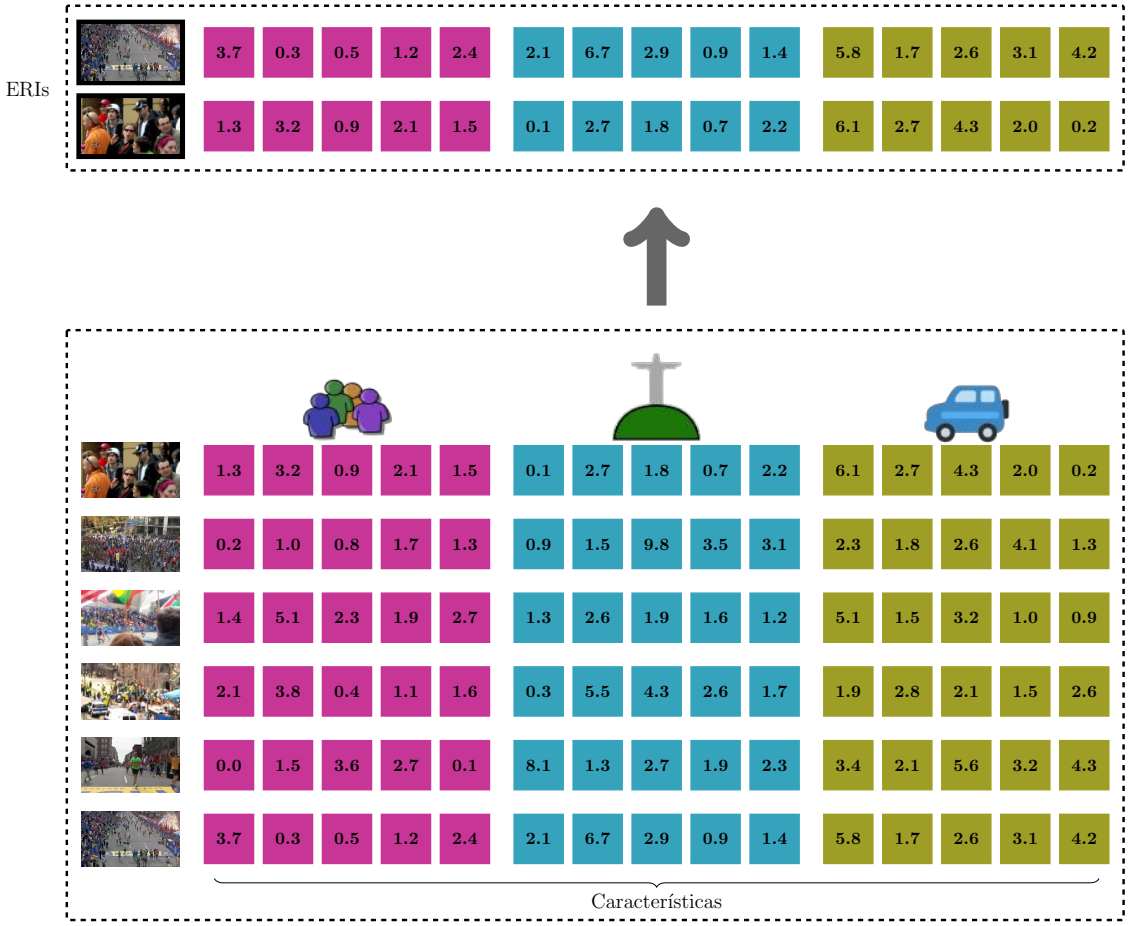


Figura 4.4: Nesse exemplo, cada um dos RCs gera um vetor de características 5-dimensional para cada uma das seis imagens. Cada cor representa vetores de características para cada RC: c_1 = rosa, c_2 = azul e c_3 = verde. Nesse caso, as duas ERIs foram obtidas do conjunto de dados (não é um requisito).

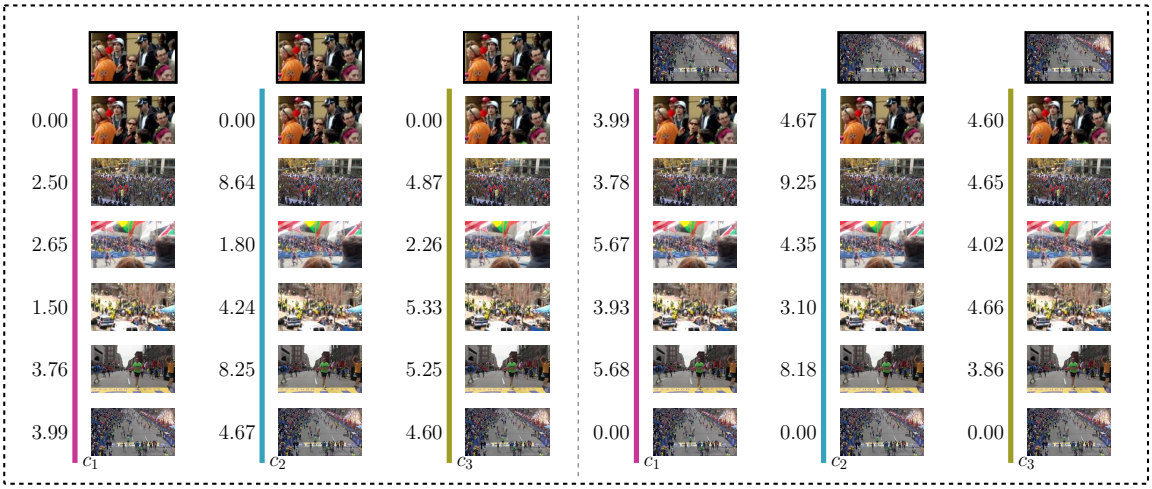


Figura 4.5: Valores de Distância Euclidiana calculados para cada ERI (indicada na primeira linha) e cada imagem, considerando as características obtidas por cada RC (c_1 = rosa, c_2 = azul e c_3 = verde).

d) Representação de Características ESS: a construção da representação ESS para uma imagem é feita utilizando todos os valores de distância (utilizando as características por RC individualmente) dessa imagem para as duas ERIs. A Figura 4.6 apresenta o vetor de características ESS para cada uma das seis imagens.

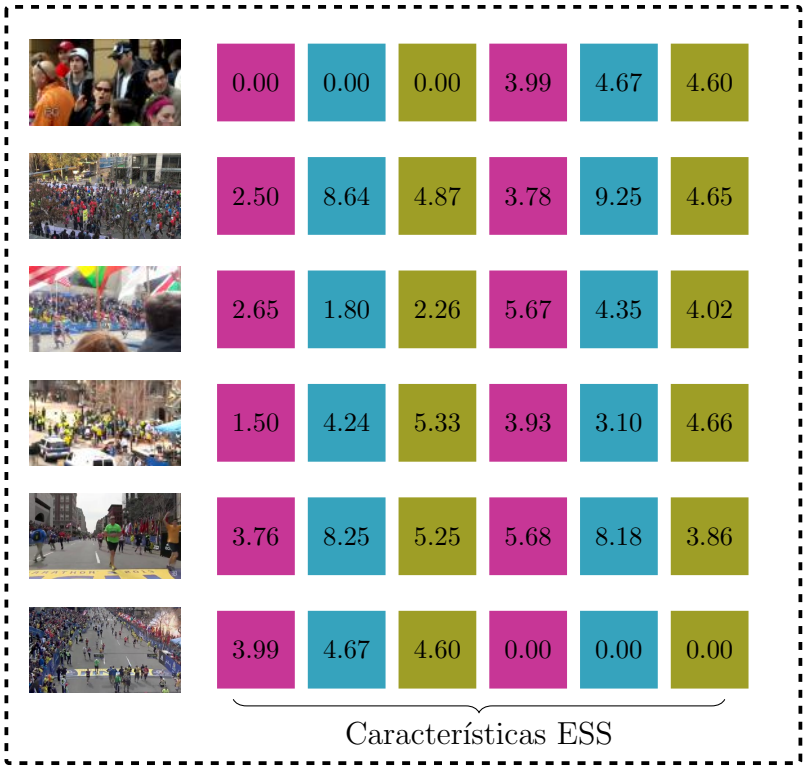


Figura 4.6: Utilizando as distâncias obtidas, construímos a representação ESS que é 6-dimensional ($\ell \times m$), já que inclui as distâncias da imagem para duas ERIs (ℓ) considerando três RCs (m).

O conjunto inicial de todas as características para representação de uma imagem possuía dimensionalidade igual a 15 (Figura 4.4). Quando construímos a representação ESS reduzimos a dimensionalidade para 6 (Figura 4.6). Ressaltamos ainda que, a flexibilidade do método permite que seja realizada a redução de dimensionalidade ao utilizarmos poucas ERIs, mas também possibilita o aumento de informações, caso seja necessário, incluindo novas ERIs e/ou RCs.

4.2 Experimentos

Três experimentos foram realizados para avaliar o método ESS. Para evitar vieses nos resultados, 100 conjuntos de 20 ERIs foram randomicamente gerados e cada experimento foi realizado para todos os conjuntos. Os resultados apresentados são a média obtida. Antes de descrever cada experimento, nós apresentaremos os extratores de características utilizados.

4.2.1 Extratores de Características

Três RCs foram escolhidos para descrever os eventos: lugares, objetos e pessoas. Para cada um dos RCs, Redes Neurais Convolucionais (*Convolutional Neural Networks* – CNNs) pré-treinadas foram utilizadas para representar as imagens dos conjuntos de dados apresentados na Seção 3.

Para compor o vetor de características relacionadas a lugares ($\mathbf{x}_{places}$), duas CNNs foram utilizadas: Inception-ResNet [68] treinada no conjunto de dados Imagenet [52] e VGG16 [67] treinada no conjunto de dados Places [85]. O vetor de características final foi obtido concatenando as saídas das duas CNNs, resultando em um vetor 5.632-dimensional. Os vetores de características relacionados a objetos ($\mathbf{x}_{objects}$) foram obtidos utilizando a Inception-Resnet [68] treinada no conjunto de dados Imagenet [52]. O vetor final de características foi um vetor 1.536-dimensional. Finalmente, o vetor de características relacionadas a pessoas ($\mathbf{x}_{people}$) foi obtido utilizando a rede de Re-identificação (*Re-Identification*) PCB [65], que foi treinada no conjunto de dados Market-1501 [82]. O vetor final de características foi um vetor 12.288-dimensional.

Para todas as CNNs, a saída da última camada antes da *Softmax* foi considerada o vetor de características. Todos os vetores de características foram escalados utilizando normalização *Z-score* para evitar incompatibilidades de escala.

Os vetores de características extraídos foram utilizados para compor as representações de imagens. A representação de características ESS foi construída utilizando os vetores $\mathbf{x}_{places}, \mathbf{x}_{objects}$, e $\mathbf{x}_{people}$. A representação de lugares (*Places*) utilizou somente $\mathbf{x}_{places}$. A representação de objetos (*Objects*) utilizou somente $\mathbf{x}_{objects}$. A representação de pessoas (*People*) utilizou somente $\mathbf{x}_{people}$. A representação concatenada (*Concatenated*) utilizou a concatenação de $\mathbf{x}_{places}$, $\mathbf{x}_{objects}$, e $\mathbf{x}_{people}$.

4.2.2 Recuperação de Imagens Representativas

Nesse experimento, nós testamos o desempenho da representação de características ESS para recuperar imagens representativas. Tanto as imagens do conjunto de dados quanto as ERIs foram representadas por vetores de características. Em seguida, para cada ERI, um *ranking* foi construído medindo a distância Euclidiana entre ela e todas as imagens do conjunto de dados. Finalmente, como utilizamos mais de uma ERI para cada conjunto de dados, os *rankings* para cada ERI foram agregados em um *ranking* final utilizando o método CombMin [59].

O ESS foi comparado com cinco representações para cálculo de *rankings*: *i*) *Places*; *ii*) *Objects*; *iii*) *People*; *iv*) *Concatenated*; e *v*) *Aggregated*, que agrega os *rankings* construídos com $\mathbf{x}_{places}$, $\mathbf{x}_{objects}$, e $\mathbf{x}_{people}$ utilizando o método CombSum [59]. A eficácia do experimento de recuperação foi medida utilizando as métricas de Precisão e Revocação.

4.2.3 Organização da Representação de Imagem

Para avaliar a discriminabilidade das características ESS no seu espaço original, o algoritmo de agrupamento k-means [35] foi empregado. Para realizar a comparação com a representação ESS, o algoritmo k-means também foi aplicado para as representações: *i*)

Places; *ii*) *Objects*; *iii*) *People*; e *iv*) *Concatenated*. Acurácia Balanceada foi utilizada para medir a eficácia do k-means já que o número de imagens representativas é (bem) menor que o número de imagens não-representativas.

Os centroides do k-means foram inicializados com dois conjuntos de imagens randomicamente selecionadas agindo como: *i*) imagens representativas e *ii*) imagens não-representativas, ambos na mesma quantidade de ERIs. Finalmente, as imagens foram consideradas pertencentes ao conjunto de imagens representativas se elas aparecessem em agrupamentos com pelo menos uma das imagens representativas selecionadas como centroides, e foram consideradas não-representativas caso contrário.

4.2.4 Representação de Imagens para Classificação

Para analisar as possibilidades de uso das características ESS na tarefa de classificação, um classificador de Floresta Aleatória [17] foi empregado. A representação ESS foi comparada com as representações: *i*) *Places*; *ii*) *Objects*; *iii*) *People*; *iv*) *Concatenated*; e *v*) *Aggregated*.

O conjunto de treinamento foi composto utilizando os *rankings* obtidos no experimento descrito na Seção 4.2.2 para cada representação. Para a classe positiva, utilizamos os vetores de características das imagens representativas indicadas pelo top@50 (primeiras 50 imagens do *ranking*). Para a classe negativa, utilizamos os vetores de características de 50 imagens não-representativas randomicamente obtidas da metade inferior do *ranking*. Como a abordagem *Aggregated* não possui um vetor único de características para representação (ela é composta pela agregação dos *rankings* obtidos por $\mathbf{x}_{places}$, $\mathbf{x}_{objects}$, e $\mathbf{x}_{people}$), o conjunto de treinamento foi gerado a partir do *ranking Aggregated* como o vetor de características da abordagem *Concatenated*. Florestas aleatórias com tamanhos $1, 50, 100, 150, 200, 250$ e 300 foram testadas. Os resultados foram avaliados utilizando Acurácia Balanceada das classes positiva e negativa (acurácia Representativa e Não-Representativa).

4.3 Resultados

Primeiramente, verificamos os resultados obtidos pela representação de características ESS na tarefa de recuperação de imagens representativas. A Figura 4.7 apresenta os resultados do ESS e dos métodos de referência com relação à curva de *Precisão* $\times$ *Top* K nos três conjuntos de dados. Nós notamos que o ESS atingiu a maior precisão quando comparado com os descritores individuais $\mathbf{x}_{places}$, $\mathbf{x}_{objects}$, e $\mathbf{x}_{people}$, e apresentou resultados similares às abordagens *Concatenated* e *Aggregated*.

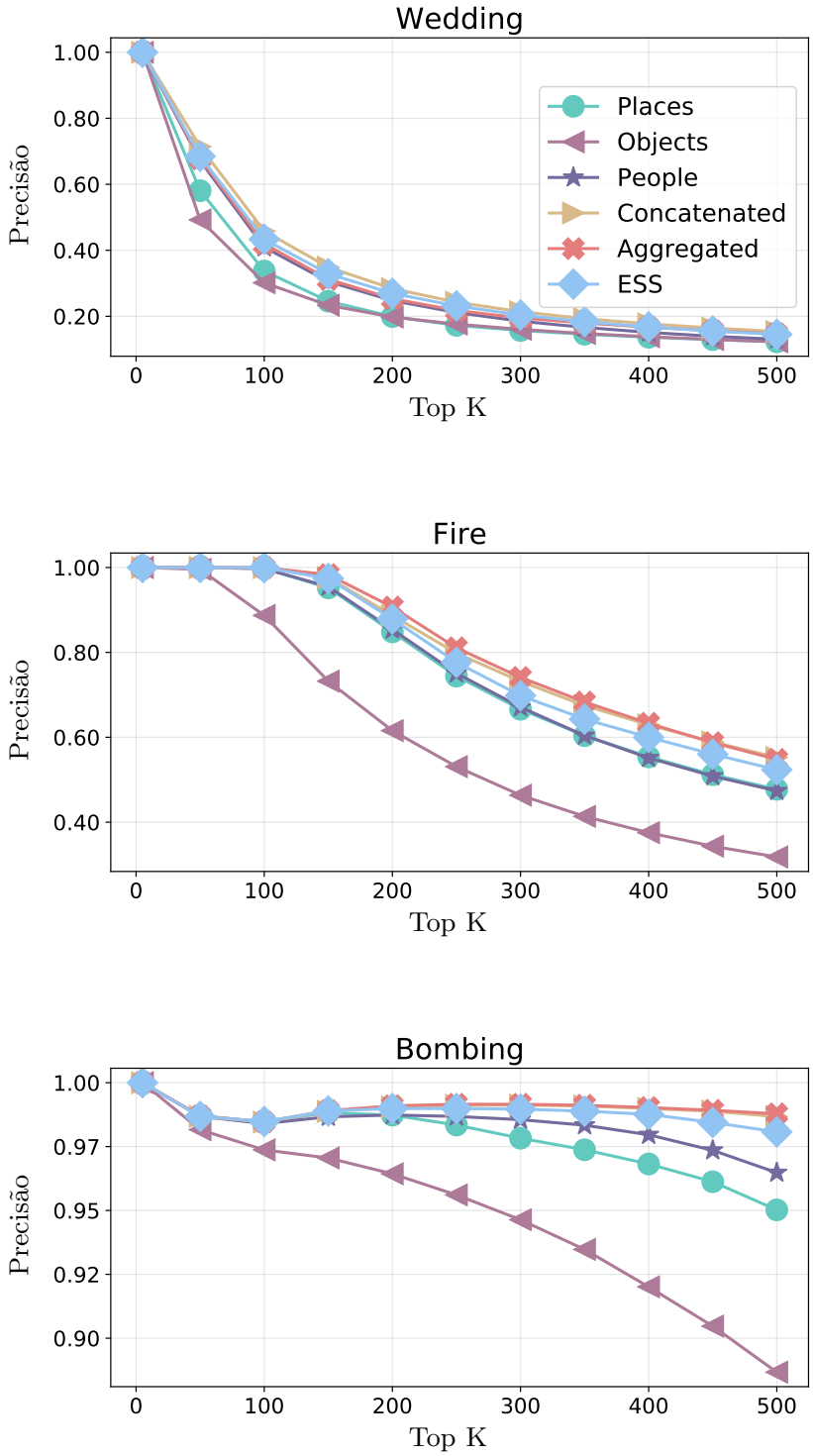


Figura 4.7: A representação de características ESS conduz a taxas mais altas de imagens representativas recuperadas do que os descritores individuais. De cima para baixo, a curva de *Precisão* × *Top K* das representações ESS, Places, Objects, People, Aggregated, e Concatenated.

Para avaliar a discriminabilidade das características ESS em seu espaço original, nós realizamos o agrupamento pelo algoritmo k-means. Na abordagem do k-means, a acurácia média (apresentada na Figura 4.8) demonstra que a representação *Concatenated* possui melhor desempenho do que as demais. Apesar de perder para as outras representações,

o ESS, que tem a menor dimensionalidade, apresenta um desempenho mais estável em todos os conjuntos de dados, atingindo por volta de 63% de acurácia.

