Classificação Multirrótulo Aplicada a Imagens Omnidirecionais

Manuel Veras
Instituto de Informática
Universidade Federal do Rio Grande do Sul
Email: manuel.veras@inf.ufrgs.br

Thiago L. T. da Silveira
Instituto de Informática
Universidade Federal do Rio Grande do Sul
Email: tltsilveira@inf.ufrgs.br

***Abstract*—Convolutional neural networks (CNNs) have been widely employed in computer vision problems, especially in applications involving conventional images captured using *pinhole* cameras. However, there is a growing demand for solutions capable of handling spherical images, and the successful adaptation of methods used for planar images to omnidirectional images is not a straightforward task. In this work, our goal is to perform a comparative analysis between two neural network architectures for multilabel classification applied to spherical images. The first network employs conventional convolutions, while the second incorporates spherical convolutions. Both were trained on a subset of the *Structured3D* dataset. Two experiments were conducted with the dataset: in the first experiment, we considered non-rotated ERP images, and in the second experiment, we used rotated ERP images, simulating inclined captures. We found that for both experiments the spherical CNN outperformed in all three metrics analyzed: *Hamming Loss* (HL), *Exact Match Ratio* (EMR), and F1-score.**

***Resumo*—Redes neurais convolucionais (CNNs) têm sido amplamente empregadas em problemas de visão computacional, especialmente em aplicações que envolvem imagens convencionais, baseadas em captura *pinhole*. No entanto, há uma crescente demanda por soluções capazes de lidar com imagens esféricas e a adaptação bem-sucedida de métodos utilizados em imagens planas para imagens omnidirecionais não é uma tarefa direta. Neste trabalho, nosso objetivo é realizar uma análise comparativa entre duas arquiteturas de redes neurais para a classificação multirrótulo aplicada a imagens esféricas. A primeira rede utiliza convoluções convencionais, enquanto a segunda incorpora convoluções esféricas. Ambas foram treinadas em um subconjunto da base de dados *Structured3D*. Foram feitos dois experimentos com o conjunto de dados: no primeiro experimento consideramos imagens ERP não-rotacionadas e no segundo experimento foram utilizadas imagens ERP rotacionadas, simulando capturas inclinadas. Constatamos que para ambos experimentos a CNN esférica obteve um desempenho mais satisfatório em relação as três métricas analisadas: *Hamming Loss* (HL), *Exact Match Ratio* (EMR) e F1-score.**

I. Introdução

Imagens omnidirecionais, também conhecidas como imagens em 360 graus, esféricas ou panoramas, têm um campo de visão completo ($180^\circ \times 360^\circ$) [1]. Diferentemente das imagens convencionais, que são baseadas no modelo de imageamento *pinhole* e possuem um campo de visão limitado, imagens omnidirecionais são definidas na superfície da esfera unitária e capturam intensidades de luz de toda a cena [2]. Imagens esféricas oferecem uma série de vantagens em relação às convencionais quando um campo de visão amplo é necessário. Por isso, essas imagens têm se tornado cada vez mais populares em aplicações inovadoras, como sistemas de vigilância inteligente, veículos autônomos e realidade aumentada [3].

Imagens omnidirecionais diferem das convencionais, exigindo atenção especial mesmo em tarefas como classificação de imagens. Classificação de imagens é uma tarefa fundamental no reconhecimento visual e tem como objetivo compreender e categorizar uma imagem sob um rótulo específico [4]. De forma semelhante, na classificação *multirrótulo*, uma instância pode estar associada a múltiplos rótulos [5]. Essa abordagem se diferencia da tarefa tradicional de classificação de único rótulo, tanto binário, como multiclasse, onde cada instância tem apenas um rótulo de classe [5].

Através da classificação multirrótulo, é possível identificar objetos presentes em uma imagem sem necessariamente localizá-los como em tarefas de detecção e segmentação [6]. As possíveis aplicações dessa tarefa incluem a busca de imagens, organização pessoal de fotos, gerenciamento de ativos digitais, reconhecimento de padrões em imagens médicas, etc. [7]. Há vários anos, muitas abordagens de sucesso aplicadas à classificação de imagens se baseiam em redes neurais convolucionais (CNNs) [8]. Essas redes utilizam a operação de convolução, que consiste em aplicar filtros com suporte fixo sobre a imagem para extrair características significativas [9].