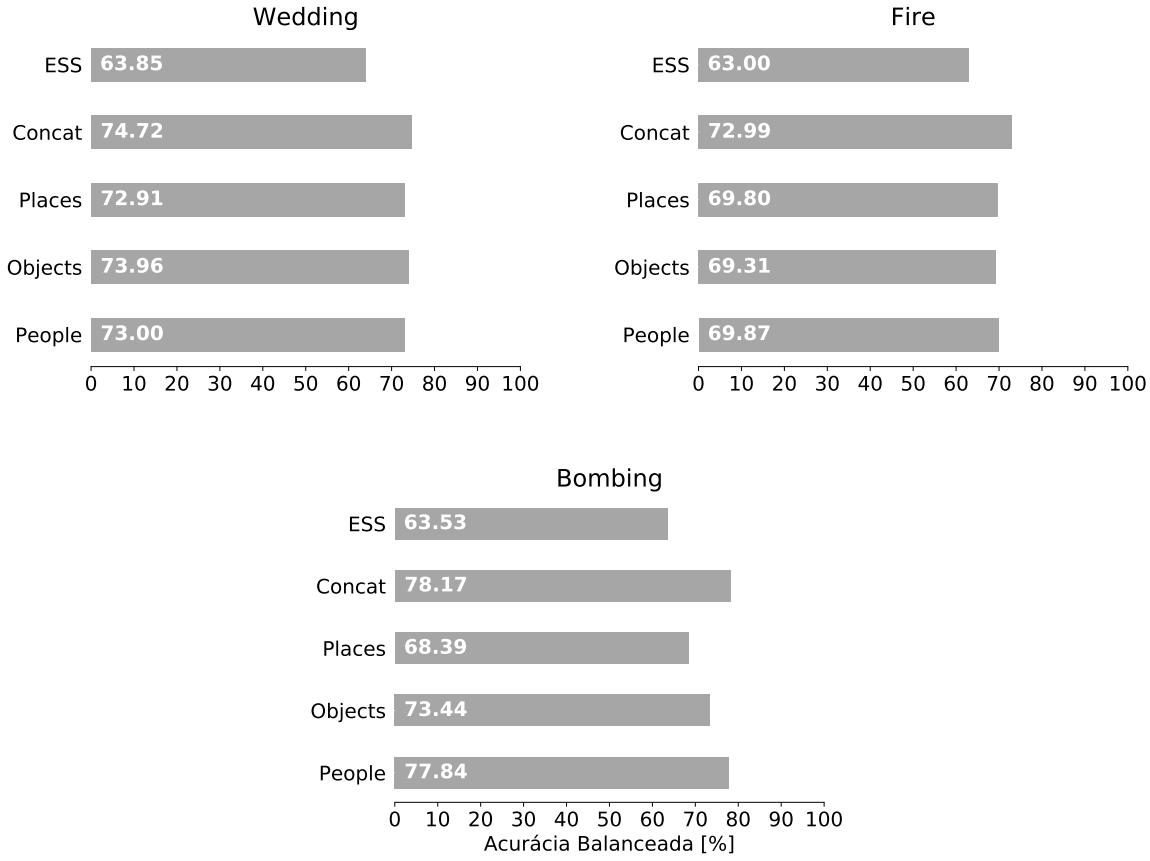


Figura 4.8: A representação ESS resulta em pior eficácia em comparação com as outras representações quando o agrupamento k-means é utilizado. De cima para baixo, a Acurácia Balanceada das representações ESS, *Concatenated* (Concat), *Places*, *Objects* e *People*.

Para analisar as possibilidaddes do uso de características ESS na tarefa de classificação, treinamos classificadores de Foresta Aleatória com múltiplos números de árvores. Nós notamos comportamentos diferentes por conjunto de dados. No conjunto *Wedding* (Figura 4.9), o ESS apresentou comportamento similar às abordagens *Concatenated* e *Aggregated*. No conjunto *Fire* (Figura 4.10), o ESS superou todas as representações. No conjunto *Bombing* (Figura 4.11), o ESS apresentou resultados similares aos resultados das representações *Places*, *People* e *Concatenated*. Em todos os conjuntos de dados, o ESS atingiu boas porcentagens de Acurácia Representativa obtendo os melhores valores dessa acurácia nos conjuntos de dados *Fire* e *Bombing*.

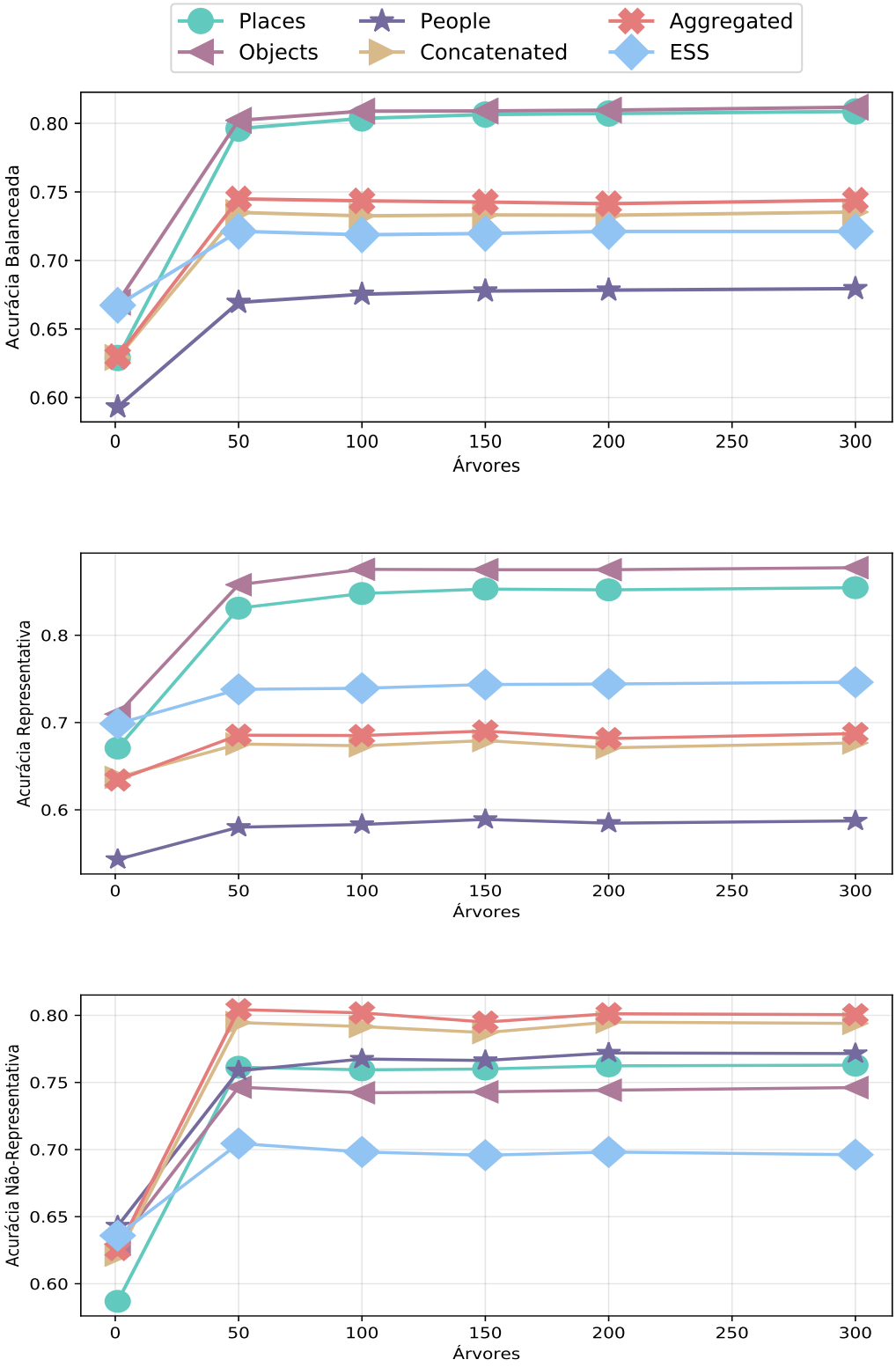


Figura 4.9: A representação de características ESS, no conjunto de dados *Wedding*, proporcionou maior Acurácia Balanceada em comparação com a representação *People* e maior Acurácia Representativa em comparação com *People*, *Concatenated* e *Aggregated*. Topo do gráfico: Acurácia Balanceada. Na sequência: acurácia considerando imagens Representativas, e acurácia considerando imagens Não-Representativas.

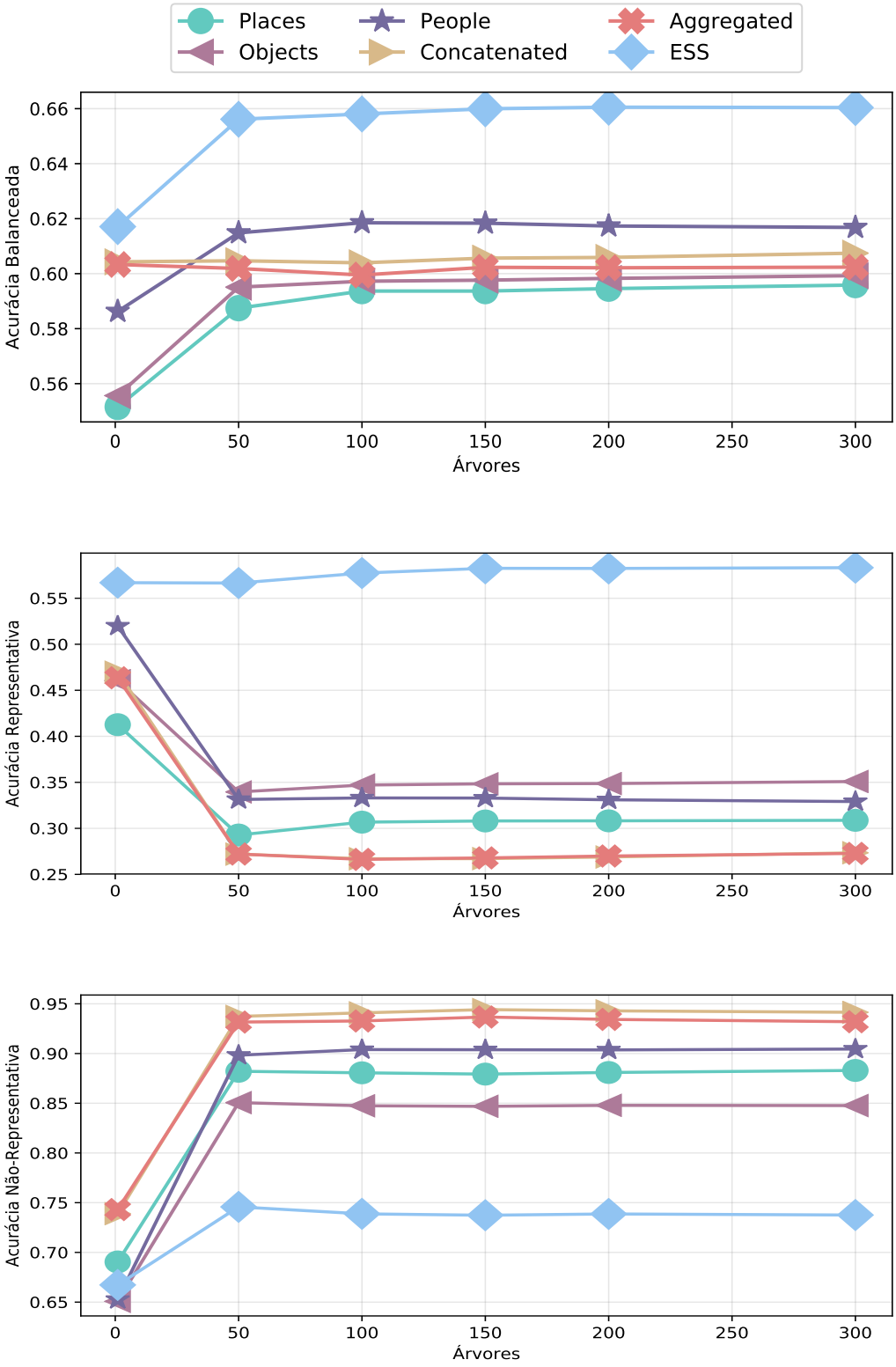


Figura 4.10: A representação de características ESS, no conjunto de dados *Fire*, proporcionou maior Acurácia Balanceada e Acurácia Representativa. Topo do gráfico: Acurácia Balanceada. Na sequência: acurácia considerando imagens Representativas, e acurácia considerando imagens Não-Representativas.

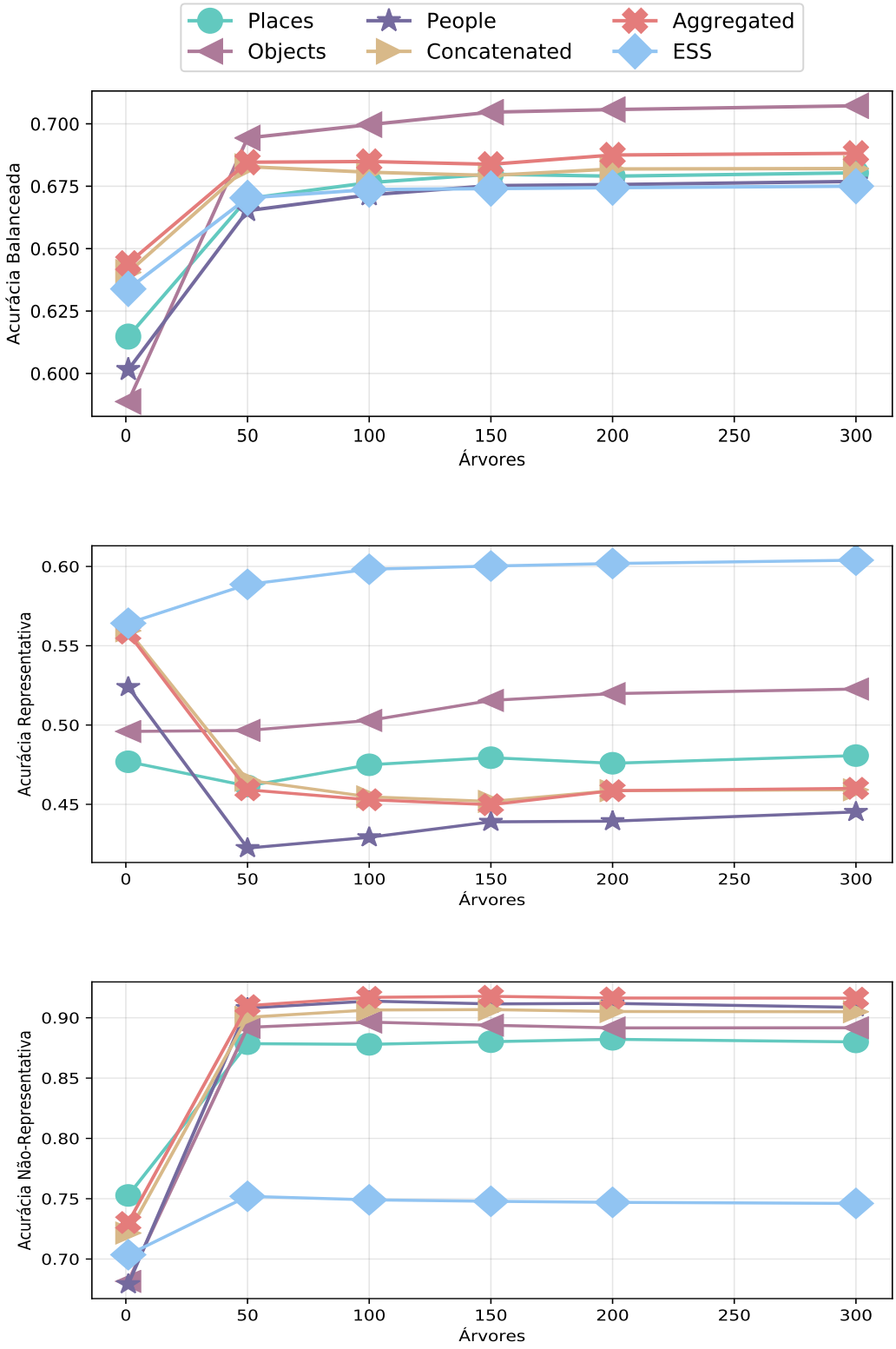


Figura 4.11: A representação de características ESS, no conjunto de dados *Bombing*, proporcionou maior Acurácia Representativa. Topo do gráfico: Acurácia Balanceada. Na sequência: acurácia considerando imagens Representativas, e acurácia considerando imagens Não-Representativas.

Vetores de Características	Dimensões
Places	5.632
Objects	1.536
People	12.288
Concatenated	19.456
Aggregated	5.632/1.536/12.288
ESS	60

Tabela 4.1: Número de dimensões dos vetores de características comparados ao vetor de características ESS de baixa dimensionalidade.

4.4 Discussões sobre o Espaço Semântico de Evento

A quantidade massiva de imagens disponíveis na internet pode ajudar o entendimento de eventos. No entanto, como separar automaticamente imagens representativas – as que ajudam esse entendimento — das não-representativas ainda é um problema em aberto. Nós apresentamos um método de representação semântica de imagem para auxiliar o entendimento de eventos em dimensionalidade reduzida. Nosso método procura capturar a semântica projetando todas as imagens em um espaço onde as coordenadas de uma representação de imagem são determinadas por Componentes Representativos de eventos e Imagens Representativas do Evento, resultando em uma meta-representação semântica de imagem.

Com os experimentos realizados confirmamos que representações convencionais não são suficientes para a tarefa de recuperação de imagens por representatividade (Figura 4.7), o que evidencia a complexidade da seleção de imagens representativas e da análise de eventos. Se tentarmos recuperar a maior parte das imagens representativas — que pode ser o objetivo do especialista tentando entender um evento —, a maioria das imagens obtidas são não-representativas, o que reduz a confiança destes métodos. Por outro lado, recuperar menos imagens para aumentar a confiança leva a uma possível perda de imagens representativas, contribuindo para a redução da interpretabilidade do evento. Esse problema pode ser explicado pela dificuldade encontrada pelas representações convencionais ao descreverem o conteúdo semântico de alto-nível em imagens usando características locais e globais de baixo nível (e.g., textura, cor e forma) [2, 31, 32]. Os resultados da recuperação de imagens utilizando representações tradicionais sugerem que estes métodos não são suficientes para determinar representatividade de imagens para o entendimento do evento e indicam a necessidade de métodos para realizar essa tarefa.

O ESS, por outro lado, lida com a semântica do evento decompondo-o em Componentes Representativos. Essa decomposição proporciona uma representação explícita de baixa dimensionalidade, conforme apresenta a Tabela 4.1, que pode indicar quais componentes contribuem mais para a análise do evento. Além disso, mesmo apresentando acurácias menores para a abordagem de agrupamento (Figura 4.8), a representação ESS provou seu valor para tarefas de recuperação e classificação, e pode ser considerada uma alternativa na tentativa de entender eventos complexos do mundo real. Durante a recuperação (Figura 4.7), mesmo com menor dimensionalidade, o ESS superou as representações *Places*,

Objects e *People*, compostas pelos descritores de Componentes Representativos individualmente. Durante a tarefa de classificação (figuras 4.9, 4.10 e 4.11), a dimensionalidade reduzida do ESS torna-se uma vantagem, já que o conjunto de treinamento é pequeno. O ESS demonstra a capacidade de proporcionar boa Acurácia Representativa, aprendendo o que é uma imagem representativa, e atingindo a melhor Acurácia Balanceada no conjunto de dados *Fire*.

Capítulo 5

Espaço Combinado de Evento

Como mostra o Capítulo 4 a decomposição de eventos em Componentes Representativos, de fato, melhora a tarefa de recuperação por representatividade (Figura 4.7). No entanto, o método anteriormente proposto (ESS), apesar da dimensionalidade reduzida, possui limitação de precisão, atingindo resultados similares à concatenação simples das características dos componentes (*Concatenated*).

Acreditamos que o fato das taxas de precisão (em especial do *Concatenated* e ESS) não atingirem resultados mais altos deve-se ao desconhecimento da contribuição de cada componente para eventos diferentes. Por isso, procuramos uma estratégia para aprender cada contribuição e a maneira de combinar as características, de forma a melhorar a qualidade da representação de imagens de um evento.

Nosso maior desafio é a falta de imagens anotadas para o processo de treinamento. Por isso, precisamos de estratégias que sejam capazes de aprender a combinar características de componentes de um evento com poucas imagens anotadas.

Hipótese 2: *É possível aprender um espaço de combinação de características de componentes (manifold) com pequenos conjuntos de treinamento de forma a melhorar a separação de imagens Representativas e Não-Representativas.*

Nós propomos um método para aprender a contribuição dos diferentes componentes do evento. Esse método chamado de Espaço Combinado de Evento (*Event Combined Space – ECS*) utiliza uma estratégia orientada a dados para aprender um *manifold* onde imagens Representativas são posicionadas juntas distanciando-se das Não-Representativas. Uma contribuição essencial da nossa metodologia é a possibilidade de usarmos conhecimento prévio de outros domínios com conjuntos de dados grandes, apenas aprendendo como combinar esse conhecimento. Exploramos três funções de perda diferentes usando a rede de combinação proposta para a tarefa de recuperação de imagens Representativas.

5.1 Construção do Espaço Combinado de Evento

Dado o evento $\mathcal{E}$, que queremos representar e entender, e n imagens disponíveis, a tarefa que queremos ser capazes de realizar é a separação das imagens Representativas e das Não-Representativas.

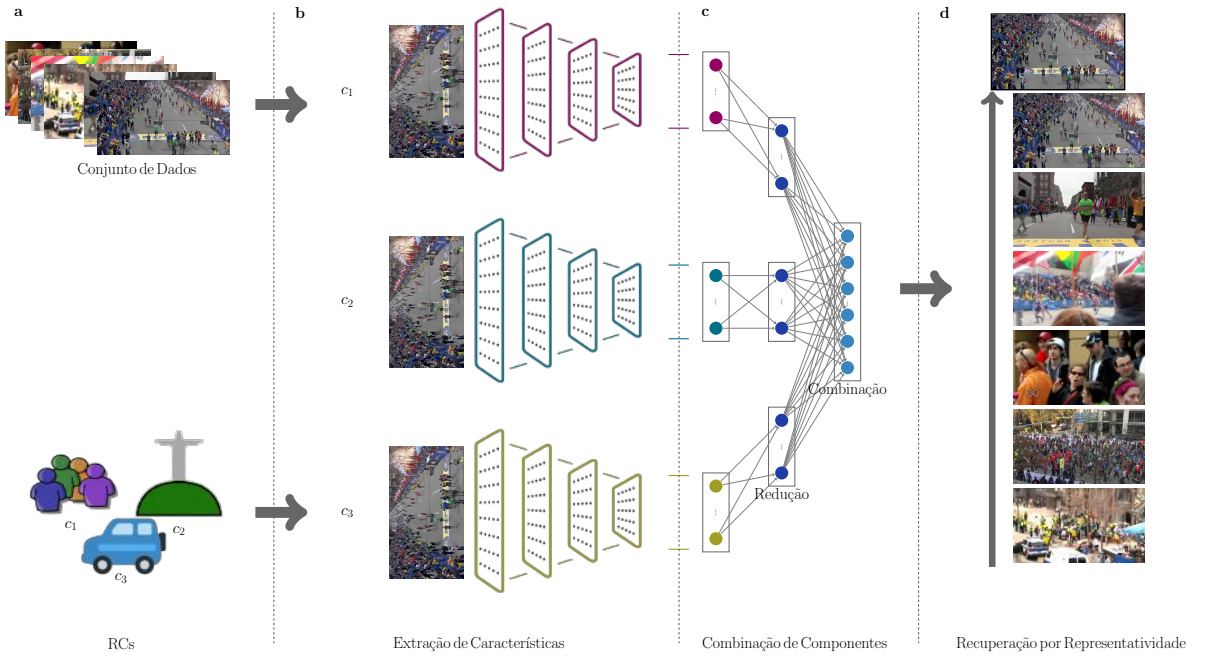


Figura 5.1: Processo de combinação de componentes por aprendizado de *manifold*. (a) Selecionamos cada uma dos Componentes Representativos (RCs) c_1, c_2, e c_3 com base no conjunto de dados do eventos de interesse para a realização da decomposição. (b) Utilizamos técnicas de transferência de aprendizado para descrever inicialmente as imagens. Isso pode ser um processo de treinamento independente (*Standalone*) usando CNNs pré-treinadas ou um processo de treinamento conjunto (*Joint*) com a próxima etapa do processo, envolvendo a combinação dos componentes; (c) Combinamos as características extraídas para cada componentes por meio do aprendizado de *manifold*; (d) Por fim, as imagens Representativas podem ser recuperadas (utilizando uma imagem Representativa de consulta) utilizando a representação aprendida.

Assim como no Capítulo 4 queremos explorar a decomposição de eventos em componentes, mas ao invés de realizar manualmente a combinação dos componentes, esperamos ser capazes de combiná-los a partir da técnica de aprendizado de *manifold*. Assim, encontraremos um espaço latente no qual os componentes possam ser ponderados e combinados dinamicamente.

5.1.1 Representação e Recuperação por Componentes

Para realizar a separação por representatividade, nosso método segue um processo composto por quatro etapas: (a) Definição dos Componentes Representativos; (b) Extração de Características; (c) Combinção de Características com base nos Componentes; e (d) Recuperação de Exemplos com base na Representatividade. A Figura 5.1 apresenta esse processo.

Componentes Representativos (RCs): A escolha dos Componentes Representativos deve refletir os pontos principais do evento, de forma a melhorar seu entendimento. Assim como apresentados no Capítulo 4, alguns dos componentes que podem ser importantes incluem: lugar do evento, objetos que aparecem na cena e pessoas que podem estar envolvidas. Esses componentes formam o conjunto $\mathcal{C} = \{c_k\}_{k=1}^{m}$ (Figura 5.1(a)).

Extração de Características: Depois de escolhidos os RCs, precisamos representar cada imagem com esses componentes. A extração de características pode ser realizada utilizando diferentes técnicas contanto que descrevam os RCs. Seja $\mathcal{X} = \{\mathbf{I}_i\}_{i=1}^{n}$ um conjunto de n imagens. Nós temos m extratores de características h para representar os m RCs, e o *k-ésimo* extrator de características obtêm a representação $\mathcal{D}$-dimensional (onde $\mathcal{D}$ depende do extrator de características adotado) para cada imagem $\mathbf{I}_i$ no conjunto de dados: $h_k : \mathcal{X} \rightarrow \mathbb{R}^{\mathcal{D}}$. Nós propomos o uso de CNNs como extratores de características h de acordo com a Figura 5.1(b).

Combinação de Componentes: Depois de obter as características para cada RC, nós queremos aprender um *manifold* para combiná-los apropriadamente (Figura 5.1(c)). Para isso, propomos uma rede densa, incluindo duas partes principais: *Redução* e *Combinação*. As camadas de *Entrada* da arquitetura têm tamanhos diferentes, de acordo com a saída da CNN utilizada como *k-ésimo* extrator de características. A primeira parte principal compreende uma camada de *Redução* para reduzir a dimensionalidade das características representando os m componentes para o mesmo comprimento de vetor. Cada vetor de características de componente tem sua sub-camada de *Redução* com tamanho igual a r neurônios, que não compartilham pesos com outra sub-camada de *Redução*. A razão por não compartilharem pesos é o fato de querermos reduzir cada representação dos componentes — de forma que a combinação não seja influenciada pelo tamanho dos vetores de características — evitando a influência dos componentes entre si.

A segunda parte principal é a *Combinação*. Essa parte compreende camadas $\mathcal{Y} = \{\mathbf{y}_p\}_{p=1}^{L}$, com a primeira camada de tamanho igual a $|\mathbf{y}_1| = r/2$ neurônios, reduzida pela metade em cada camada subsequente (i.e, $|\mathbf{y}_p| = r/2^p$). A parte de *Combinação* recebe todas as representações reduzidas dos componentes para serem combinadas (vide Seção 5.1.2). A saída é um vetor único que representa as imagens de um evento de forma mais compacta e efetiva. O processo de treinamento é dependente da função de perda adotada (vide Seção 5.1.3).

Recuperação por Representatividade: Considere o conjunto $\mathcal{X} = \{\mathbf{I}_i\}_{i=1}^{n}$ de n imagens, que precisamos separar como Representativas ou Não-Representativas para o evento $\mathcal{E}$, e $\mathcal{Q} = \{\mathbf{q}_j\}_{j=1}^{\ell}$ o conjunto de imagens Representativas usadas como consultas. Para recuperar imagens Representativas de $\mathcal{X}$, procuramos obter a imagem menos distante (mais similar) a uma consulta $\mathbf{q}_j$. Para obter essas distâncias (dissimilaridades), nós precisamos definir uma função descritora f e uma métrica válida δ (e.g., distância Euclidiana), de forma que $\delta(f(\mathbf{q}_j), f(\mathbf{I}_i)) = d_{ij}$ onde d_{ij} representa um valor de distância entre as representações de $\mathbf{q}_j$ e $\mathbf{I}_i$. Um *ranking* final $\mathcal{R}_{\mathbf{q}_j}$ é uma permutação das imagens em $\mathcal{X}$, conforme apresentado na Seção 2.8 (Figura 5.1(d)). Neste capítulo, a função de descrição f é obtida pela etapa de *Combinação de Componentes* (Figura 5.1(c)).

5.1.2 Aprendizado de *Manifold*

Nossa metodologia realiza o aprendizado de um *manifold* para combinação de características de diferentes RCs. Podemos realizar esse processo de duas maneiras diferentes: usando um processo independente (*Standalone*) de combinação de características, ou usando um processo de otimização conjunta (*Joint*) para extração e combinação. Ambas as aborda-

gens tem pontos positivos e negativos.

Quando utilizamos o processo de treinamento independente (*Standalone*), extraímos características (Figura 5.1(b)) usamos transferência de aprendizado de CNNs pré-treinadas para tarefas de visão computacional (geralmente com centenas ou milhares de exemplos de treinamento) e, posteriormente usamos esses vetores de características para treinar (independentemente) a rede de combinação de componentes (Figura 5.1(c)) para aprender o *manifold*. A principal vantagem dessa abordagem é a necessidade de menos dados de treinamento para treinar a rede de combinação e para refinar a arquitetura no geral, já que lidamos com um número muito menor de parâmetros. Além disso, podemos escolher redes para extração de características de imagem que foram treinadas com grandes conjuntos de dados, contanto que descrevam razoavelmente os componentes adotados. Por outro lado, a principal desvantagem é a falta de especialização dos extratores de características de imagem para o evento de interesse.

Quanto utilizamos o processo de otimização conjunta (*Joint*), podemos treinar os extratores de características (CNNs) (Figura 5.1(b)) junto com a rede de combinação de componentes (Figura 5.1(c)). Dessa maneira, o processo de aprendizado de características e combinação de componentes é unificado. A principal vantagem dessa abordagem é a especialização dos extratores de características para os eventos de interesse. Nesse caso, o objetivo é obter o extrator que melhor represente os RCs de cada evento. No entanto, nessa abordagem, existe um número significantemente maior de parâmetros para treinamento, consequentemente, é necessário aumentar o número de imagens de treinamento.

Como a falta de imagens anotadas é uma característica intrínseca do nosso problema, optamos pela abordagem de treinamento independente, usando transferência de aprendizado de CNNs pré-treinadas como extratores de características. Nesse contexto, nós também precisamos definir os tamanhos das partes de *Redução* e *Combinação* da rede densa considerando apenas um conjunto pequeno de amostras de treinamento para cada evento.

5.1.3 Função de Perda

Propomos o aprendizado de *manifold* de duas maneiras diferentes, por aprendizado de classificação ou de distância. Para o aprendizado de classificação, nós utilizamos a função de perda Entropia Cruzada. Para aprendizado de distância, utilizamos as funções de perda Contrastiva e Tripla (vide Seção 2.6.2).

Entropia Cruzada: A primeira função de perda realiza a combinação apresentada na Figura 5.1(c) como um classificador de duas classes de acordo com a Figura 5.2: Representativa e Não-Representativa. Nesse caso, nós adotamos a função de perda Entropia Cruzada. Depois do treinamento, obtemos a representação da última camada de *Combinação*. Nós chamaos o espaço aprendido de espaço *Cross-Entropy*. Nós treinamos a rede por 50 épocas usando o otimizador Adam [26] com uma taxa de aprendizado $\alpha = 0,00001$ e uma taxa de decréscimo de $1e^{-5}$.

Contrastiva: A segunda função de perda adotada é a Contrastiva. Essa função de perda usa uma rede Siamesa utilizando o modelo base (Figura 5.1(c)) em cada ramo, seguido pelo cálculo da distância Euclidiana entre as representações obtidas pelos ra-

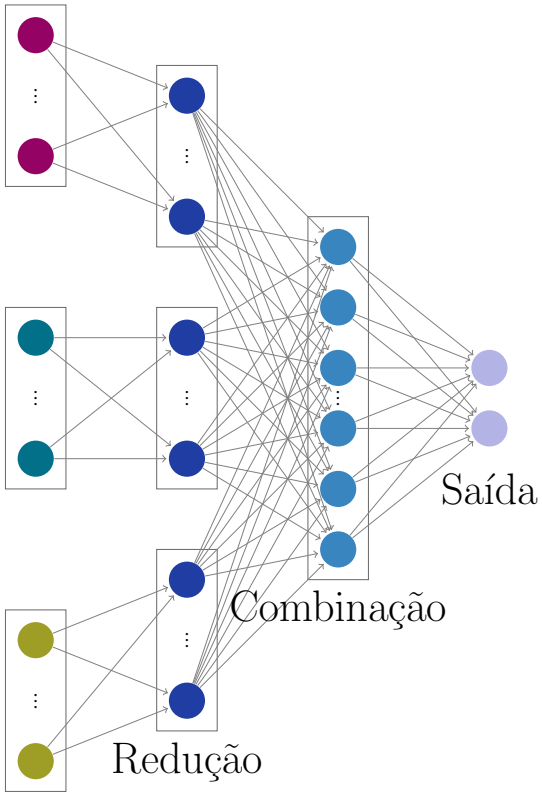


Figura 5.2: Rede para uma tarefa de classificação utilizando a função de perda Entropia Cruzada. A camada de *Entrada* recebe vetores de características com dimensionalidades diferentes, de acordo com as CNNs utilizadas para descrever os RCs; a camada de *Redução* reduz as características independentemente (sem compartilhamento de pesos entre os vetores dos componentes); as camadas de *Combinação* combinam as características reduzidas; a camada de *Saída* é um vetor 2-dimensional, classificando as imagens em Representativa ou Não-Representativa.

mos, como apresentado na Figura 5.3. Essa abordagem compara duas representações de imagem. Por essa razão, nós utilizamos dois conjuntos de vetores de entrada, um representando a imagem $\mathbf{I}_a$ e outro representando a imagem $\mathbf{I}_b$. Depois do treinamento, obtemos a representação da última camada de *Combinação*. Nós chamamos esse espaço aprendido de espaço *Contrastive*. Nós treinamos a rede por 50 épocas usando o otimizador Adam [26] com uma taxa de aprendizado $\alpha = 0,00001$ e uma taxa de decréscimo de $1e^{-5}$, usando uma margem igual a 1,0.

Para treinar a rede, adotamos *batches* de 128 pares. Duas imagens da mesma classe foram definidas como um par positivo, enquanto duas imagens de classes diferentes são um par negativo. Para cada imagem do conjunto de treinamento, geramos dez pares positivos e dez negativos.

Tripla: A terceira função de perda adotada consiste na função de perda tripla formulada utilizando o modelo base (Figura 5.1(c)) e uma camada de perda tripla no fim. Esse modelo utiliza três ramos para o treinamento, cada um com um grupo de características. O primeiro grupo representa uma imagem $\mathbf{I}_a$ (a imagem âncora para a tripla); o segundo grupo representa a imagem $\mathbf{I}_p$, que é uma imagem positiva para a âncora; e o terceiro grupo representa a imagem $\mathbf{I}_n$, que é uma imagem negativa para a âncora. Depois do

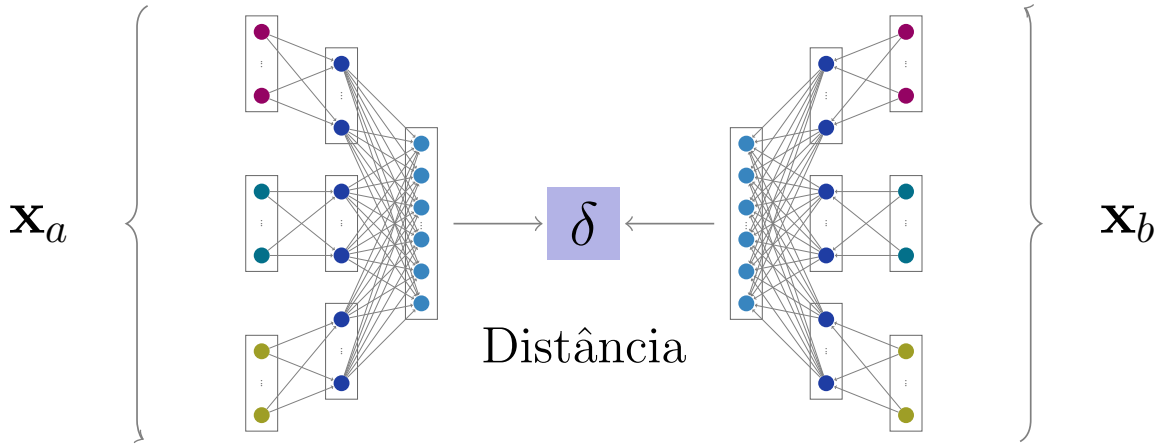


Figura 5.3: Rede para aprendizado de métricas de distância utilizando a função de perda Contrastiva. Para treinar a rede, nós utilizamos dois ramos com pesos compartilhados para o conjunto de características ($\mathbf{x}_a$ e $\mathbf{x}_b$) das imagens $\mathbf{I}_a$ e $\mathbf{I}_b$ com o objetivo de reduzir a distância entre características de imagens da mesma classe enquanto aumenta a distância entre características de imagens de classes diferentes. A camada de *Entrada* recebe vetores de características com dimensionalidades diferentes, de acordo com as CNNs utilizadas para descrever os RCs; a camada de *Redução* reduz as características independentemente (sem compartilhamento de pesos entre os vetores dos componentes); as camadas de *Combinação* combinam as características reduzidas. Depois da última camada de *Combinação*, temos a função de distância para a computação da função de perda.

treinamento, obtemos a representação da última camada de *Combinação*. Nós chamamos o espaço aprendido de espaço *Triplet*. Nós treinamos a rede por 50 épocas usando o otimizador Adam [26] com uma taxa de aprendizado $\alpha = 0,01$ e uma taxa de decréscimo de $1e^{-5}$, usando uma margem igual a 1,0.

Utilizamos *batches* de 128 triplas criadas considerando uma âncora, uma amostra positiva da mesma classe, e uma amostra negativa de classe diferente. A literatura de visão computacional sugere o uso de triplas comuns com triplas difíceis. Uma tripla comum é aleatoriamente construída respeitando a formação de tripla descrita. Uma tripla difícil consiste em exemplos positivos e negativos muito próximos no espaço de representação. Nós consideramos 64 triplas selecionadas aleatoriamente e 64 triplas difíceis. As triplas difíceis foram selecionadas a cada vez de um conjunto de 512 triplas aleatórias e ordenadas de acordo com o nível de confusão no espaço de características.

5.2 Experimentos

Para avaliar o método proposto, realizamos cinco experimentos considerando os três conjuntos de dados descritos no Capítulo 3. Para isso, dividimos os conjuntos de dados em subconjuntos de treinamento, validação e teste gerando *rankings* para análises quantitativas e qualitativas. Usamos o subconjunto de treinamento como conjunto de consultas para recuperar imagens do subconjunto de teste. Os resultados de Precisão são reportados utilizando a média dos *rankings* gerados. Antes de apresentar os detalhes de cada experimento, nós descrevemos: a divisão e aumentação dos conjuntos de dados (Seção 5.2.1), os

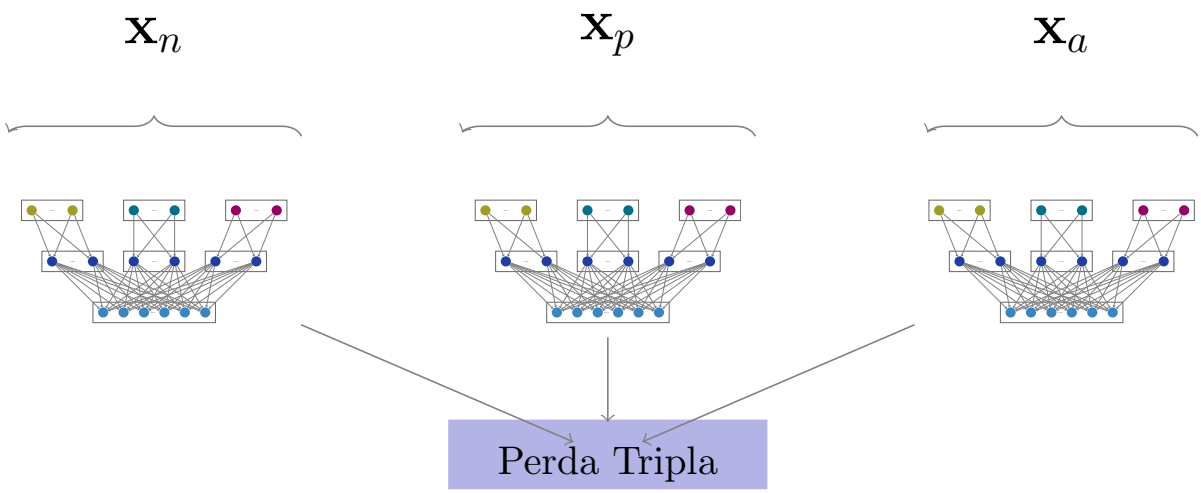


Figura 5.4: Rede para aprendizado de métricas de distância utilizando a formulação da função de perda Tripla. Para treinar a rede, nós usamos três ramos com compartilhamento de pesos para cada conjunto de características ($\mathbf{x}_a$, $\mathbf{x}_p$ e $\mathbf{x}_n$) das imagens $\mathbf{I}_a, \mathbf{I}_p$ and $\mathbf{I}_n$. O objetivo é reduzir a distância final entre características das imagens $\mathbf{I}_a$ e $\mathbf{I}_p$ e aumentar a distância das características das imagens $\mathbf{I}_a$ e $\mathbf{I}_n$. A camada de *Entrada* recebe vetores de características com dimensionalidades diferentes, de acordo com as CNNs utilizadas para descrever os RCs; a camada de *Redução* reduz as características independentemente (sem compartilhamento de pesos entre os vetores dos componentes); as camadas de *Combinação* combinam as características reduzidas. Depois da última camada de *Combinação*, temos a camada de perda Tripla para a computação da função de perda.

extratores de características utilizados (Seção 5.2.2), e o protocolo de treinamento para os espaços aprendidos (Seção 5.2.3).

5.2.1 Divisão e Aumentação dos Dados

Avaliamos o método proposto nos três conjuntos de dados apresentados no Capítulo 3: *Wedding*, *Fire* e *Bombing*. Cada conjunto de dados descreve um evento, incluindo imagens Representativas e Não-Representativas.