Ao projetar sinais esféricos para o domínio planar, é possível aplicar as ferramentas convencionais para visão computacional [8]. A projeção equirretangular da esfera (ERP) é a representação padrão da esfera no plano, sendo amplamente usada na indústria e na pesquisa [10]. No entanto, uma imagem em formato ERP ("imagem ERP") sofre de distorções intensas – especialmente nas regiões polares [1]. Tal efeito resulta da amostragem não uniforme da esfera [2] e implica que um objeto tem aparência dependente de sua posição latitudinal. Isso apresenta um desafio para algoritmos modernos de visão computacional, tais como as CNNs convencionais, que não adaptam o suporte do filtro enquanto convolvem com a imagem [1].

Diferentes trabalhos propõem adaptar técnicas de aprendizado profundo em imagens para considerar as distorções introduzidas por um dado mapeamento esfera-plano [1], [8], [11]–[13]. Para tratar imagens ERP, as diferentes abordagens baseadas em CNN fazem com que o suporte do filtro trabalhe com uma amostragem uniforme na superfície esférica ao invés do plano. Embora existam diversas abordagens para tratar imagens ERP, ao melhor do nosso conhecimento, nenhuma avalia o problema de classificação multirrótulo. Ademais, sabe-se que a depender do problema, a contribuição da convolução esférica pode ser marginal ainda que levando a um aumento da complexidade de processamento [11]. Assim, o objetivo deste estudo é realizar uma análise comparativa entre CNNs convencional e esférica. Para nossa avaliação, consideramos CNNs simples e um subconjunto da base de dados *Structured3D* [14] com aproximadamente 5 mil imagens ERP fotorrealistas e anotações semânticas.

O restante deste artigo está organizado como segue. A Seção II revisa trabalhos relacionados às tarefas de classificação multiclasse e multirrótulo. Essa seção também discute convoluções aplicáveis a imagens ERP e, em mais detalhes, a abordagem adotada [1] na presente análise. A avaliação proposta é apresentada na Seção III. Experimentos e resultados são discutidos na Seção IV. Considerações

finais, incluindo trabalhos futuros, são feitas na Seção V.

II. Trabalhos Relacionados

Ao melhor do nosso conhecimento, não há trabalhos que lidam com a classificação multirrótulo de imagens omnidirecionais. Portanto, na Seção II-A, revisamos brevemente o estado da arte nesse problema com imagens convencionais. A classificação multiclasse de imagens esféricas está presente nos trabalhos [1], [15], [16] e é abordada na Seção II-B. A Seção II-C discute convoluções esféricas e dá enfoque particular no trabalho de Coors et al. [1].

A. Classificação Multirrótulo em Imagens Regulares

Uma imagem frequentemente contém múltiplos alvos de classificação, como objetos, o que requer recortes trabalhosos para criar anotações de um único rótulo [17]. Essas situação realista representa a classificação multirrótulo, na qual uma amostra de imagem não se limita a ter um único rótulo, mas sim vários rótulos simultaneamente.

Nos últimos anos, a pesquisa acadêmica de classificação multirrótulo vem se desenvolvendo no sentido de explorar correlações entre rótulos diferentes [18]. Isso geralmente é feito utilizando redes neurais de grafos (GNNs) que exploram os relacionamentos entre rótulos. Além disso, estratégias utilizando modelos recorrentes ou com estrutura *transformer encoder* também têm se popularizado [19].

Em Wang et al. [20], é proposto uma única rede *end-to-end* que combina uma rede neural recorrente (RNN) com CNN. Essa rede unificada de tipo CNN-RNN aprende uma incorporação conjunta de imagem-rótulo para caracterizar a dependência semântica entre rótulos, assim como a relevância entre imagem e rótulo. Os resultados experimentais em vários conjuntos de dados de referência mostram que essa abordagem alcança um desempenho superior em relação ao estado da arte.

Além das aplicações mencionadas na Seção I, bases de dados com múltiplos rótulos por imagem permitem o desenvolvimento de abordagens para detecção de objetos fracamente supervisionada (WSOD) [21]. O leitor interessado em "Uma revisão completa dos métodos para classificação multirrótulo em imagens e outros domínios pode ser encontrada na literatura [22].

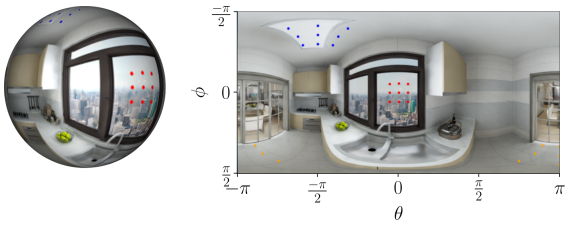
B. Classificação Multiclasse em Imagens Omnidirecionais

Ao examinar a literatura, não foram identificados estudos que abordam a classificação multirrótulo em imagens omnidirecionais. Entretanto, alguns trabalhos abordam o problema multiclasse [1], [15], [16].

Coors et al. [1] propõem adaptar o suporte das convoluções de tal forma que o kernel atue sobre regiões equiespaçadas na superfície da esfera. Os autores avaliam a abordagem em uma base de dados produzida a partir do MNIST, a qual eles denominam OMNI-MNIST. Nessa versão, os dígitos do conjunto MNIST são dispostos em planos tangentes à representação esférica da imagem, de onde são extraídas imagens planares. Coors et al. revelaram que as CNNs esféricas que eles propuseram tiveram melhores resultados quando comparadas à CNNs regulares atuando sobre imagens ERP (*EquirectCNN*) e mapas cúbicos da esfera (*CubeMapCNN*). A proposta de Coors e colegas também foi superior à proposta de [15], uma versão híbrida da *EquirectCNN* com *transformers* (*SphereTN*) e a GNN de [16].

Cohen et al. [15] propõem uma definição para a correlação cruzada esférica, que é simultaneamente expressiva e rotacionalmente equivariante. A partir disso, é desenvolvida uma CNN esférica que é adaptada para diferentes tarefas e cujos resultados numéricos confirmam uma alta acurácia e estabilidade do algoritmo. Mais especificamente, no trabalho de Cohen e colegas, são alcançados resultados próximos ao estado da arte em desafios de reconhecimento de modelos 3D e regressão de energia molecular.

Khasanova e Frossard [16] propõem usar representações baseadas em grafos e arquiteturas de aprendizado profundo para dados em grafos. Esses autores introduzem uma maneira fundamentada de construir grafos de modo que os filtros convolucionais respondam de forma semelhante para o mesmo padrão em diferentes posições da imagem, independentemente das distorções da lente. Por fim, os experimentos conduzidos em conjuntos de dados análogos ao MNIST, mas adaptados ao domínio esférico, demonstram que o método proposto por Khasanova e Frossard obtém resultados superiores em relação às redes agnósticas à geometria de imagens para o problema de classificação de imagens omnidirecionais.



(a) (b)

Figura 1. Amostragem do kernel de convolução em $\phi = 0$ (vermelho), $\phi = -\frac{3}{7}\pi$ (azul) e $\phi = \frac{2}{5}\pi$ (laranja) na (a) esfera e (b) imagem ERP.

Para este trabalho, nos baseamos na convolução esférica desenvolvida por Coors et al. [1] para desenvolver uma rede esférica capaz de lidar com o problema de classificação multirrótulos. Nota-se que o problema de classificação multirrótulos é consideravelmente mais complexo que o multiclasse [23]. No último, pode ser suficiente que um conjunto pequeno de informações locais seja determinante para o resultado – enquanto que no último não. Como exemplo, o leitor é convidado a pensar no problema de classificação binário que rotula imagens como *indoor* ou *outdoor*: identificar padrões de céu, árvore, lâmpada ou parede pode ser suficiente. Espera-se, portanto, que a contribuição das convoluções esféricas sejam mais efetivas na tarefa multirrótulo que na multiclasse. Na próxima subseção, introduzimos o conceito de convoluções esféricas e mostramos o método desenvolvido por [1].

C. Convolução Esférica em Imagens Omnidirecionais

A utilização de convoluções esféricas para lidar com problema de aprendizado profundo é vista em diferentes trabalhos [1], [8], [11], [12]. Tais trabalhos adaptam redes convencionais para lidar com diferentes tarefas, como detecção de objetos, segmentação de imagens, classificação multiclasse. A abordagem padrão consiste em desenvolver filtros de convolução que se deformam conforme a posição da imagem em que se encontram; e, dessa forma, permitem invariância às transformações geométricas induzidas pela ERP.

A seguir, apresentamos a formulação de convolução esférica proposta por Coors et al. que permitiu a implementação da CNN esférica utilizada para fins de comparação com uma CNN regular neste trabalho. Cohen et al. também se valem desses princípios para formulação de uma operação de *pooling* esférico.