Dividimos cada conjunto de dados em três subconjuntos — treinamento, validação e teste — conforme apresentado na Tabela 5.1. Como a classe Não-Representativa tem sub-categorias — Muito Próximas (MP), Próximas (P) e Distantes (D) — também apresentamos o número de imagens de cada sub-categoria designadas para o subconjunto de teste na Tabela 5.2.

Em cenários forenses, é comum ter acesso a um número limitado de imagens de um dado evento no início da investigação. Técnicas de aumentação de dados tornam-se cruciais para a expansão das capacidades de algoritmos de aprendizado. Por isso, consideramos técnicas de aumentação geométrica para melhorar o processo de aprendizado da combinação (etapa de aprendizado de *manifold*). Utilizamos as seguintes tranformações: recorte (*crop*), rotação (*rotation*) e ampliação (*zooming*). Realizamos as transformações antes do processo de treinamento. Cinco recortes aleatórios foram realizados para cada imagem, mantendo metade do tamanho original. Cinco rotações aleatórias entre $-30°$ e $30°$ foram

	Representativa			Não-Representativa			
Conjunto de Dados	Trein.	Val.	Teste	Trein.	Val.	Teste	Total
Wedding	280	75	84	19.055	6.101	6.438	32.033
Fire	507	76	390	28.222	7.465	8.420	45.080
Bombing	2.153	709	320	32.936	5.858	6.430	48.406

Tabela 5.1: Conjuntos de dados *Wedding, Fire* e *Bombing* separados em subconjuntos de treinamento (Trein.), validação (Val.) and teste (Teste).

		Categorias Não-Representativas			
Conjunto de dados	Representativa	MP	P	D	Total
Wedding	84	587	3.565	2.286	6.522
Fire	390	182	7.217	1.021	8.810
Bombing	320	0	3.216	3.214	6.750

Tabela 5.2: Subconjunto de teste dos conjuntos de dados *Wedding, Fire* e *Bombing* considerando a classe Representativa e categorias da classe Não-Representativa (Muito Próxima (MP), Próxima (P) e Distante (D)).

aplicadas a cada imagem. Por fim, cinco ampliações no intervalo de 1× a 5× foram realizadas em cada imagem. A Tabela 5.3 apresenta o número final de imagens em cada subconjunto dos conjuntos de dados depois da aumentação.

	Representativa			Não-Representativa		
Conjunto de Dados	Trein.	Val.	Teste	Trein.	Val.	Teste
Wedding	4.426	1.188	1.325	19.055	6.101	6.438
Fire	8.019	1.204	6.176	28.222	7.465	8.420
Bombing	34.080	11.231	5.066	32.936	5.858	6.430

Tabela 5.3: Número final de imagens em cada conjunto de dados — *Wedding, Fire* e *Bombing* — depois da aumentação nos subconjuntos de treinamento (Trein.), validação (Val.) e teste (Teste).

5.2.2 Extratores de Características

Para a extração de características, utilizamos os mesmos três componentes utilizados no Capítulo 4: *Places, Objects*, e *People*. Utilizamos CNNs para extrair as características relacionadas a cada um dos três componentes, obtendo as saídas da camada anterior à *Softmax* em cada rede adotada. Para o componente *Places*, nós aplicamos VGG16 [67] treinada no conjunto de dados Places [85] para extrair as características ($\mathbf{x}_{places}$) obtendo um vetor 4.096-dimensional. Para o componente *Objects*, nós utilizamos a Inception-ResNet [68] treinada no conjunto de dados Imagenet [52] para obter as características ($\mathbf{x}_{objects}$) gerando um vetor 1.536-dimensional. Por fim, para o componente *People*, nós

aplicamos uma rede de Re-identificação (*Re-Identification*) PCB [65] treinada no conjunto de dados Market-1501 [82] para descrever as características ($\mathbf{x}_{people}$) obtendo um vetor 12.288-dimensional.

5.2.3 Espaços Aprendidos

Para comparação, nós adotamos três referências: o vetor de características final extraído diretamente das três CNNs pré-treinadas (*Concatenated*); características de Espaço Semântico de Evento (ESS) como proposto no Capítulo 4; e o vetor de características extraídas das três CNNs depois de realizado o *fine-tuning* para cada conjunto de dados utilizados. Nós nos referimos a esta última referência como *Fine-tuned*. O *fine-tuning* das CNNs foi realizado por 50 épocas considerando as classes Representativa e Não-Representativa. Nós usamos o otimizador Adam [26] com $\alpha = 0,00001$ e taxa de decréscimo de $1e^{-5}$.

Chamamos as três redes de combinação propostas como *Cross-Entropy, Contrastive* e *Triplet*, de acordo com a função de perda usada no processo de treinamento. Nós dividimos as seis abordagens (três propostas + três referências) em dois grupos: aprendizado de classificação e aprendizado de distância. Os métodos inclusos no aprendizado de classificação usam uma rede de classificação em alguma parte do processo de aprendizado de características e incluem os métodos *Concatenated, ESS, Fine-Tuned*, e *Cross-Entropy*. As abordagens de aprendizado de distância, apesar de utilizar CNNs para extrair as características originais, aprendem o *manifold* usando uma abordagem de comparação, e incluem os métodos *Contrastive* e *Triplet*.

5.2.4 Explorando Largura e Profundidade da Rede

Nosso primeiro experimento explora a arquitetura de rede considerando quatro redes com diferentes profundidades e larguras. Como a Figura 5.1 mostra, a rede proposta compreende uma camada densa para redução de características de cada vetor de entrada e uma camada densa para a combinação das características reduzidas. As variações experimentadas dessa rede foram chamadas de $N(512, 128)$, $N(512, 128, 64)$, $N(1024, 512)$ e $N(1024, 512, 128)$, nas quais o primeiro valor (entre parênteses) determina o número de neurônios na camada de *Redução* e os outros valores representam o número de neurônios na(s) camada(s) de *Combinação*.

A rede $N(512, 128)$ apresenta 512 neurônios em cada uma das sub-camadas de *Redução* e 128 neurônios na camada densa de *Combinação*. A rede $N(512, 128, 64)$ segue a configuração inicial da $N(512, 128)$ mas aumenta o número de camadas acrescentando uma camada com 64 neurônios depois da primeira camada de *Combinação*. A rede $N(1024, 512)$ explora uma variação de largura. Essa rede tem 1.024 neurônios em cada sub-camada densa de *Redução* (ao invés dos 512 da $N(512, 128)$) e 512 neurônios na camada de *Combinação* (ao invés dos 128 da $N(512, 128)$). Por fim, a rede $N(1024, 512, 128)$ aumenta tanto largura quanto profundidade usando a $N(1024, 512)$ como base e adicionando uma camada densa com 128 neurônios depois da primeira camada de *Combinação*.

O experimento com essas quatro arquiteturas foi realizado usando as características extraídas da última camada, antes da camada de saída, obtendo 128 características para

a rede $N(512, 128)$, 64 características para a rede $N(512, 128, 64)$, 512 características para a rede $N(1024, 512)$ and 128 características para a rede $N(1024, 512, 128)$. As arquiteturas foram treinadas usando as três diferentes funções de custo adotadas — *Entropia Cruzada, Contrastiva* e *Tripla* — para os três conjuntos de dados. As imagens de teste foram recuperadas usando imagens Representativas do treinamento como consultas. Os resultados foram avaliados usando curvas de Precisão × Revocação.

5.2.5 Variando Tamanho do Treinamento

É essencial avaliar os resultados dos métodos considerando um cenário com exemplos limitados de imagens de treinamento. Esse experimento foi realizado utilizando dados aumentados de imagens Representativas aleatoriamente escolhidas para compor conjuntos de treinamento de diferentes tamanhos. O número de imagens de treinamento Representativas originais (desconsiderando a aumentação) foram $10, 20, 50, 100$ e 200. As imagens Não-Representativas foram aleatoriamente selecionadas considerando o mesmo número de imagens no conjunto final de Representativas (originais + aumentadas). A seleção das imagens de treinamento está relacionada com a qualidade do *manifold* aprendido. Por exemplo, se realizarmos o treinamento com dez imagens Representativas muito parecidas, provavelmente o *manifold* não será capaz de representar a complexidade do evento (pouca diversidade). Por essa razão, nós selecionamos dez conjuntos de treinamento diferentes para cada tamanho de treinamento. Usando essas imagens de treinamento selecionadas, nós treinamos dez redes com cada tamanho de conjunto de forma a observar o comportamento sob mudanças de imagens de treinamento.

Nós treinamos as duas melhores arquiteturas (determinadas no experimento anterior) usando as funções de perda *Entropia Cruzada, Contrastiva* e *Tripla*. Comparamos os resultados obtidos pelos diferentes tamanhos de treinamento com a representação ESS (Capítulo 4) recuperando as imagens de teste usando as imagens de treinamento Representativas como consulta. Para gerar a representação ESS, nós usamos as mesmas imagens dos conjuntos de treinamento de diferentes tamanhos como Event Representative Images (ERIs). Utilizamos a média de dez valores de MAP (obtidos pelos *rankings* gerados para cada um dos cinco tamanhos de conjunto de treinamento) para avaliar a qualidade dos *rankings*.

5.2.6 Visualizando Separação das Classes

Nesse experimento, nós exploramos a discriminabilidade de cada um dos seis espaços de características analisados, com e sem aumentação de dados. As representações *ESS, Fine-Tuned, Cross-Entropy, Contrastive* e *Triplet* foram obtidas usando todas as imagens Representativas de treinamento (Tabelas 5.1 e 5.3) e a mesma quantidade de imagens Não-Representativas.

Para a visualização, projetamos todos os vetores de características de teste em uma espaço 2-dimensional utilizando a técnica de redução de dimensionalidade UMAP (vide Seção 2.7.1). O número de vizinhos utilizado como parâmetro foi 25, e a distância dentre amostras foi 0,01. Como a métrica de distância utilizada para a composição dos *rankings*

foi a distância Euclidiana, também a utilizamos como parâmetro do UMAP. Os vetores foram reduzidos para três dimensões. A primeira e segunda dimensões foram usadas como eixos x e y do gráfico de dispersão, enquanto a terceira dimensão foi utilizada como ordem de exibição das amostras no gráfico. Essa ordem significa que a amostra com o valor mais significante de terceira dimensão é desenhada primeiro (longe do observador), e a amostra com menor valor é desenhada por último (mais perto do observador). Consideramos quatro classes de discriminação – *Representativa, Não-Representativa Muito Próxima (MP), Não-Representativa Próxima (P)* e *Não-Representativa Distante (D).*

5.2.7 Recuperando Imagens Representativas

Nós realizamos esse experimento utilizando dados aumentados e sem aumentação. Utilizamos todas as imagens Representativas de treinamento como consultas para o método de recuperação. Foram calculadas as distâncias Euclidianas de todas as imagens de teste para as consultas, e as distâncias foram utilizadas para ordenar as imagens mais similares para as consultas. A Precisão dos *rankings* gerados foi avaliado por pontos de Revocação $(10\%, 20\%, ..., 90\%, 100\%)$, e a média e os valores de desvio padrão da Precisão foram obtidos com base nos *rankings*. Nós geramos a curva de Precisão $\times$ Revocação baseada nos valores das médias e dos desvios padrão.

Nós também calculamos os valores de média e variância para cada classe — Representativa e Não-Representativa — baseando-se nas distâncias das imagens dessas classes para as consultas.

5.2.8 Analisando Qualidade Visual dos *Rankings*

Por fim, este experimento tem como objetivo a comparação qualitativa de exemplos do top recuperado (*top*@5) pelos *rankings* gerados pelos seis espaços de características para a análise de qualidade. Os *rankings* foram obtidos usando os dados de treinamento e seus exemplos aumentados.

5.3 Resultados

O primeiro experimento avalia os resultados obtidos por diferentes profundidades/larguras na rede de combinação. As figuras 5.5, 5.7 e 5.9 apresentam os resultados utilizando as funções de perda Entropia Cruzada, Contrastiva e Tripla, respectivamente, para o conjunto de dados não aumentado. Já as figuras 5.6, 5.8 e 5.10 apresentam os resultados utilizando as funções de perda Entropia Cruzada, Contrastiva e Tripla, respectivamente, para o conjunto de dados aumentado. As redes $N(512, 128)$ e $N(1024, 512)$ apresentaram as melhores taxas de precisão no conjuntos de dados com menos imagens de treinamento como o conjunto *Wedding*, e também apresentando bons resultados para os conjuntos de treinamento sem aumentação (figuras 5.5, 5.7 e 5.9).

As redes com menos camadas na parte de *Combinação* apresentaram melhores resultados, possivelmente devido ao menor número de amostras requeridas para treinamento.

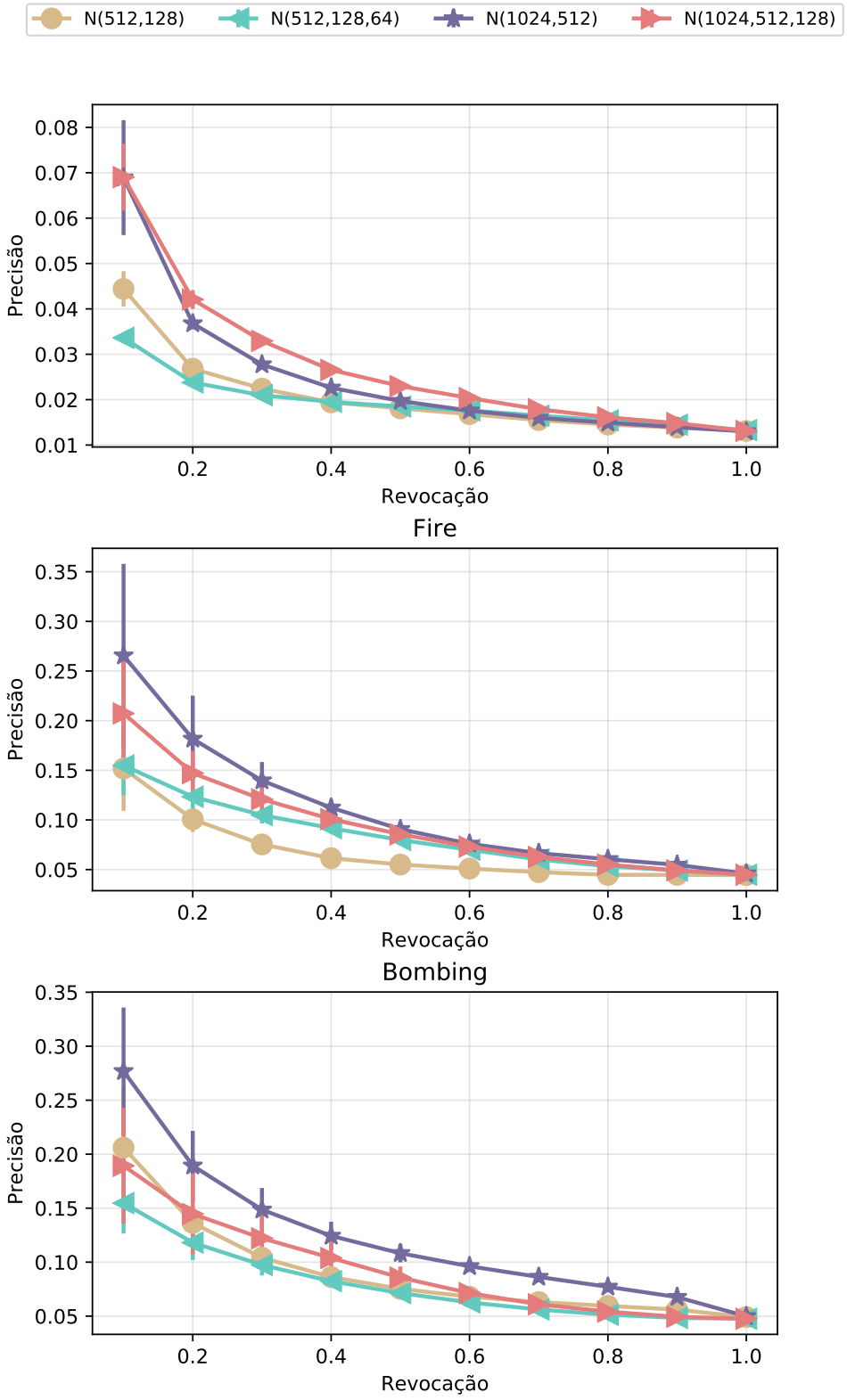


Figura 5.5: Resultados no conjunto de teste com as redes treinadas utilizando a função de perda Entropia Cruzada e o conjunto de dados sem aumentação. As redes $N(1024, 512)$ e $N(1024, 512, 128)$ apresentaram a melhor precisão em quase todos os conjuntos de dados.

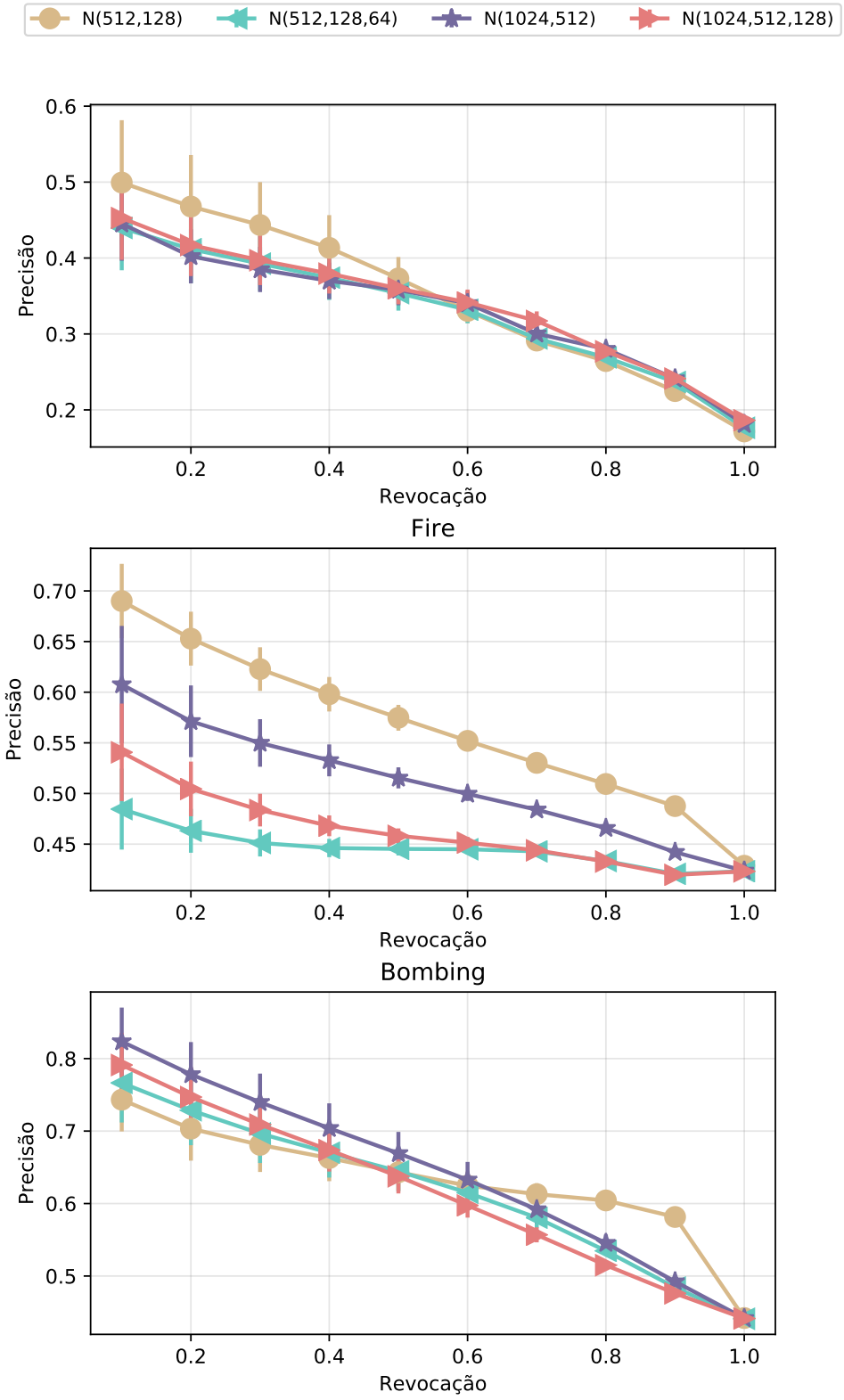


Figura 5.6: Resultados no conjunto de teste com as redes treinadas utilizando a função de perda Entropia Cruzada e o conjunto de dados com aumentação. As redes $N(512, 128)$ e $N(1024, 512)$ apresentaram a melhor precisão em quase todos os conjuntos de dados.