Seja S a esfera unitária com S^2 sendo sua superfície. Cada ponto $s = (\phi, \theta) \in S^2$ é definido de forma única por sua latitude $\phi \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ e longitude $\theta \in [-\pi, \pi]$. Além disso, Π denota o plano tangente localizado em $s_\Pi = \phi_\Pi, \theta_\Pi$. Denotamos um ponto em Π por suas coordenadas $x \in \mathbb{R}^2$. O sistema de coordenadas local de Π é centrado em s e orientado verticalmente. Seja Π_0 o plano tangente localizado em $s = (0, 0)$. Um ponto s na esfera está relacionado às suas coordenadas no plano tangente x via uma projeção gnômica.

Sem perda de generalidade, considere um filtro de convolução 3×3. A forma do filtro de convolução é definida de modo que suas localizações de amostragem $s(j, k)$, onde $j, k \in \{-1, 0, 1\}$, se alinhem com os tamanhos de passo $\Delta\theta$ e $\Delta\phi$ da imagem ERP no equador. Com isso, pode-se amostrar a imagem em Π_0 sem interpolação usando

$$s(0, 0) = (0, 0), \tag{1}$$

$$s(\pm 1, 0) = (\pm\Delta\phi, 0), \tag{2}$$

$$s(0, \pm 1) = (0, \pm\Delta\theta) \tag{3}$$

e

$$s(\pm 1, \pm 1) = (\pm\Delta\phi, \pm\Delta\theta). \tag{4}$$

Para obter a posição dessas localizações de filtro no plano tangente Π_0 é utilizada a projeção gnômica. Assim, para o padrão de amostragem $s(j, k)$, isso resulta no seguinte padrão de filtro $x(j, k)$ em Π_0:

$$x(0, 0) = (0, 0), \tag{5}$$

$$x(\pm 1, 0) = (\pm \tan\Delta\theta, 0), \tag{6}$$

$$x(0, \pm 1) = (0, \pm \tan\Delta\phi) \tag{7}$$

e

$$x(\pm 1, \pm 1) = (\pm \tan\Delta\theta, \pm \sec\Delta\theta \tan\Delta\phi). \tag{8}$$

A forma do filtro na tangente é mantida fixa. Ao aplicar o filtro em uma localização diferente $s_\Pi = (\phi_\Pi, \theta_\Pi)$ da esfera, é aplicada a projeção gnômica inversa:

$$\phi(x, y) = \sin^{-1}\left(\frac{\cos\nu \sin\phi_\Pi + y \sin\nu \cos\phi_\Pi}{\rho}\right), \tag{9}$$

$$\theta(x, y) = \theta_\Pi + \tan^{-1}\left(\frac{x \sin\nu}{\rho \cos\phi_\Pi \cos\nu - y \sin\phi_\Pi \sin\nu}\right), \tag{10}$$

onde $\rho = \sqrt{x^2 + y^2}$ e $\nu = \tan^{-1}\rho$.

Assim, esse método permite que os filtros de convolução tangentes à esfera sofram distorção da mesma maneira que objetos em um plano tangente da esfera se distorcem ao serem projetados de diferentes elevações para uma representação de imagem ERP. A Figura 1 ilustra a distorção do filtro conforme sua latitude na imagem ERP.

Coors et al. [1] mostra que, através de sua abordagem, é possível que a amostragem dos filtros seja feita ao redor da esfera, através da fronteira da imagem. Isso elimina a descontinuidade presente ao processar imagens omnidirecionais com uma rede neural convolucional convencional e melhora o reconhecimento de objetos que estão divididos nas fronteiras laterais da imagem ERP.

III. Materiais e métodos

Essa seção é divida em duas partes. A Seção III-A apresenta o conjunto de dados utilizado para treino, teste e validação. A Seção III-B define a arquitetura dos modelos utilizados bem como seus hiperparâmetros.

A. Preparação do dataset

Neste trabalho, consideramos um subconjunto da base de dados *Structured3D* [14] – que contém, dentre outras informações, imagens coloridas e mapas semânticos associados. Mais precisamente, utilizamos as partições 0, 1, 2, 3 e 4 que somadas contabilizam um total de 4907 imagens. As imagens da base de dados *Structured3D* foram geradas de maneira sintética, estão no espaço de cores RGB e têm resolução de 256×512 pixels. As anotações dos mapas semânticos da base de dados totalizam 40 classes – de acordo com a taxonomia de . Zheng et al. [14] – com diferentes níveis de frequência, abrangendo desde objetos contáveis (*things*), como sofá, cama, geladeira, mesa e cadeira, até objetos não-contáveis (*stuff*), como chão, paredes e teto.