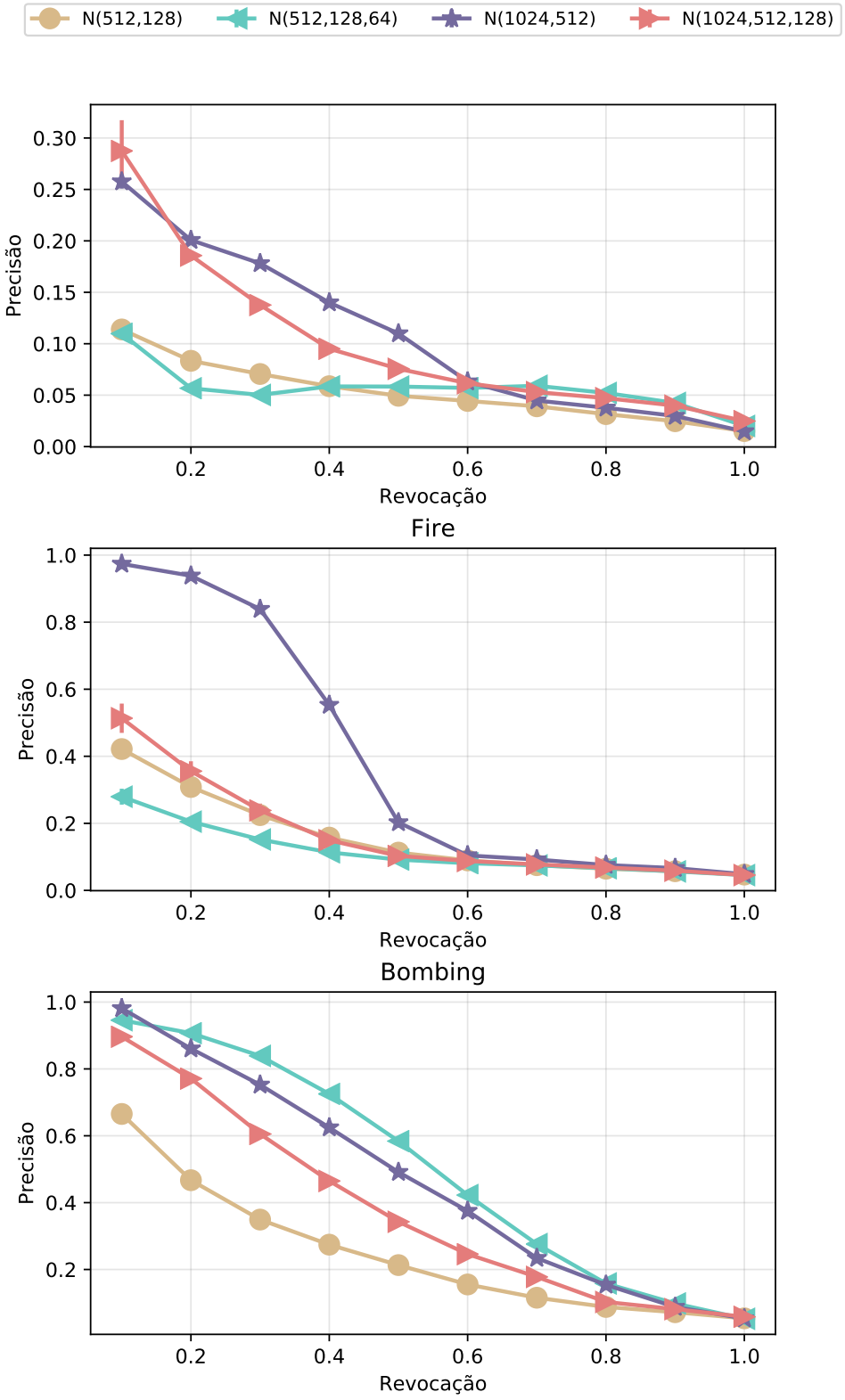


Figura 5.7: Resultados no conjunto de teste com as redes treinadas utilizando a função de perda Contrastiva. A rede $N(1024, 512)$ apresentou a melhor precisão para dois conjuntos de dados sem o uso de técnicas de aumentação.

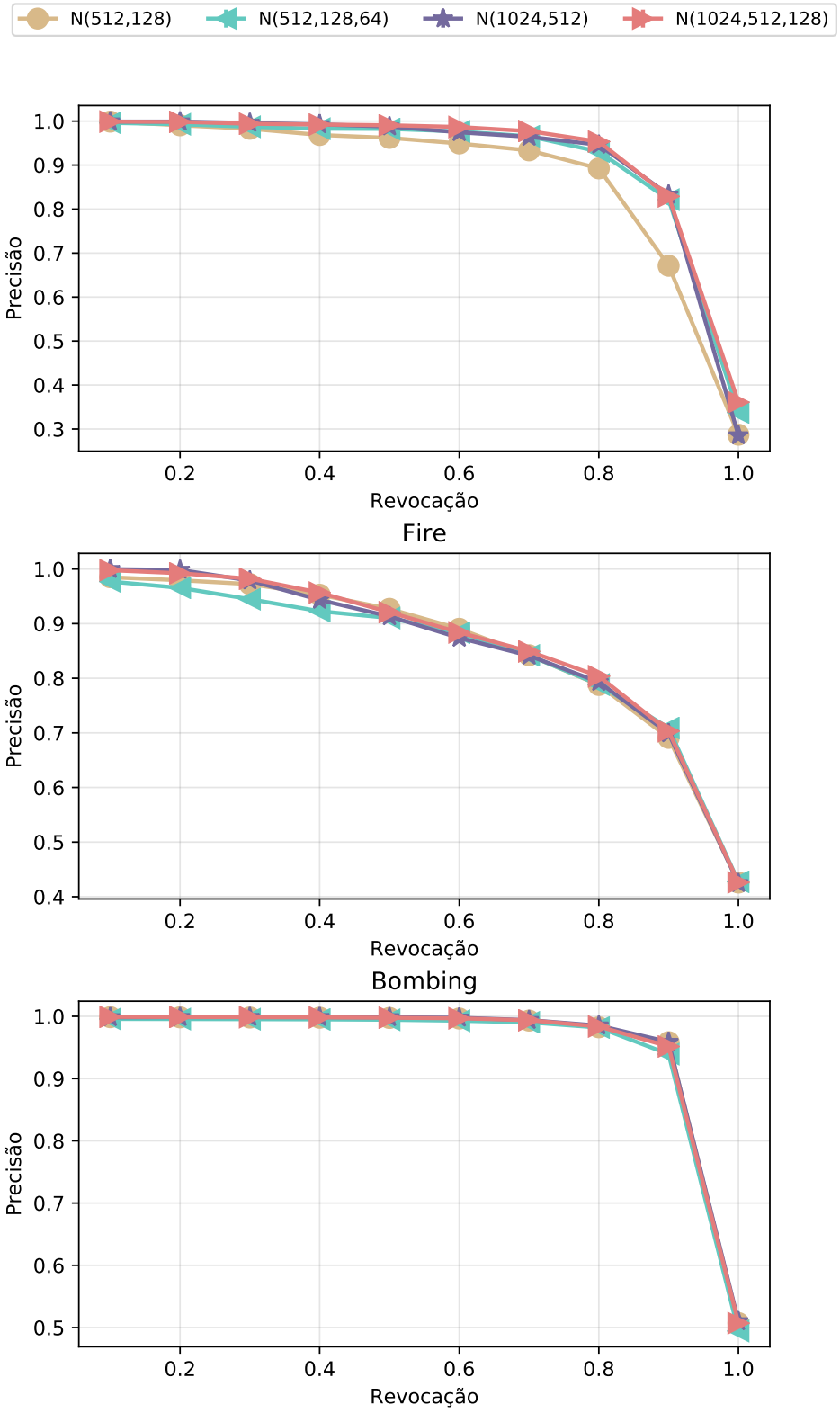


Figura 5.8: Resultados no conjunto de teste com as redes treinadas utilizando a função de perda Contrastiva. As rede apresentaram resultados similares com o uso dos dados aumentados.

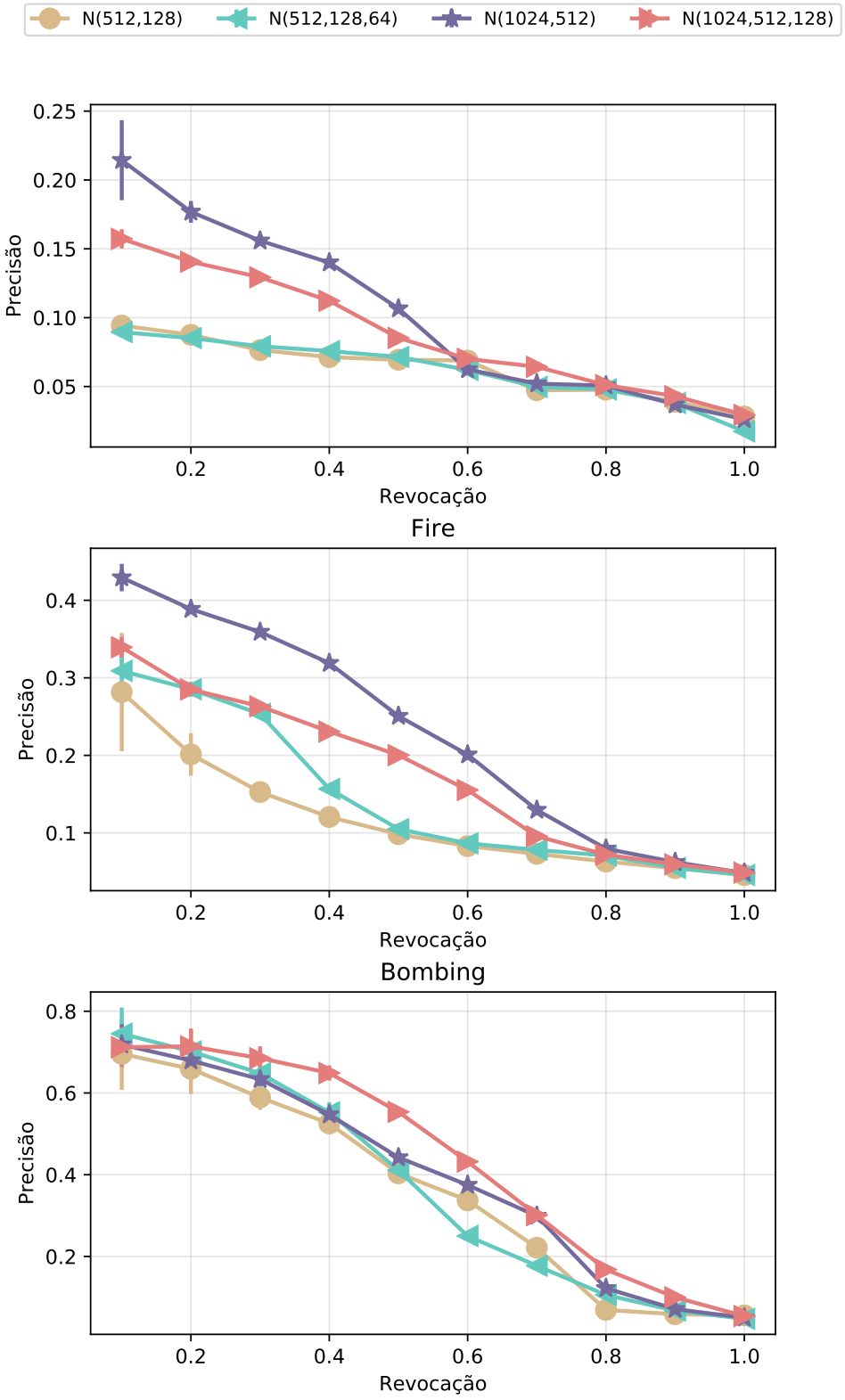


Figura 5.9: Resultados no conjunto de teste com as redes treinadas utilizando a função de perda Tripla. A rede $N(1024, 512)$ apresentou melhor precisão em dois dois eventos, sem o uso de técnicas de aumentação.

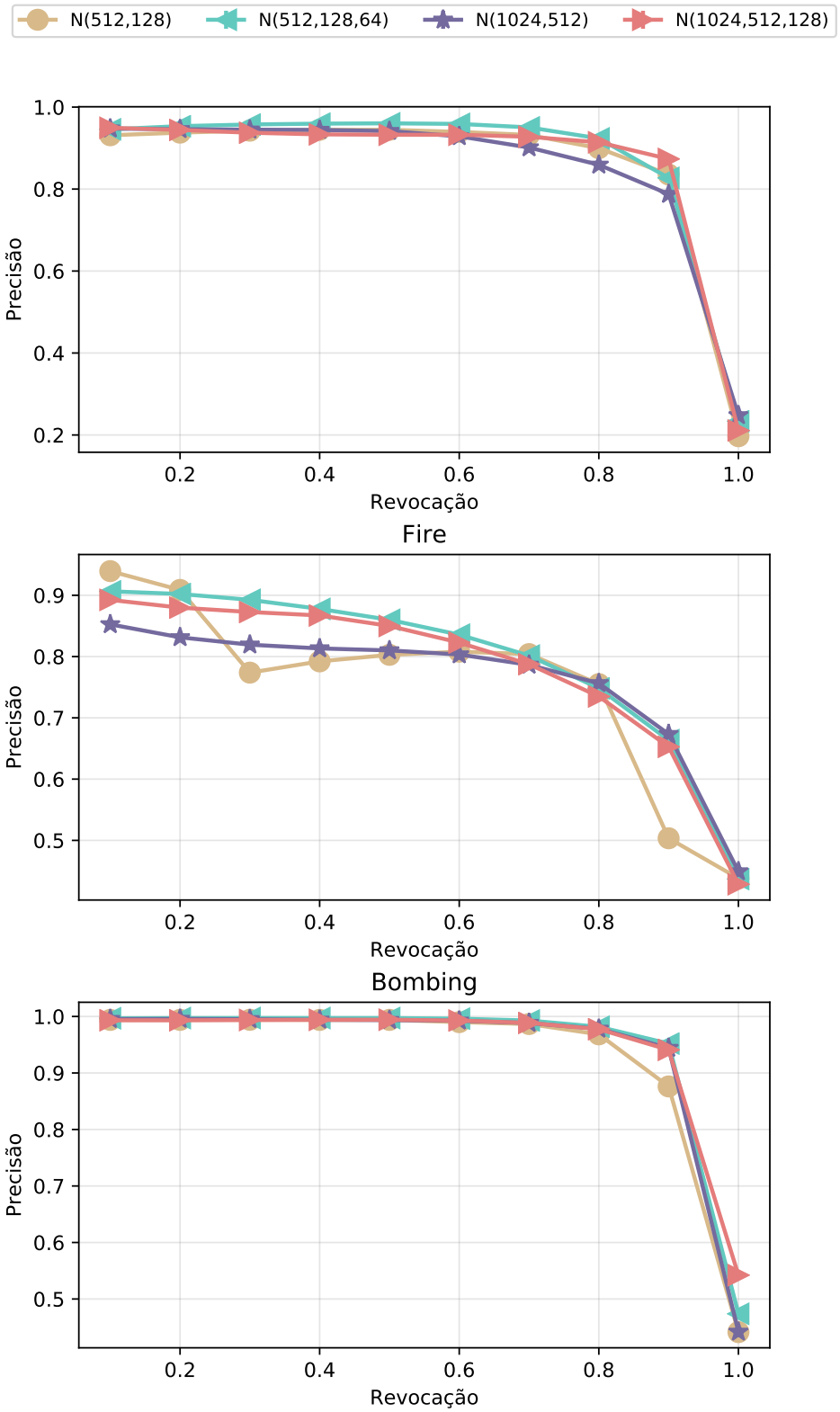


Figura 5.10: Resultados no conjunto de teste com as redes treinadas utilizando a função de perda Tripla. As rede apresentaram resultados similares com o uso dos dados aumentados.

Nós selecionamos as redes $N(512, 128)$ e $N(1024, 512)$ para o próximo experimento no qual nós exploramos o comportamento de diferentes tamanhos de conjuntos de treinamento.

As figuras 5.11, 5.12 e 5.13 apresentam os resultados obtidos quando treinamos as redes com menor número de imagens originais representativas, usando as funções de perda Entropia Cruzada, Contrastiva e Tipla, respectivamente. Nós comparamos $N(512, 128)$ e $N(1024, 512)$ com o método semi-supervisionado ESS apresentado no Capítulo 4. Mesmo com menos imagens de treinamento, a combinação de componentes produzida por $N(512, 128)$ e $N(1024, 512)$ apresentaram os melhores valores de MAP, resultando na melhoria de mais de 10 pontos percentuais, com relação ao ESS, no menor conjunto de imagens de treinamento testadas (com apenas 10 imagens Representativas originais). Nós também percebemos resultados um pouco superiores para a $N(1024, 512)$. Com base nesses resultados, continuamos os experimentos adotando a rede $N(1024, 512)$ como opção padrão.

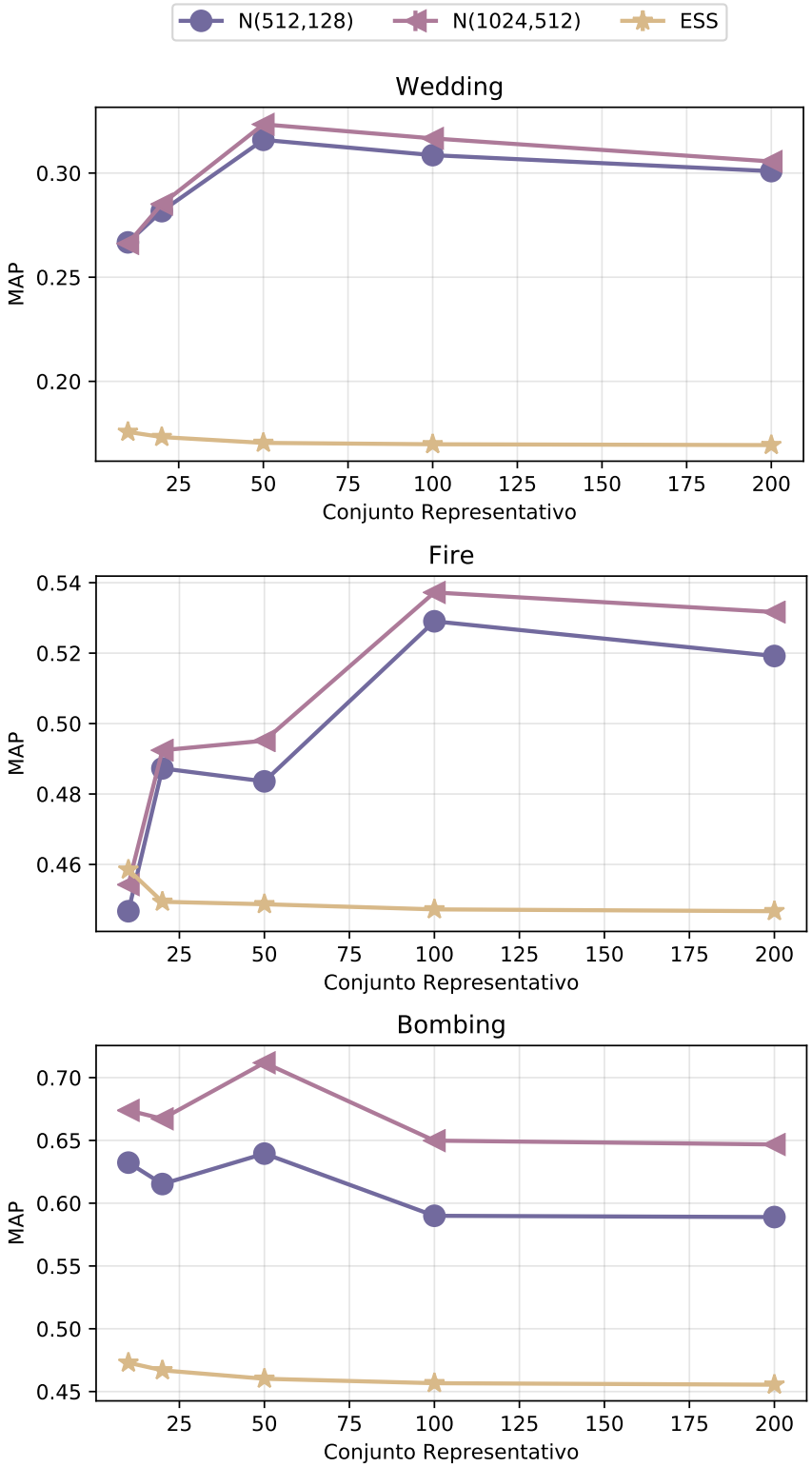


Figura 5.11: Resultados de MAP para as redes treinadas com a função de perda Entropia Cruzada com conjuntos pequenos de treinamento. $N(512, 128)$ e $N(1024, 512)$ apresentaram comportamento similar com resultados superiores ao MAP obtido pelo ESS mesmo para conjuntos de treinamento muito pequenos (i.e., dez imagens representativas originais).