Dado o desbalanceamento entre as classes, variando de classes ausentes em várias imagens àquelas presentes em todas, foi realizada uma seleção de um subconjunto dessas classes. Especificamente, foram eliminadas classes com menos de 800 ocorrências ou com mais de 4000 ocorrências no conjunto de dados completo. Dessa forma, um total de 10 classes foi utilizado para conduzir o treinamento, validação e teste das redes. Estas classes são armário, cama, cadeira, sofá, mesa, janela, quadro, cortina, televisão e lâmpada.

(a) (b)

Figura 2. Exemplo de imagem extraída da base de dados *Structured 3D*: (a) imagem colorida e (b) mapa semântico. O mapa semântico indica as seguintes categorias: armário, janela, lâmpada, parede, chão, teto e porta. Dessas, são mantidas apenas as 3 primeiras em nossos experimentos.

Tabela I
Categorias e distribuições nas áreas 0 a 4 da base de dados *Structured3D*.

Categoria	Treino	Validação	Teste	Total
Armário	2226	479	475	3180
Cama	890	187	188	1265
Cadeira	1034	250	231	1515
Sofá	895	199	181	1275
Mesa	915	209	201	1325
Janela	2441	494	514	3449
Quadro	1547	346	347	2240
Cortina	1619	344	338	2301
Televisão	761	188	163	1112
Lâmpada	1824	420	385	2629
Total	17232	3916	3613	24781

Neste trabalho, convertemos os mapas semânticos em anotações categóricas considerando as 10 classes existentes. A Figura 2 exemplifica uma imagem da base de dados *Structured3D* e o seu mapa semântico correspondente. Aqui, as anotações são feitas considerando o padrão *one hot encoding*. Para avaliação, definimos uma divisão aleatória do conjunto de dados considerados em treino, teste e validação na proporção de 70%, 15% e 15%, respectivamente. A distribuição de instâncias por classe em cada conjunto de dados é dada na Tabela I.

Em nossos experimentos (confira a Seção IV) também consideramos uma versão "rotacionada" da base do *dataset* aqui descrito. Para tanto, aplicamos rotações aleatórias em cada imagem do conjunto de treino, teste e validação. As rotações são feitas variando os ângulos α, β e γ no intervalo de 0 a 30 graus. As imagens resultantes não estão mais perfeitamente alinhadas com o chão – o que simula capturas com câmeras levemente inclinadas [2].

B. Definição e treinamento dos modelos de classificação

Neste trabalho, consideramos duas arquiteturas *vanilla* de CNNs com mesmas configurações: uma rede esférica e uma rede convencional. A escolha por CNNs compostas basicamente por camadas convolucionais, de *pooling* e densas foi feita para avaliar o impacto dos dois primeiros componentes – sem interferência de outros como blocos residuais, conexões de atalho, mecanismos de atenção, etc.

As duas CNNs têm seis camadas de convolução intercaladas com operações de *pooling* de máximo. A CNN convencional utiliza convolução e *pooling* regulares, ao passo que a CNN esférica utiliza o ferramental revisado na Seção II-C e proposto em [1]. As duas redes aplicam a função de ativação ReLU em cada camada convolucional e

Tabela II
ARQUITETURAS DAS CNNS CONVENCIONAL E ESFÉRICA.

Camada	Entrada	Saída	Parâmetros
Convolução	3×256×512	6×256×512	168
Pooling	6×256×512	6×128×256	0
Convolução	6×128×256	12×128×256	660
Pooling	12×128×256	12×64×128	0
Convolução	12×64×128	24×64×128	2616
Pooling	24×64×128	24×32×64	0
Convolução	24×32×64	48×32×64	10416
Pooling	48×32×64	48×16×32	0
Convolução	48×16×32	96×16×32	41568
Pooling	96×16×32	96×8×16	0
Convolução	96×8×16	192×8×16	166080
Pooling	192×8×16	192×4×8	0
Densa	6144	3072	5907456
Densa	3072	10	30730

têm três camadas densas (totalmente conectadas) com 61.442, 3.072 e 10 neurônios – a qual é ativada por uma função sigmoide. Se a saída da função sigmoide é maior que λ para uma classe, então considera-se que a rede indica que o rótulo está presente na imagem. Consideramos $\lambda = 0{,}5$. A Tabela II mostra os detalhes de cada camada das CNNs (convolucionais, de *pooling* e totalmente conectadas). Note que as duas arquiteturas possuem exatamente a mesma quantidade de parâmetros: 19.129.678. Dada a natureza do problema, considerou-se a *binary cross entropy* [6] como função de perda. A arquitetura das CNNs é esquematizada na Figura 3.

Para o treinamento dos modelos, foram considerados *batch size* de tamanho 16, 100 épocas com parada antecipada (paciência de 10 épocas sem melhoras) e uma taxa de aprendizado de 10^{-4}. O otimizador Adam [24] é empregado neste trabalho.

IV. RESULTADOS EXPERIMENTAIS

Neste trabalho, usamos duas métricas para avaliar os resultados, a *Hamming loss* (HL) [25] e a *exact match ratio* (EMR) [25] que serão discutidas na seção IV-A. Os Resultados quantitativos e qualitativos, apresentados nas seções IV-B e IV-C, respectivamente, demonstram um desempenho superior da CNN esférica em relação a CNN convencional.

A. Métricas de Avaliação

A avaliação de algoritmos de classificação consiste em assinalar um escore para medir o quão distantes estão as previsões do algoritmo em relação aos rótulos reais, testadas em dados não vistos anteriormente. Quando lidamos com a avaliação de algoritmos de classificação multirrótulo temos a dificuldade adicional derivada do fato de que as predições podem ser parcialmente corretas [23], caso que ocorre quando apenas parte das classes presentes em uma instância são previstas corretamente pelo modelo.

Para avaliar o desempenho das redes, foi utilizada a *Hamming loss* (HL), definida como [25]

$$\text{HL} = \frac{1}{nk}\sum_{i=1}^{n}|Y_i \Delta Z_i|.$$

A operação Δ retorna a diferença simétrica entre Y_i, o conjunto real de rótulos da i-ésima instância, e Z_i, o conjunto previsto de rótulos da i-ésima instância. O operador $|\cdot|$ conta o número de valores 1 nessa diferença, ou seja, o número de previsões incorretas. O número total de erros nas n instâncias é agregado e normalizado, levando em consideração também o número de rótulos k [25].

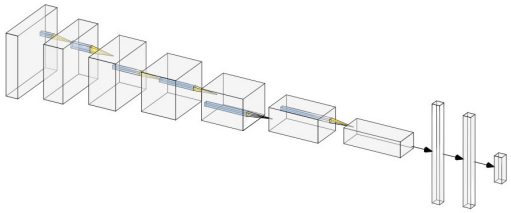


Figura 3. Arquitetura das CNNs utilizadas: a imagem de entrada é processada por seis camadas de convolução alternadas com *pooling* de máximo. Como resultado, são gerados mapas de características de resolução progressivamente menor e profundidade progressivamente maior. Após as camadas convolucionais, há três camadas totalmente conectadas.

A HL relata quantas vezes, em média, a relevância de um exemplo para um rótulo é prevista de forma incorreta. Portanto, a HL considera o erro de previsão (uma etiqueta incorreta é prevista) e o erro de omissão (uma etiqueta relevante não é prevista) [23]. Os dois erros são normalizados pelo número total de classes e pelo número total de instâncias [23]. Os valores da HL variam de 0 a 1, sendo que HL $= 0$ indica que não houve erros de classificação. Na prática, quanto menor o valor dessa métrica, melhor o desempenho do modelo [25].

O *exact match ratio* (EMR) [25] é uma métrica bastante estrita, uma vez que considera que uma instância foi predita corretamente apenas quando todos os rótulos verdadeiros foram preditos pelo modelo e que nenhuma classe verdadeira foi omitida pela predição [23]. O ERM pode ser calculada por [25]

$$\text{EMR} = \frac{1}{n}\sum_{i=1}^{n}[[Y_i = Z_i]],$$

onde o operador $[[\cdot]]$ denota o colchete de Iverson, que retorna 1 quando a expressão argumento é verdadeira e 0 caso contrario [25].

Além disso, é possível calcular as métricas usuais de classificação para cada classe de forma independente. A precisão é dada pelo número de rótulos preditos corretamente dividido pelo total de rótulos preditos e a revocação é dada pelo número de rótulos preditos corretamente dividido pelo número de rótulos reais [20]. A partir dessas duas métricas pode-se calcular o *F1-score* de cada classe, dado pela média geométrica da precisão e da revocação.