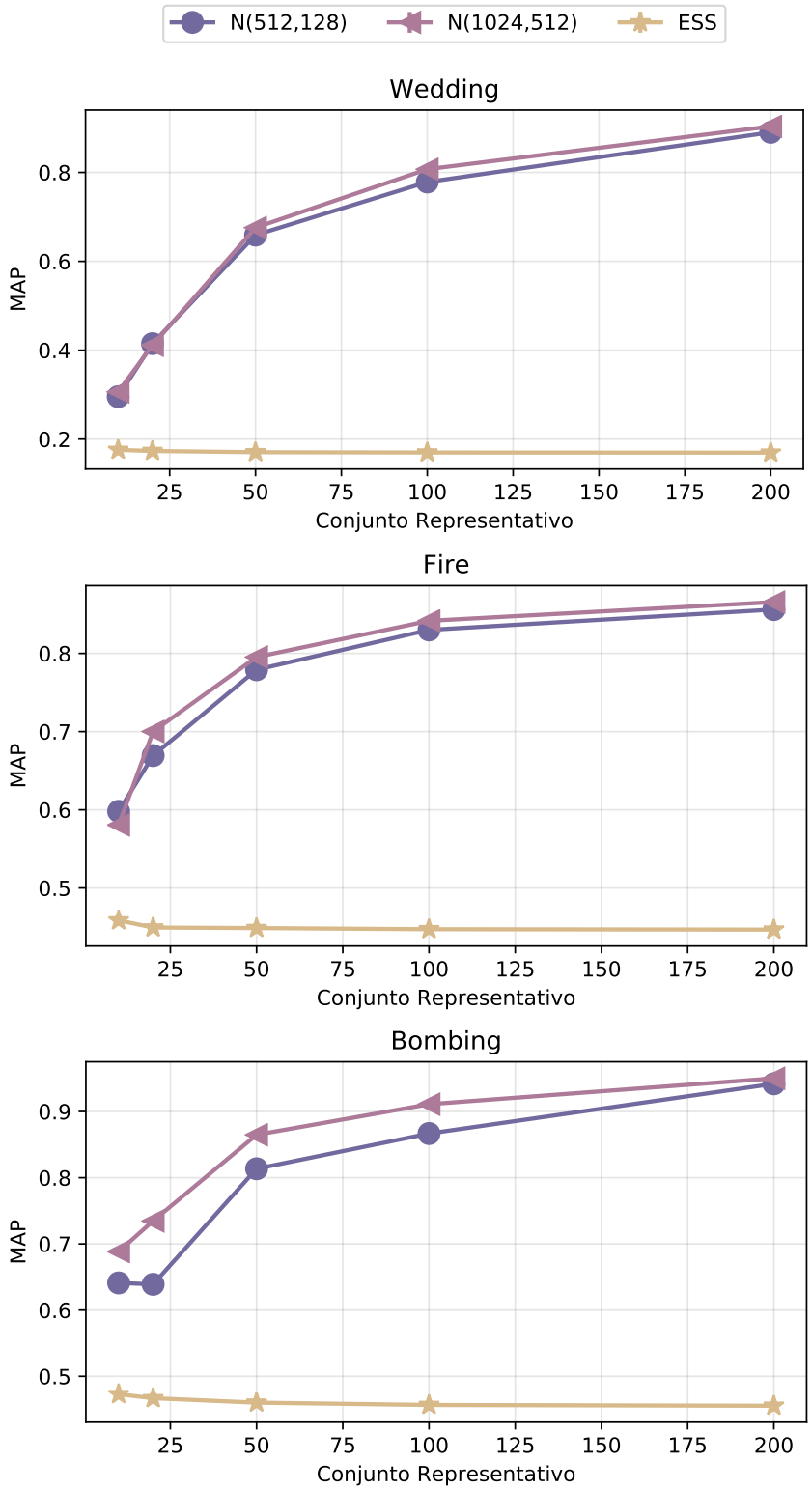


Figura 5.12: Resultados de MAP para as redes treinadas com a função de perda Contrastiva com conjuntos pequenos de treinamento. $N(512, 128)$ e $N(1024, 512)$ apresentaram comportamento similar com resultados superiores ao MAP obtido pelo ESS mesmo para conjuntos de treinamento muito pequenos (i.e., dez imagens representativas originais).

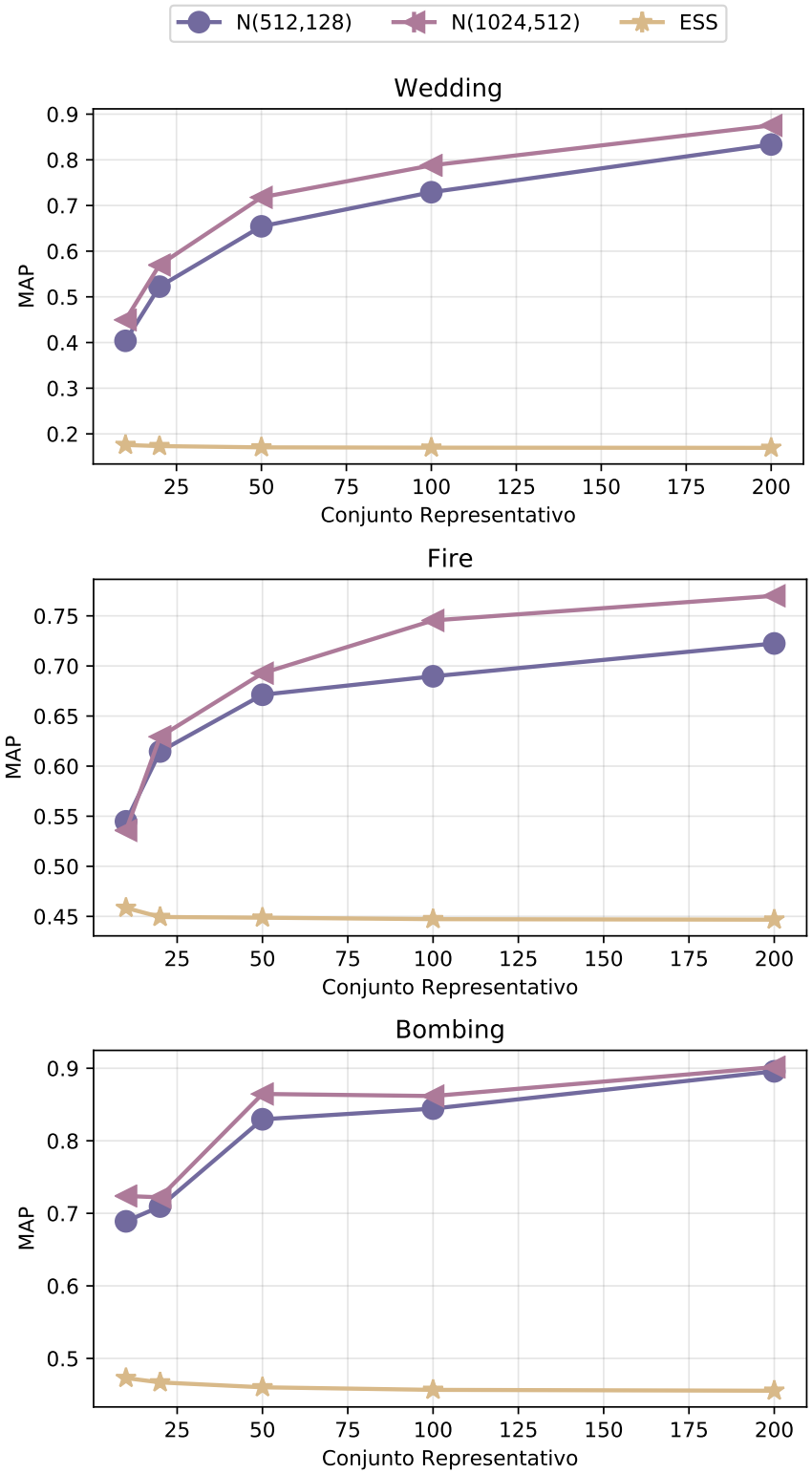


Figura 5.13: Resultados de MAP para as redes treinadas com a função de perda Tripla com conjuntos pequenos de treinamento. $N(512, 128)$ e $N(1024, 512)$ apresentaram comportamento similar com resultados superiores ao MAP obtido pelo ESS mesmo para conjuntos de treinamento muito pequenos (i.e., dez imagens representativas originais).

Com a escolha da rede de combinação $N(1024, 512)$, procuramos analisar como as imagens se comportam no espaço de representação. Para isso, projetamos as imagens em um espaço $2-$dimensional, verificando se a combinação proporciona melhor separação entre classes. As figuras 5.14 e 5.15 apresentam a projeção de todas as imagens dos conjuntos de dados nos espaços aprendidos, com e sem a aplicação de técnicas de aumentação de dados, usando as classes Representativa e Não-Representativa. Nós percebemos um comportamento similar entre as abordagens de classificação — *Concatenated, Fine-Tuned* e *Cross-Entropy* — e até para o ESS, apresentando maior dispersão entre imagens de ambas as classes. Por outro lado, as abordagens de aprendizado de distância — *Contrastive* e *Triplet* — apresentaram separações muito melhores, agrupando melhor imagens da mesma classe. As técnicas de aumentação mostraram ainda menor mistura entre classes, especialmente para as abordagens de aprendizado de distância. O conjunto de dados *Bombing* parece ser o mais facilmente separável quando consideramos os espaços 2-dimensionais apresentados.

Nas figuras 5.14 e 5.15 (com e sem aumentação, respectivamente), se considerarmos os círculos não roxos como as categorias MP, P e D, podemos realizar uma análise de níveis de representatividade. Os *manifolds* construídos pelas abordagens *Contrastive* e *Triplet* aparentemente separam também esses níveis, em alguns casos como no conjunto de dados *Wedding* afastando ainda mais imagens D das imagens Representativas.

Com base nas visualizações, podemos observar que possivelmente as abordagens de aprendizado de distância são capazes de separar melhor as imagens Representativas das Não-Representativas, especialmente quando usamos dados aumentados para o treinamento. Para verificar os resultados dos métodos na tarefa de recuperação, nós partimos para o quarto experimento, recuperando imagens Representativas utilizando os seis métodos descritos de representação de imagens.

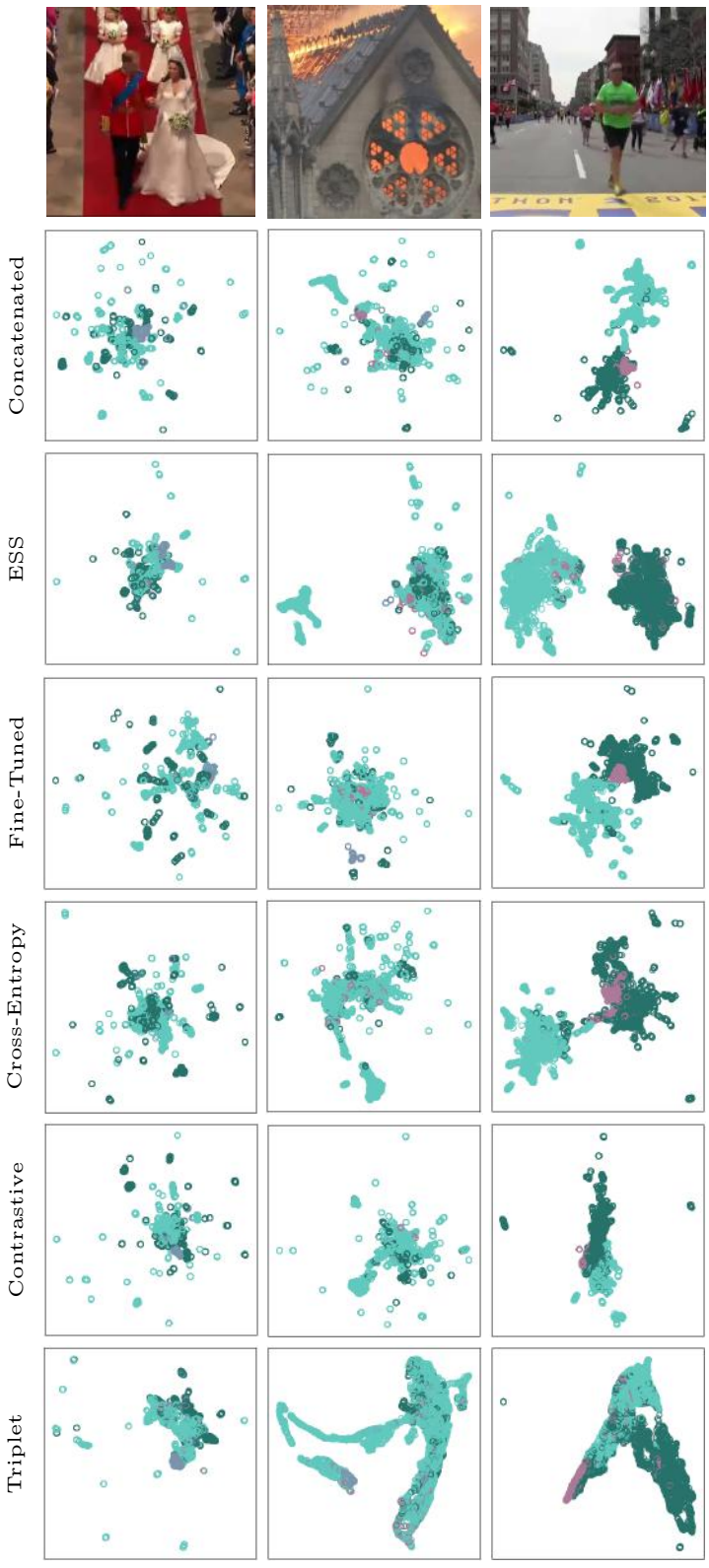


Figura 5.14: Gráficos de dispersão das imagens Representativas (círculos roxos), Não-Representativas MP (círculos azul acinzentados – em quantidades muito pequenas), Não-Representativas P (círculos verde claros), e Não-Representativas D (círculos verde escuros) de todos os conjuntos de dados nos espaços aprendidos sem aumentação de dados. As quatro classes de visualização destacam a capacidade do espaço *Triplet* em separar imagens Não-Representativas D das imagens Representativas.

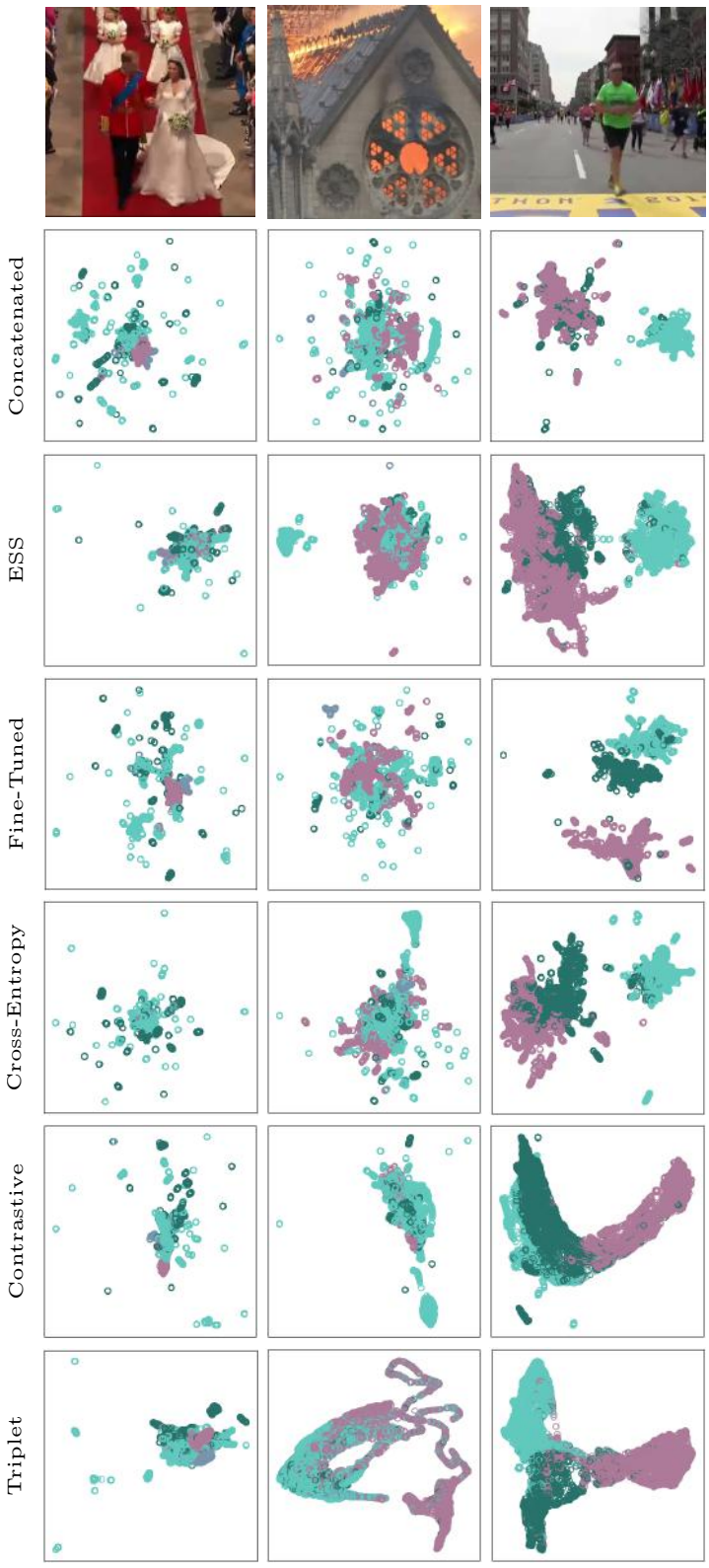


Figura 5.15: Gráficos de dispersão das imagens Representativas (círculos roxos), Não-Representativas MP (círculos azul acinzentados – em quantidades muito pequenas), Não-Representativas P (círculos verde claros), e Não-Representativas D (círculos verde escuros) de todos os conjuntos de dados nos espaços aprendidos com aumentação de dados. As quatro classes de visualização destacam a capacidade dos espaços *Contrastive* e *Triplet* em separar exemplos não relacionados das imagens Representativas.

As figuras 5.16 e 5.17 apresentam as curvas de Precisão × Revocação dos diferentes métodos, sem e com a aplicação de de técnicas de aumentação de dados. Todos os espaços gerados pelo aprendizado da combinação (*Cross-Entropy, Contrastive* e *Triplet*) superaram a abordagem de concatenação em todos os conjuntos de dados. Além disso, as técnicas propostas também apresentaram melhores resultados que o espaço *Fine-Tuned* e o ESS. Notamos ainda que, as abordagens de aprendizado de distância (*Contrastive* e *Triplet*) contribuiram mais para a tarefa de recuperação melhorando a precisão em mais de 20 pontos percentuais, para todos os conjuntos de dados.

Os resultados dos *rankings* são confirmados pela média das distâncias das imagens Representativas e Não-Representativas. As figuras 5.18 e 5.19 apresentam a média e variância das distâncias das imagens para as consultas dos conjuntos de dados *Wedding*, *Fire* e *Bombing*. Nesse experimento, nós esperamos que as imagens Representativas estejam mais próximas das consultas (médias baixas). Por essa razão, um resultado considerado bom mostra imagens Representativas com média e variância baixas. Como esperado, os três métodos propostos (*Cross-Entropy, Contrastive* e *Triplet*) mostram a média de distâncias para as imagens Representativas (com baixa variância) menor que a média de distâncias das imagens Não-Representativas. Os resultados obtidos indicam que o aprendizado da combinação permite melhor separação entre as imagens do evento.

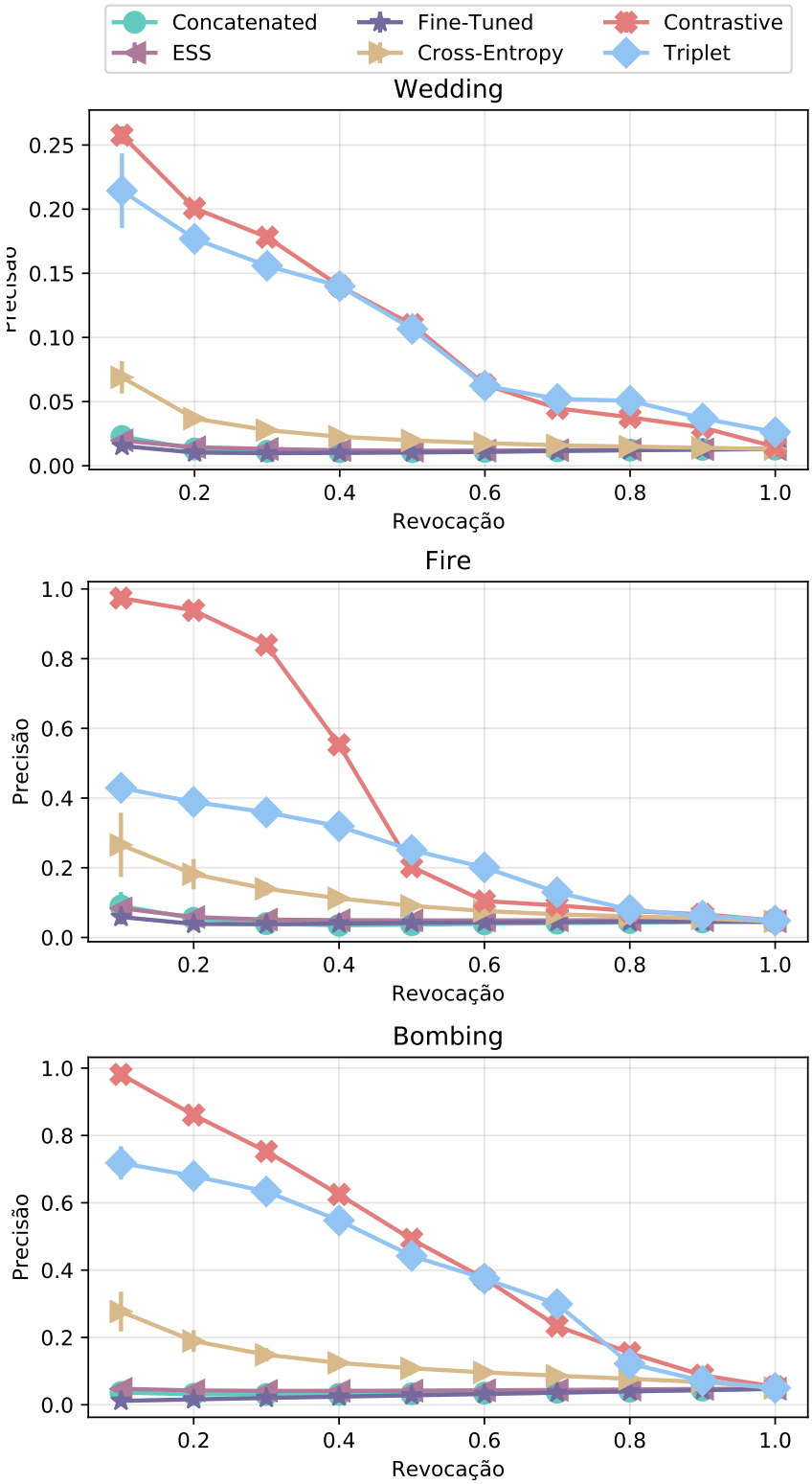


Figura 5.16: Abordagens de aprendizado de distâncias (*Contrastive* e *Triplet*) apresentaram os melhores espaços aprendidos. Aprender a combinar componentes mostrou melhorias na precisão dos *rankings*. Mesmo a abordagem de combinação por classificação (*Cross-Entropy*) mostrou resultados superiores aos espaços *Pre-trained* e *Fine-Tuned*. Os resultados apresentados correspondem às técnicas sem o uso de aumentação de dados.