B. Resultados Quantitativos

O *dataset* Structured 3D contém apenas imagens alinhadas ao horizonte. Isso faz com que grande parte dos objetos em nosso esteja localizada em regiões de baixa latitude (próximas ao equador). Nestas áreas, a deformação é mínima, fazendo com que a vantagem da rede esférica em relação à rede convencional possa não ser tão evidente. Portanto, neste trabalho, consideramos dois experimentos:

- Experimento 1: treinamento, validação e teste das CNNs com imagens ERP não-rotacionadas (alinhadas ao horizonte).
- Experimento 2: treinamento, validação e teste das CNNs com imagens ERP rotacionadas (simulando capturas inclinadas).

A Tabela III apresenta os resultados de HL, EMR e *F1-score* médio para as CNNs convencional e esférica. Note que em ambos os experimentos, a rede esférica demonstra desempenho superior à rede convencional em todas as métricas avaliadas. Especificamente no experimento 2, em que são utilizadas imagens rotacionadas, esta superioridade da rede esférica torna-se ainda mais evidente. Acreditamos que isso ocorra devido às deformações mais acentuadas em objetos de imagens rotacionadas, com as quais a rede convencional parece ter dificuldade em lidar.

Tabela III
RESULTADOS EXPERIMENTAIS PARA CNNS CONVENCIONAL E ESFÉRICA. MELHORES RESULTADOS SÃO DESTACADOS EM NEGRITO.

Experimento	Métrica	CNN convencional	CNN esférica
Experimento 1	HL	0.2454	**0.2345 (-4.45%)**
	EMR	0.1440	**0.1671 (+16.04%)**
	F1-score	0.6420	**0.6730 (+4.83%)**
Experimento 2	HL	0.2789	**0.2547 (-8.67%)**
	EMR	0.1114	**0.1427 (+28.11%)**
	F1-score	0.5738	**0.6283 (+9.49%)**

(a) (b)

Figura 4. Resultados qualitativos considerando a CNN esférica em (a) com 10/10 e (b) 8/10 predições corretas e utilizando a CNN convencional em (a) com 7/10 e (b) 5/10 predições corretas. As categorias presentes nas imagens da esquerda e da direita são armário, janela e lâmpada, e armário, cadeira, sofá, mesa, janela, quadro, cortina, televisão e lâmpada, respectivamente.

C. Resultados Qualitativos

Por fim, a Figura 4 apresenta os resultados das CNNs convencionais e esféricas para duas imagens de teste do experimento 1. Em ambas as imagens a rede esférica obtém maior número de predições corretas quando comparada com sua contraparte convencional.

V. CONCLUSÃO

O presente trabalho realiza a comparação de uma CNN convencional com uma CNN esférica para a classificação multirrótulo de imagens esféricas em formato ERP. Nossa abordagem considerou 4907 imagens sintéticas com 10 possíveis classes. Os resultados experimentais para o conjunto de dados não rotacionados mostram que a CNN esférica apresenta uma redução de 4.45% da *Hamming Loss*, um aumento de 16.04% da EMR e um aumento de 4.83% de F1-score em relação a CNN convencional. Para o conjunto de dados rotacionados, a CNN esférica apresenta resultados ainda mais expressivos, com uma redução de $8,67\%$ da *Hamming Loss*, um aumento de 28.11% da EMR e um aumento de 9.49% de F1-score em relação a CNN convencional.

Para trabalhos futuros, planejamos ampliar a variedade de arquiteturas avaliadas, incluindo outras formulações para processamento de imagens ERP. Além disso, planejamos investigar como essas diferentes abordagens ativam regiões das imagens para cada classe utilizando de conceitos como *Class Activation Maps* [26]. Esse comportamento também pode ser explorado em problemas de detecção de objetos. Por fim, planejamos abordar o problema de segmentação panóptica utilizando o conjunto de dados *Structured3D* que já possui as anotações semânticas das classes.

AGRADECIMENTOS

Agradecemos o apoio financeiro da Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS), Brasil, do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brasil, e da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brasil – Código Financeiro 001, Brasil.

REFERÊNCIAS

[1] B. Coors, A. P. Condurache, and A. Geiger, "Spherenet: Learning spherical representations for detection and classification in omnidirectional images," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[2] T. Silveira and C. Jung, "Visual computing in 360°: Foundations, challenges, and applications," Natal/RN, 2022, in Portuguese: Anais do XXXV Congresso de Gráficos, Padrões e Imagens.