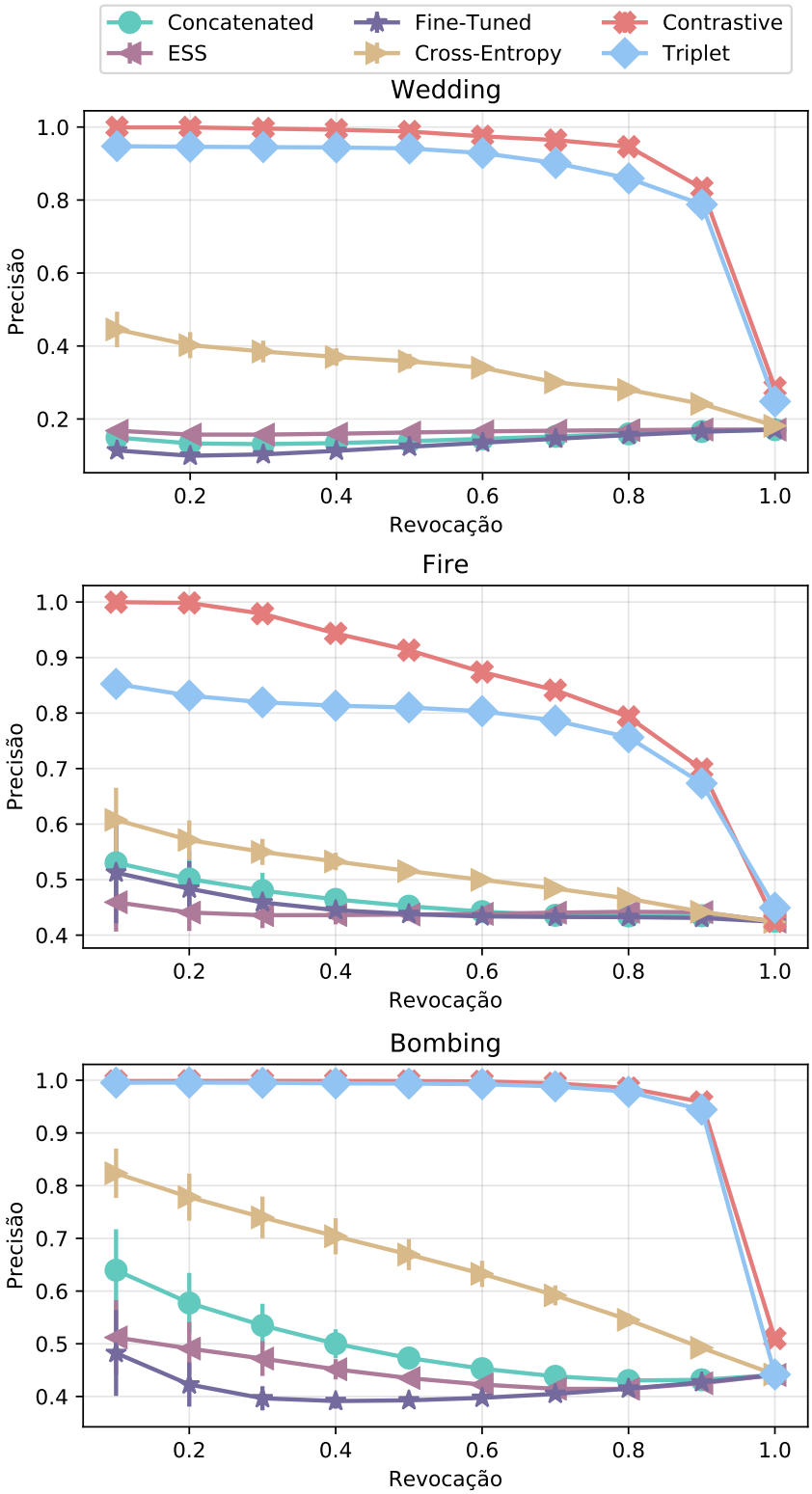


Figura 5.17: Abordagens de aprendizado de distâncias (*Contrastive* e *Triplet*) apresentaram os melhores espaços aprendidos. Aprender a combinar componentes mostrou melhorias na precisão dos *rankings*. Mesmo a abordagem de combinação por classificação (*Cross-Entropy*) mostrou resultados superiores aos espaços *Pre-trained* e *Fine-Tuned*. Os resultados apresentados correspondem às técnicas utilizando dados aumentados.

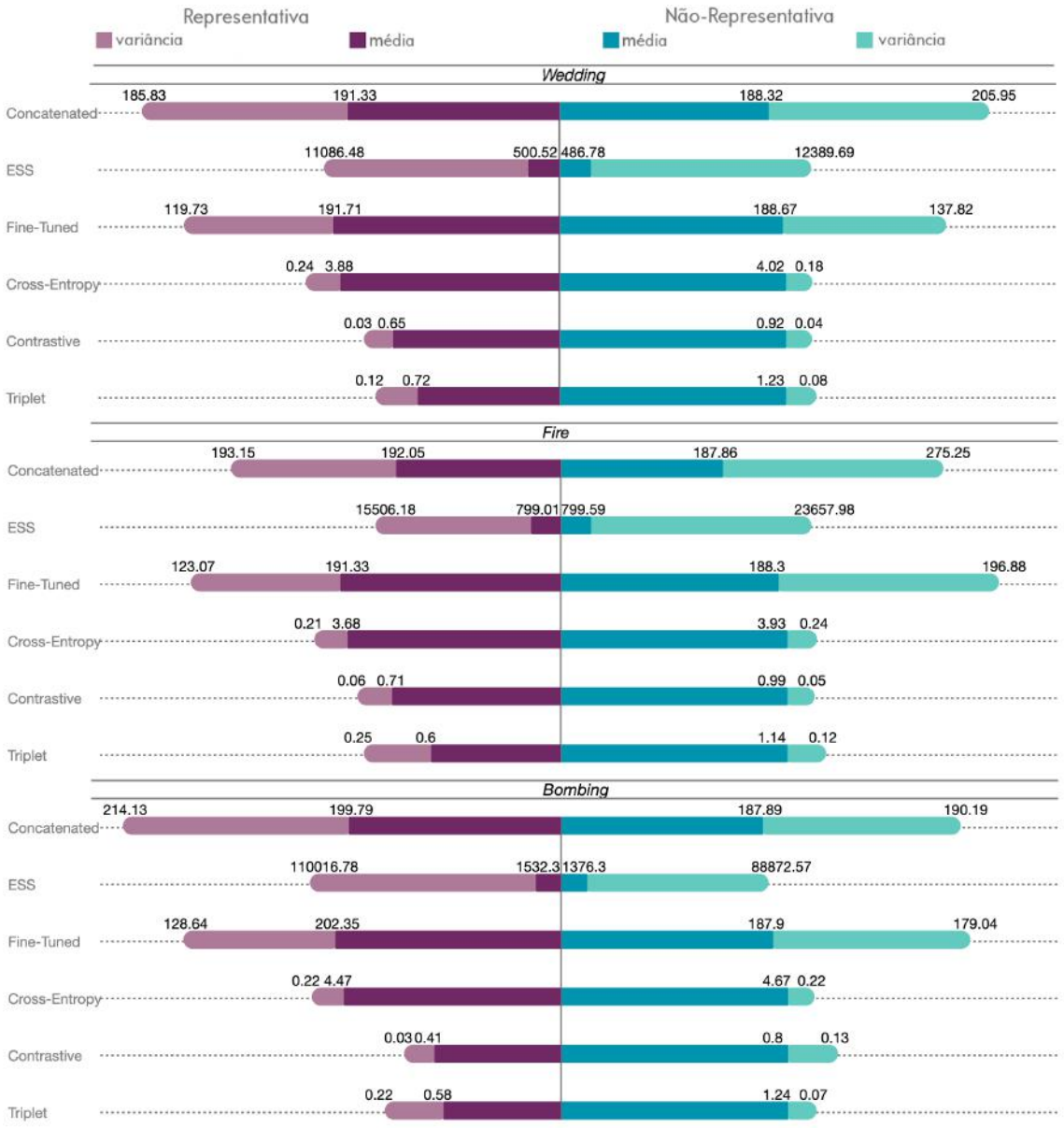


Figura 5.18: Média e variância das distâncias das imagens de cada classe no conjunto de teste com relação às consultas dos conjuntos de dados *Wedding*, *Fire* e *Bombing*, sem o uso de técnicas de aumentação de dados.

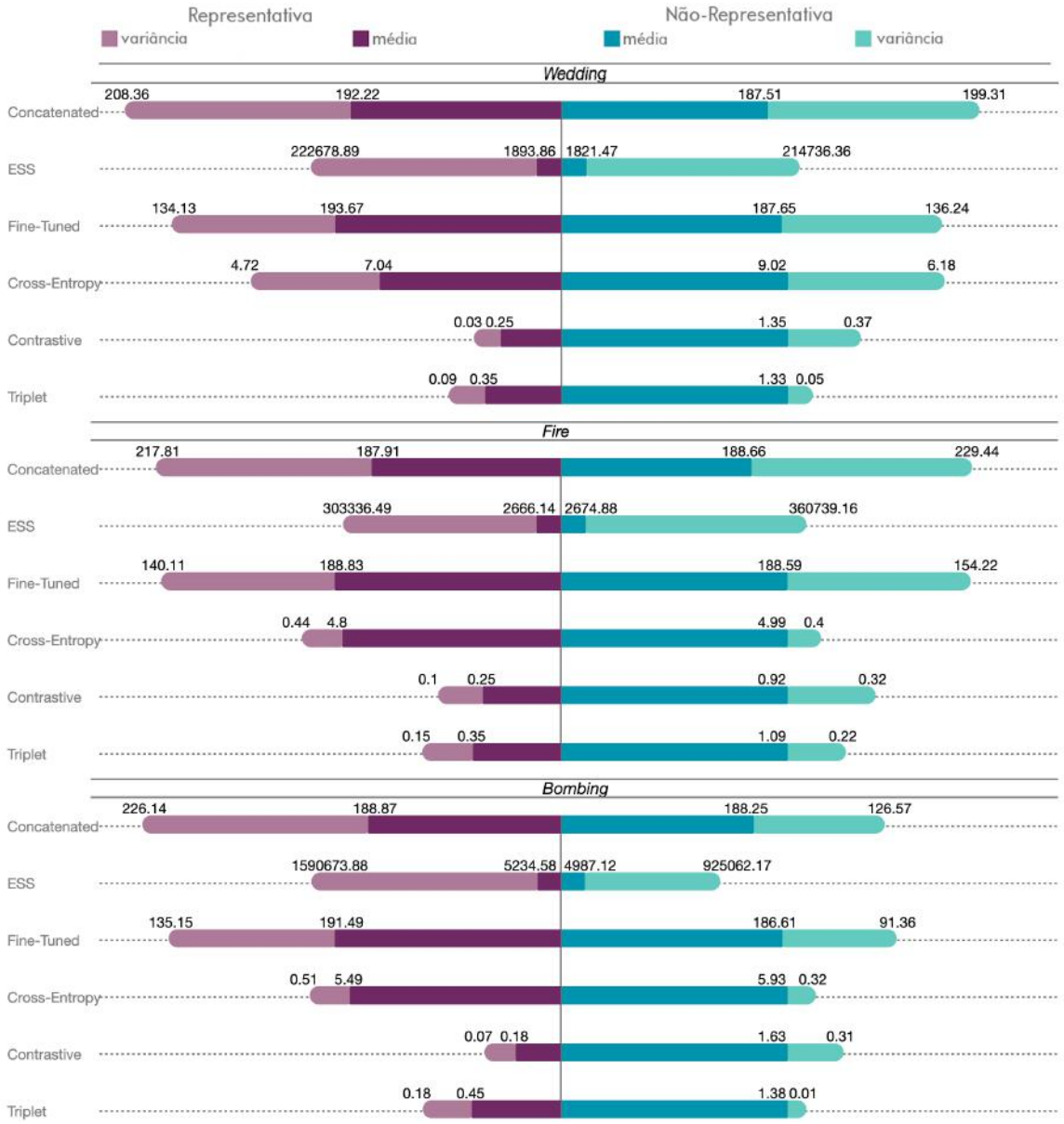


Figura 5.19: Média e variância das distâncias das imagens de cada classe no conjunto de teste com relação às consultas dos conjuntos de dados *Wedding*, *Fire* e *Bombing*, utilizando técnicas de aumentação de dados.

Por fim, esperamos ver as diferenças entre as imagens recuperadas no topo dos *rankings*, utilizando os seis diferentes tipos de características. Para essa análise qualitativa, nós observamos os *rankings* apresentados nas figuras 5.20, 5.21 e 5.22 para os conjuntos de dados *Wedding*, *Fire* e *Bombing*, respectivamente. A primeira linha apresenta a imagem de consulta utilizada, seguida pelas imagens recuperadas utilizando as características *Concatenated, ESS, Fine-Tuned, Cross-Entropy, Contrastive* e *Triplet* em cada coluna (*top*@5), respectivamente. As bordas roxas indicam imagens Representativas, enquanto as bordas verdes indicam as imagens Não-Representativas. Notamos que os *manifold aprendidos* pela combinação de componentes (métodos propostos) proporcionam características mais discriminativas, melhorando a precisão e diversidade do topo recuperado. *Contrastive* e *Triplet* apresentam os melhores *rankings* considerando o caso de precisão e diversidade.

Figura 5.20: Comparação dos resultados do topo dos *rankings* obtidos por diferentes métodos de representação utilizando o conjunto de dados aumentado. *Rankings* obtidos utilizando os *manifolds* aprendidos – *Cross-Entropy, Contrastive* e *Triplet* – apresentaram melhor precisão e variabilidade nas posições do *top*@5. A primeira linha apresenta a imagem de consulta seguida pelas *top*@5 imagens recuperadas (por colunas) para o conjunto de dados *Wedding*. Bordas roxas nas imagens indicam imagens Representativas e bordas verdes indicam imagens Não-Representativas.

Figura 5.21: Comparação dos resultados do topo dos *rankings* obtidos por diferentes métodos de representação utilizando o conjunto de dados aumentado. *Rankings* obtidos utilizando os *manifolds* aprendidos – *Cross-Entropy, Contrastive* e *Triplet* – apresentaram melhor precisão e variabilidade nas posições do *top*@5. A primeira linha apresenta a imagem de consulta seguida pelas *top*@5 imagens recuperadas (por colunas) para o conjunto de dados *Fire*. Bordas roxas nas imagens indicam imagens Representativas e bordas verdes indicam imagens Não-Representativas.

Figura 5.22: Comparação dos resultados do topo dos *rankings* obtidos por diferentes métodos de representação utilizando o conjunto de dados aumentado. *Rankings* obtidos utilizando os *manifolds* aprendidos – *Cross-Entropy, Contrastive* e *Triplet* – apresentaram melhor precisão e variabilidade nas posições do *top*@5. A primeira linha apresenta a imagem de consulta seguida pelas *top*@5 imagens recuperadas (por colunas) para o conjunto de dados *Bombing*. Bordas roxas nas imagens indicam imagens Representativas e bordas verdes indicam imagens Não-Representativas.

5.4 Discussões sobre o Espaço Combinado de Evento

Notamos no Capítulo 4 que a decomposição de eventos em componentes representativos — *Places, Objects* e *People* — foi capaz de produzir uma representação de baixa dimensionalidade para recuperação de imagens. No entanto, a contribuição de cada componente para tipos diferentes de eventos não foi analisada principalmente pela falta de supervisão. Essa falta de supervisão acarretou restrições de eficácia na recuperação quando utilizada a representação ESS.

Por essa razão, decidimos introduzir supervisão, com um conjunto limitado de imagens e algumas técnicas de aumentação de dados, para aprender a gerar melhores representações. Como primeira tentativa de adaptar a descrição de imagens ao nosso problema, utilizamos a técnica de *fine-tuning* das redes extratoras de características (CNNs) com os conjuntos de dados deste trabalho (*Wedding, Fire* e *Bombing*). Os resultados dos *rankings* obtidos com o uso da representação *Fine-Tuned* demonstram o problema em utilizar conjuntos pequenos de treinamento atingindo baixa precisão (Figura 5.16) mesmo com o uso de dados aumentados (Figura 5.17). Assim, propomos nossas abordagens de combinação, que não pretendem aprender as características diretamente do evento alvo, mas sim, combiná-las.

Os métodos de combinação propostos apresentaram altas taxas de precisão mesmo com o uso de conjuntos de treinamento pequenos. Fatores importantes para a precisão do método proposto consistem na escolha correta das redes convolucionais utilizadas como base (*backbone*) para extração de características e na combinação das características obtidas por meio do aprendizado de um *manifold* aliado a uma função de perda que objetiva discriminar imagens de classes diferentes.

Durante o processo de análise das abordagens de representação comparadas (referências) e propostas, percebemos ainda diferenças no desempenho das abordagens de classificação e de aprendizado de distâncias. Dentre as abordagens de classificação — incluindo as características originais *Concatenated*, o *ESS*, o *Fine-Tuned*, e o *Cross-Entropy* — as características do espaço *Cross-Entropy* proporcionaram melhores resultados durante a recuperação. Esse resultado confirma a hipótese de que o aprendizado da combinação das características dos componentes pode ser realizado com poucos dados de treinamento, especialmente por utilizarmos arquiteturas de rede pequenas. Esse resultado é reforçado pelos experimentos de comparação de profundidade/largura das redes (figuras 5.5, 5.6, 5.7, 5.8, 5.9 e 5.10).

Resultados ainda melhores foram obtidos pelos métodos baseados em aprendizado de distância — *Contrastive* e *Triplet* — que geraram espaços de características mais discriminativos (figuras 5.14 e 5.15), proporcionando grupos mais concisos de imagens Representativas e maior proximidade entre imagens Representativas e consultas (figuras 5.18 e 5.19). Essas conclusões são reforçadas pelos *rankings* obtidos (figuras 5.20, 5.21 e 5.22), nos quais as últimas duas colunas (*Contrastive* e *Triplet*) apresentam maior variabilidade de imagens recuperadas mantendo a precisão das imagens do topo.

Capítulo 6

Conclusões

O trabalho apresentado foi desenvolvido durante o mestrado e faz parte do Projeto Temático FAPESP *"DéjàVu: Feature-Space-Time Coherence from Heterogeneous Data for Media Integrity Analytics and Interpretation of Events"*. Um dos objetivos desse projeto é analisar e interpretar dados de eventos, principalmente, obtidos de redes sociais. Um grande desafio, anterior à interpretação, é a determinação de quais dados são realmente pertencentes e capazes de representar um evento de interesse. Esse desafio foi o foco deste trabalho e também foi discutido (juntamente com outros desafios encontrados no projeto *DéjàVu*) no artigo *"Forensic Event Analysis: From Seemingly Unrelated Data to Understanding"* submetido à revista *IEEE Security & Privacy*.

A tarefa de separação (filtragem) dos dados que pertencem a um evento (dados representativos) é manualmente inviável de ser realizada quando consideramos a grande quantidade de dados coletados. Além disso, o treinamento de técnicas completamente supervisionadas para a tarefa enfrenta a limitação de dados anotados.