[3] S. Cho, R. Jung, and J. Kwon, "Spherical transformer," 2022.

[4] S. Wang and Z. Su, "Metamorphic testing for object detection systems," *CoRR*, vol. abs/1912.12162, 2019. [Online]. Available: http://arxiv.org/abs/1912.12162

[5] J. Read and F. Perez-Cruz, "Deep learning for multi-label classification," 2014.

[6] R. Szeliski, *Computer Vision*. Springer International Publishing, 2022. [Online]. Available: https://doi.org/10.1007/978-3-030-34372-9

[7] S. Liu, L. Zhang, X. Yang, H. Su, and J. Zhu, "Query2label: A simple transformer way to multi-label classification," 2021.

[8] N. M. Bidgoli, R. G. de A. Azevedo, T. Maugey, A. Roumy, and P. Frossard, "OSLO: On-the-sphere learning for omnidirectional images and its application to 360-degree image compression," *IEEE Transactions on Image Processing*, vol. 31, pp. 5813–5827, 2022. [Online]. Available: https://doi.org/10.1109%2Ftip.2022.3202357

[9] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[10] Y.-C. Su and K. Grauman, "Learning spherical convolution for fast features from 360deg $imagery$," 2018.

[11] C. Fernandez-Labrador, J. M. Facil, A. Perez-Yus, C. Demonceaux, J. Civera, and J. J. Guerrero, "Corners for layout: End-to-end layout recovery from 360 images," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1255–1262, 2020.

[12] C.-O. Artizzu, G. Allibert, and C. Demonceaux, "OMNI-CONV: Generalization of the Omnidirectional Distortion-Aware Convolutions," *Journal of Imaging*, vol. 9, no. 2, 2023. [Online]. Available: https://www.mdpi.com/2313-433X/9/2/29

[13] K. Tateno, N. Navab, and F. Tombari, "Distortion-Aware Convolutional Filters for Dense Prediction in Panoramic Images," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11220 LNCS, pp. 732–750, 2018.

[14] J. Zheng, J. Zhang, J. Li, R. Tang, S. Gao, and Z. Zhou, "Structured3d: A large photo-realistic dataset for structured 3d modeling," in *Proceedings of The European Conference on Computer Vision (ECCV)*, 2020.

[15] T. S. Cohen, M. Geiger, J. Köhler, and M. Welling, "Spherical cnns," *CoRR*, vol. abs/1801.10130, 2018. [Online]. Available: http://arxiv.org/abs/1801.10130

[16] R. Khasanova and P. Frossard, "Graph-based classification of omnidirectional images," *CoRR*, vol. abs/1707.08301, 2017. [Online]. Available: http://arxiv.org/abs/1707.08301

[17] T. Kobayashi, "Two-way multi-label loss," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[18] Y. Kim, J. M. Kim, Z. Akata, and J. Lee, "Large loss matters in weakly supervised multi-label classification," 2022.

[19] E. Ben-Baruch, T. Ridnik, N. Zamir, A. Noy, I. Friedman, M. Protter, and L. Zelnik-Manor, "Asymmetric loss for multi-label classification," 2021.

[20] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "CNN-RNN: A unified framework for multi-label image classification," *CoRR*, vol. abs/1604.04573, 2016. [Online]. Available: http://arxiv.org/abs/1604.04573

[21] F. Shao, L. Chen, J. Shao, W. Ji, S. Xiao, L. Ye, Y. Zhuang, and J. Xiao, "Deep learning for weakly-supervised object detection and localization: A survey," *Neurocomputing*, vol. 496, pp. 192–207, Jul. 2022. [Online]. Available: https://doi.org/10.1016/j.neucom.2022.01.095

[22] J. Bogatinovski, L. Todorovski, S. Džeroski, and D. Kocev, "Comprehensive comparative study of multi-label classification methods," *Expert Systems with Applications*, vol. 203, p. 117215, Oct. 2022. [Online]. Available: https://doi.org/10.1016/j.eswa.2022.117215

[23] M. S. Sorower, "A literature survey on algorithms for multi-label learning," 2010. [Online]. Available: https://api.semanticscholar.org/CorpusID:13222909

[24] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2017.

[25] F. Herrera, F. Charte, A. J. Rivera, and M. J. del Jesus, *Multilabel Classification*. Cham: Springer International Publishing, 2016, pp. 17–31. [Online]. Available: https://doi.org/10.1007/978-3-319-41111-8_2

[26] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.