Neste trabalho, decidimos restringir nossa análise a imagens. Portanto, a questão que tentamos responder é: *dado um evento, como separar automaticamente imagens Representativas de imagens Não-representativas?* Para responder a essa questão, focamos no desenvolvimento de técnicas de aprendizado de representação de imagens, por meio do aprendizado semi-supervisionado e supervisionado, utilizando o menor conjunto possível de imagens anotadas. Essas técnicas buscam facilitar a comparação por representatividade, melhorando o desempenho de abordagens de recuperação de imagens baseada em conteúdo.

Para a realização dos experimentos e avaliação dos métodos propostos, foi necessário construir conjuntos de dados anotados por representatividade. Coletamos e anotamos três conjuntos de dados de eventos específicos descritos no Capítulo 3. Um deles contendo imagens de um evento no contexto geral e os outros dois de eventos no contexto forense.

Primeiramente, testamos a hipótese de que a decomposição de eventos em componentes representativos poderia auxiliar na determinação de representatividade. De fato, ao utilizarmos características extraídas de CNNs pré-treinadas para representar os componentes *Places, Objects* e *People* obtivemos resultados promissores (Figura 4.7).

Nessa linha, propomos um método capaz de utilizar essas características de componentes e algumas imagens já determinadas como Representativas (ERIs) para gerar uma representação de baixa dimensionalidade chamada Espaço Semântico do Evento (ESS). A

representação ESS, apresentada no Capítulo 4, foi publicada nos anais do *IEEE International Workshop on Information Forensics and Security* (WIFS 2019), no artigo intitulado *"Image Semantic Representation for Event Understanding"* [49].

Com base nos resultados obtidos com o uso da representação ESS, percebemos a limitação de discriminabilidade das representações, principalmente pela falta de treinamento em eventos específicos. Quando combinamos os Componentes Representativos (RCs), não sabemos, *a priori*, a contribuição de cada um deles na descrição de eventos diferentes.

Por essa razão, testamos nossa segunda hipótese de que conjuntos pequenos de treinamento podem auxiliar no aprendizado da contribuição de cada RC para um evento. Para isso propomos o método Espaço Combinado do Evento (ECS) onde, com base em características extraídas por descritores que representam cada RC — no nosso caso CNNs pré-treinadas — aprendemos a combinar os RCs para representar imagens de um evento específico. Uma das desvantagens desse processo de aprendizado de combinação é a perda da interpretabilidade semântica dos RCs.

No entanto, percebemos uma melhoria significativa nos resultados com o uso do método. Durante os experimentos com diferentes arquiteturas de redes para combinação (variando profundidade e largura) e com diferentes funções de perda, percebemos que o aprendizado de um *manifold* para combinação dos componentes proporciona melhores resultados de precisão na tarefa de recuperação de imagens (figuras 5.16 e 5.17). Esse método, descrito no Capítulo 5, foi apresentado no artigo intitulado *"Manifold Learning for Real-World Event Understanding"* submetido à revista *IEEE Transactions on Information Forensics and Security*.

Avaliamos, assim, duas hipóteses que, com base nos experimentos realizados, se mostraram verdadeiras. A decomposição de eventos em componentes representativos auxilia na representação, assim como o apredizado da combinação de características desses componentes, orientado a uma pequena quantidade de dados do evento de interesse. Em especial, as representações obtidas pelas redes de combinação por aprendizado de distâncias — *Contrastive* (Figura 5.3) e *Triplet* (Figura 5.4) — mostraram-se mais discriminativas na tarefa de recuperação de imagens por representatividade.

Assim, com base nos dois métodos de representação propostos, percebemos um compromisso entre interpretabilidade e precisão. O método ESS surge como uma alternativa interpretável da proximidade entre imagens *Representativas* considerando características específicas de eventos, os RCs. Um evento poderia ser analisado em cada eixo do espaço semântico, observando a importância das características representando pessoas, lugares e objetos, por exemplo. No entanto, apesar da interpretabilidade, a contribuição de cada RC não é automaticamente inferida para qualquer evento, portanto existe uma limitação nos valores de precisão da recuperação de imagens. O método ECS surge então como forma de aprender a contribuição dos RCs, com um prejuízo para a interpretabilidade, melhorando a precisão na recuperação.

Em uma aplicação forense, os métodos poderiam ser utilizados em paralelo para: obter imagens *Representativas* com alta precisão utilizando o ECS; e explicar como essas imagens *Representativas* se relacionam entre si, destacando os eixos do ESS onde a proximidade semântica ocorre.

Destacamos então as principais contribuições deste trabalho de mestrado: a coleta

e anotação de três conjuntos de dados, de eventos em diferentes contextos; o desenvolvimento de métodos de representação de imagens para a tarefa de recuperação por representatividade, em cenários com poucas imagens anotadas; a proposta de redes de combinação de componentes para representação; e a avaliação e comparação de funções de perda de classificação e de aprendizado de distância.

Como projetos futuros visualizamos três vertentes principais: inclusão de novos componentes representativos, avaliação de qualidade de *rankings*, e explicabilidade no aprendizado das combinações. Acreditamos que, incorporando novos componentes que representem o evento, as características extraídas serão enriquecidas. Para isso, é também necessário estudar os eventos e encontrar descritores que possam ser utilizados pelos novos componentes de forma a acrescentar informação não redundante para as representações.

Como nosso objetivo é separar imagens relacionadas a eventos encontradas em mídias sociais, nós acreditamos que algumas das imagens Representativas já recuperadas poderiam auxiliar na recuperação de outras. Por isso, se avaliarmos a qualidade dos *rankings* obtidos, poderíamos obter indicações de quais possíveis imagens poderiam ser exploradas como consultas para próximas recuperações.

Por fim, temos ainda o interesse em entender a maneira como as redes de combinação aprendem a ponderar a influência de cada um dos componentes. Dessa maneira, poderíamos ser capazes de indicar fatores mais importantes para eventos. Assim, uma linha promissora seria trabalhar com técnicas de explicabilidade em aprendizado de máquina (X-AI).

Referências Bibliográficas

[1] Defense advanced research projects agency. `https://www.darpa.mil`. Acessado em: 26-06-2018.

[2] Ahmad Alzu'bi, Abbes Amira, and Naeem Ramzan. Semantic content-based image retrieval: A comprehensive study. *Journal of Visual Communication and Image Representation*, 32:20–54, 2015.

[3] André Arnes. *Digital Forensics*. Wiley, Norway, 1st edition, 2018.

[4] Sourour Brahimi, Najib Aoun, and Chokri Ben Amar. Boosted convolutional neural network for object recognition at large scale. *Neurocomputing*, 330:337 – 354, 2019.

[5] Kay H. Brodersen, Cheng Soon Ong, Klaas E. Stephan, and Joachim M. Buhmann. The balanced accuracy and its posterior distribution. In *20th International Conference on Pattern Recognition*, pages 3121–3124, 2010.

[6] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using "siamese" time delay neural network. *Advances in Neural Information Processing Systems*, 6:737 – 744, 1994.

[7] Tao Chen, Dongyuan Lu, Min-Yen Kan, and Peng Cui. Understanding and classifying image tweets. In *21st ACM Multimedia Conference (MM)*, pages 781–784, 10 2013.

[8] Hyun chong Cho, Lubomir Hadjiiski, Berkman Sahiner, Heang-Ping Chan, Mark Helvie, Alexis V. Nees, and Chintana Paramagul. Similarity evaluation between query and retrieved masses using a content-based image retrieval (CBIR) CADx system for characterization of breast masses on ultrasound images: an observer study. In *Medical Imaging 2011: Computer-Aided Diagnosis*, volume 7963, pages 715 – 722. International Society for Optics and Photonics, SPIE, 2011.

[9] Ricardo da Silva Torres and Alexandre X. Falcão. Content-based image retrieval: Theory and applications. *Latin-America Learning Technologies (RITA)*, 13:161–185, 2006.

[10] Duc-Tien Dang-Nguyen, Luca Piras, Michael Riegler, Liting Zhou, Mathias Lux, and Cathal Gurrin. Overview of ImageCLEFlifelog 2018: Daily Living Understanding and Lifelog Moment Retrieval. In *CLEF2018 Working Notes*, CEUR Workshop Proceedings, Avignon, France, September 10-14 2018.

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei Fei Li. Imagenet: a large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[12] Jiangfan Feng, Yuanyuan Liu, and Lin Wu. Bag of visual words model with deep spatial features for geographical scene classification. *Computational Intelligence and Neuroscience*, 2017:1–14, 06 2017.

[13] Michael Friendly and Daniel Denis. The early origins and development of the scatterplot. *Journal of the History of the Behavioral Sciences*, 41(2):103 – 130, 2005.

[14] Aditi Gupta, Hemank Lamba, and Ponnurangam Kumaraguru. 1.00 per RT #BostonMarathon #PrayForBoston: Analyzing fake content on Twitter. *eCrime Researchers Summit, eCrime*, pages 1 – 12, 09 2013.

[15] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *19th IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, page 1735–1742, USA, 2006. IEEE.

[16] Cameron Appel Hannah Ritchie, Joe Hasell and Max Roser. Terrorism. *Our World in Data*, 2020. https://ourworldindata.org/terrorism.

[17] Tin Kam Ho. Random decision forests. In *3rd International Conference on Document Analysis and Recognition (ICDAR)*, pages 278–282, Canada, 1995. IEEE.

[18] Amanda Hughes and Leysia Palen. Twitter adoption and use in mass convergence and emergency events. *International Journal of Emergency Management*, 6:248–260, 02 2009.

[19] Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. Processing social media messages in mass emergency: A survey. *ACM Computing Surveys*, pages 1–37, 06 2015.

[20] Bogdan Ionescu, Henning Müller, Mauricio Villegas, Alba García Seco de Herrera, Carsten Eickhoff, Vincent Andrearczyk, Yashin Dicente Cid, Vitali Liauchuk, Vassili Kovalev, Sadid A. Hasan, Yuan Ling, Oladimeji Farri, Joey Liu, Matthew Lungren, Duc-Tien Dang-Nguyen, Luca Piras, Michael Riegler, Liting Zhou, Mathias Lux, and Cathal Gurrin. Overview of ImageCLEF 2018: Challenges, datasets and evaluation. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Proceedings of the 9th International Conference of the CLEF Association (CLEF 2018), Avignon, France, September 10-14 2018. LNCS Lecture Notes in Computer Science, Springer.

[21] Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, Teddy Furon, and Ondrej Chum. Efficient diffusion on region manifolds: Recovering small objects with compact CNN representations. *30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.

[22] Philip T. Jackson, Amir Atapour-Abarghouei, Stephen Bonner, Toby P. Breckon, and Boguslaw Obara. Style augmentation: Data augmentation via style randomization. In *32th IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 83–92, 2019.

[23] Ian T. Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A*, 374(2065):1–16, 2016.

[24] Andrea L. Kavanaugh, Edward A. Fox, Steven D. Sheetz, Seungwon Yang, Lin Tzy Li, Donald J. Shoemaker, Apostol Natsev, and Lexing Xie. Social media use by government: from the routine to the critical. *Government Information Quarterly*, 29:480–491, 2012.

[25] Michael Peter Kennedy and Leon O. Chua. Neural networks for nonlinear programming. *IEEE Transactions on Circuits and Systems*, 35(5):554–562, 1988.

[26] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, pages 1–15, 12 2015.

[27] S. B. Kotsiantis. Supervised machine learning: A review of classification techniques. *Informatica*, 31(3):249–268, 2007.

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[29] Brian Kulis. Metric learning: A survey. *Foundations and Trends in Machine Learning*, 5(4):287 – 364, 2013.

[30] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, volume 86, pages 2278–2324, 1998.

[31] Xirong Li, Tiberio Uricchio, Lamberto Ballan, Marco Bertini, Cees G. M. Snoek, and Alberto Del Bimbo. Socializing the semantic gap: A comparative survey on image tag assignment, refinement, and retrieval. *ACM Computing Surveys*, 49(14):1–38, 2016.

[32] Zijia Lin, Guiguang Ding, Mingqing Hu, and Jianmin Wang. Semantics-preserving hashing for cross-view retrieval. In *28th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3864–3872, Boston, USA, 2015.

[33] Haiming Liu, Dawei Song, Stefan Rüger, Rui Hu, and Victoria Uren. Comparing dissimilarity measures for content-based image retrieval. In *Information Retrieval Technology*, pages 44–50, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

[34] James H. Liu, Rebekah Goldstein-Hawes, Denis Hilton, Li-Li Huang, Cecilia Gastardo-Conaco, Emma Dresler-Hawke, Florence Pittolo, Ying-Yi Hong, Colleen Ward, Sheela Abraham, Yoshihisa Kashima, Emiko Kashima, Megumi M. Ohashi, Masaki Yuki, and Yukako Hidaka. Social representations of events and people in world history across 12 cultures. *Jounal of Cross-Cultural Psychology*, 36(2):171–191, 2005.

[35] Stuart Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[36] David G. Lowe. Object recognition from local scale-invariant features. In *7th International Conference on Computer Vision (ICCV)*, volume 2, pages 1150–1157, 1999.

[37] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004.

[38] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv:1802.03426*, 2018.

[39] Kevin P. Murphy. *Machine learning: a probabilistic perspective*. MIT Press, 2013.

[40] Mor Naaman and Luis Gravano. Beyond trending topics: Real-world event identification on Twitter. In *5th International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 438–441, Barcelona, Spain, 2011.

[41] Dat Tien Nguyen, Firoj Alam, Ferda Ofli, and Muhammad Imran. Automatic image filtering on social networks using deep learning and perceptual hashing during crises. In *14th International Conference on Information Systems for Crisis Response And Management (ISCRAM)*, pages 499–511, 2017.

[42] Daniel Pedronette and Ricardo Torres. Unsupervised effectiveness estimation for image retrieval using reciprocal rank information. In *28th Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 321–328, 08 2015.

[43] Daniel Carlos Guimarães Pedronette, Lucas Pascotti Valem, Jurandy Almeida, and Ricardo da S. Torres. Multimedia retrieval through unsupervised hypergraph-based manifold ranking. *IEEE Transactions on Image Processing*, 28(12):5824–5838, Dec 2019.

[44] Robin Peters and João Porto de Albuquerque. Investigating images as indicators for relevant social media messages in disaster management. In *12th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*, pages 1–8, 2015.

[45] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. In *27th IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 512–519, 2014.

[46] Jérôme Revaud, Matthijs Douze, Cordelia Schmid, and Hervé Jégou. Event retrieval in large video collections with circulant temporal encoding. In *26th IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2459–2466, Portland, Oregon, USA, 2013.

[47] C. J. Van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edition, 1979.

[48] Hannah Ritchie and Max Roser. Natural disasters. *Our World in Data*, 2020. https://ourworldindata.org/natural-disasters.

[49] Caroline Mazini Rodrigues, Luis Pereira, Anderson Rocha, and Zanoni Dias. Image semantic representation for event understanding. In *11th IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2019.

[50] Frank Rosenblatt. The perceptron – a perceiving and recognizing automaton. Technical Report 85-460-1, Cornell Aeronautical Laboratory, 1957.

[51] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533 – 536, 1986.

[52] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[53] S. Rasoul Safavian and David Landgrebe. A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics*, 21(3):660 – 674, 1991.

[54] Surajit Saikia, Eduardo Fidalgo, Enrique Alegre, and Laura Fernández-Robles. Object detection for crime scene evidence analysis using deep learning. In *19th International Conference on Image Analysis and Processing (ICIAP)*, pages 14–24, Cham, 2017. Springer International Publishing.

[55] Amaia Salvador, Xavier Giró-i-Nieto, Ferran Marques, and Shin'ichi Satoh. Faster R-CNN features for instance search. In *29th IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 394–401, 2016.

[56] Veit Sandfort, Ke Yan, Perry J. Pickhardt, and Ronald M. Summers. Data augmentation using generative adversarial networks (cyclegan) to improve generalizability in ct segmentation tasks. *Nature Scientific Reports 9*, 9(16884):1 – 12, 2019.

[57] Walter J. Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772, 2013.

[58] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *28th IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823. IEEE, 2015.

[59] Joseph A. Shaw and Edward A. Fox. Combination of multiple searches. In *2nd Text Retrieval Conference (TREC)*, pages 243–252. NIST Special Publication 500-215, 1994.

[60] Guan-Lin Shen and Xiao-Jun Wu. Content based image retrieval by combining color texture and centrist. In *International Workshop on Signal Processing (SiPS)*, pages 1–4, London, UK, 2013.

[61] Connor Shorten and Taghi M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data 6*, 60:1 – 48, 2019.

[62] Nishant Shrivastava and Vipin Tyagi. Content based image retrieval based on relative locations of multiple regions of interest using selective regions matching. *Information Sciences*, 259:212–224, 2014.

[63] Neelabh Shanker Singh, S. Hariharan, and Monika Gupta. Facial recognition using deep learning. In *Advances in Data Sciences, Security and Applications*, pages 375–382, Singapore, 2020. Springer Singapore.

[64] Kate Starbird, Leysia Palen, Amanda Hughes, and Sarah Vieweg. Chatter on the red: What hazards threat reveals about the social life of microblogged information. In *10th ACM Conference on Computer Supported Cooperative Work (CSCW)*, pages 241–250, 10 2010.

[65] Yifan Suny, Liang Zhengz, Yi Yangz, Qi Tianx, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling and a strong convolutional baseline. In *15th European Conference on Computer Vision (ECCV)*, pages 1–17, Munich, Germany, 2018.

[66] Daniel Svozil, Vladimír Kvasnicka, and Jiří Pospichal. Introduction to multi-layer feed-forward neural networks. *Chemometrics and Intelligent Laboratory Systems*, 39(1):43–62, 1997.

[67] Christian Szegedy, Sergey Ioffe, and Vincent Vanhoucke. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations (ICLR)*, pages 1–14, San Diego, California, USA, 2015.

[68] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, Inception-ResNet and the impact of residual connections on learning. In *31st International Conference on Artificial Intelligence (AAAI)*, pages 4278–4284, San Francisco, California, USA, 2017.

[69] Ahmed Talib, Massudi Mahmuddin, Husniza Husni, and Loay E. George. A weighted dominant color descriptor for content-based image retrieval. *Journal of Visual Communication and Image Representation*, 24:345–360, 2013.

[70] Robert J Topinka. Terrorism, governmentality and the simulated city: the boston marathon bombing and the search for suspect two. *Visual Communication*, 15(3):351–370, 2016.

[71] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[72] Jesper E. van Engelen and Holger H. Hoos. A survey on semi-supervised learning. *Machine Learning*, 109:373–440, 2020.

[73] Anna Volokitin, Radu Timofte, and Luc Van Gool. Deep features or not: Temperature and time prediction in outdoor scenes. In *29th IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1136–1144, 2016.

[74] Jacob Walker, Abhinav Gupta, and Martial Hebert. Dense optical flow prediction from a static image. In *20th IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, page 2443–2451, USA, 2015. IEEE Computer Society.

[75] Xiang-Yang Wang, Yong-Wei Li, Hong-Ying Yang, and Jing-Wei Chen. An image retrieval scheme with relevance feedback using feature reconstruction and SVM reclassification. *Neurocomputing*, 127:214–230, 2014.

[76] Meihong Wu, Wenbin Xiao, and Zhiling Hong. Similar image retrieval in large-scale trademark databases based on regional and boundary fusion feature. *PLOS ONE*, 13(11):1–25, 11 2018.

[77] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. Lift: Learned invariant feature transform. In *7th European Conference on Computer Vision (ECCV)*, pages 467–483, 2016.

[78] Muhammad Yousuf, Zahid Mehmood, Hafiz Adnan Habib, Toqeer Mahmood, Tanzila Saba, Amjad Rehman, and Muhammad Rashid. A novel technique based on visual words fusion analysis of sparse features for effective content-based image retrieval. *Mathematical Problems in Engineering*, 2018:1–13, 2018.

[79] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *13th European Conference on Computer Vision*, pages 818–833. Springer International Publishing, 2014.

[80] Fan Zhang, Bolei Zhou, Carlo Ratti, and Yu Liu. Discovering place-informative scenes and objects using social media photos. *Royal Society Open Science*, 6, 03 2019.

[81] Qingchen Zhang, Laurence T. Yang, Zhikui Chen, and Peng Li. A survey on deep learning for big data. *Information Fusion*, 42:146 – 157, 2018.

[82] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *20th IEEE International Conference on Computer Vision (ICCV)*, pages 1116–1124, Santiago, Chile, 2015.

[83] Liang Zheng, Yi Yang, and Qi Tian. SIFT meets CNN: A decade survey of instance retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(5):1224–1244, 2018.

[84] Nanning Zheng and Jianru Xue. Manifold learning. In *Statistical Learning and Pattern Analysis for Image and Video Processing*. Springer, London, 2009.

[85] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Antonio Torralba, and Aude Oliva. Places: An image database for deep scene understanding. *Journal of Vision*, 17(10):1–12, 2017.

[86] Wengang Zhou, Houqiang Li, and Qi Tian. Recent advance in content-based image retrieval: A literature survey. *CoRR*, abs/1706.06064, 2017.