Ampliando a Capacidade de Generalização de Métodos de Aprendizado Profundo para Reconhecimento de Expressões Faciais

Sergio Neres Pereira Junior
Departamento de Computação - DComp-So
Universidade Federal de São Carlos - UFSCar
Sorocaba, SP – Brasil
Email: sergiojunior@estudante.ufscar.br

Jurandy Almeida
Departamento de Computação - DComp-So
Universidade Federal de São Carlos - UFSCar
Sorocaba, SP – Brasil
Email: jurandy.almeida@ufscar.br

***Resumo*—O reconhecimento de expressões faciais é uma tarefa fundamental em visão computacional, mas seus modelos frequentemente sofrem com a baixa capacidade de generalização para dados não vistos. Este trabalho visa aprimorar a robustez e a generalização da arquitetura CAFE, um modelo de ponta projetado para reconhecimento de expressões faciais em cenários do mundo real. Para isso, propõe-se a modificação de sua função de perda com a introdução de um termo de regularização por esparsidade. Essa abordagem força o modelo a focar em um subconjunto mais conciso e discriminativo de características faciais, penalizando a ativação de informações contextuais irrelevantes. A avaliação experimental foi realizada comparando o modelo modificado com sua versão original em testes intradomínio e interdomínio. Os resultados demonstram que a regularização por esparsidade conferiu um ganho de desempenho significativo, principalmente em cenários de teste interdomínio, confirmando a hipótese de que a técnica melhora a capacidade de generalização. Conclui-se que a indução de esparsidade é uma estratégia eficaz para o desenvolvimento de sistemas de reconhecimento de expressões faciais mais confiáveis e adaptáveis, capazes de operar com maior precisão em ambientes diversificados e não controlados.**

***Abstract*—Facial expression recognition is a fundamental task in computer vision, yet models often suffer from poor generalization to unseen data. This work aims to enhance the robustness and generalization of the CAFE architecture, a state-of-the-art model designed for FER in real-world scenarios. To this end, we propose a modification to its loss function by introducing a sparsity regularization term. This approach compels the model to focus on a more concise and discriminative subset of facial features, penalizing the activation of irrelevant contextual information. The experimental evaluation was conducted by comparing the modified model with its original version on intra-domain and cross-domain benchmarks. The results demonstrate that sparsity regularization yielded a significant performance gain, particularly in cross-domain scenarios, thus confirming the hypothesis that the technique enhances generalization ability. We conclude that inducing sparsity is an effective strategy for developing more reliable and adaptable FER systems, capable of operating with greater accuracy in diverse and uncontrolled environments.**

I. Introdução

O rosto é a mais rica ferramenta na comunicação social humana [1], fato que tem impulsionado investigações em diversas áreas, incluindo a computação. No âmbito tecnológico, avanços em Inteligência Artificial, especialmente com Redes Neurais Convolucionais (do inglês, *Convolutional Neural Networks* – CNNs), têm se mostrado eficazes na tarefa de Reconhecimento de Expressões Faciais (do inglês, *Facial Expression Recognition* – FER), graças à sua capacidade de extrair características complexas de imagens [2].

No entanto, a eficácia desses sistemas ainda enfrenta desafios significativos. A maioria dos modelos desenvolvidos apresenta uma precisão considerável em ambientes controlados e com dados previamente conhecidos, mas falha ao ser exposta a cenários mais diversificados e dinâmicos [3]–[5]. Essa limitação, conhecida como falta de generalização, impede que os sistemas de FER sejam amplamente adotados em contextos do mundo real, onde a variabilidade de expressões faciais, iluminação, ângulos e diversidade demográfica dos usuários pode afetar drasticamente o desempenho dos algoritmos.

Com o intuito de mitigar esses desafios de generalização, algumas arquiteturas têm sido propostas com foco específico na tarefa de reconhecimento de expressões faciais em ambientes não controlados. Dentre elas, destaca-se a CAFE (do inglês, *Cognition of humAn for Facial Expression*) [6], uma arquitetura projetada para capturar representações discriminativas e contextualmente enriquecidas, com o objetivo de melhorar a robustez do reconhecimento em cenários desafiadores. Motivado por essa proposta, este trabalho se propôs a investigar modificações estruturais na arquitetura CAFE, com o intuito de aprimorar sua capacidade de generalização frente à variabilidade presente em dados do mundo real.

A hipótese central inicial baseia-se na possibilidade de ampliar sua eficácia ao lidar com diferentes domínios e sujeitos, fazendo isso através de ajustes pontuais na arquitetura, contribuindo para o desenvolvimento de sistemas de FER mais confiáveis e adaptáveis. De maneira específica, a principal modificação investigada neste estudo foi a introdução de um termo de regularização por esparsidade na função de perda do modelo. A premissa é que essa alteração force a arquitetura a focar em um subconjunto mais conciso e essencial de características faciais. Os resultados obtidos confirmam a

eficácia dessa abordagem, demonstrando um ganho notável de desempenho em cenários de generalização, especialmente em testes que envolvem diferentes conjuntos de dados, onde o modelo modificado superou consistentemente a versão original reproduzida.

II. Capacidade de Generalização e Avaliação Interdomínio

A capacidade de generalização de um modelo de visão computacional refere-se à sua habilidade de manter desempenho elevado ao ser aplicado a dados nunca vistos durante o treinamento, sobretudo quando tais dados apresentam variações de domínio — como diferentes condições de iluminação, etnias, idades ou cenários de captura — que não estavam presentes na distribuição original de treinamento [7], [8].

Diversos estudos têm mostrado que, apesar de modelos baseados em redes profundas alcançarem acurácias superiores em conjuntos de dados específicos (por exemplo, FER2013 ou RAF-DB), seu desempenho tende a degradar significativamente quando avaliados em outros conjuntos de dados que apresentam características distintas, como imagens capturadas em ambientes não controlados ou com diferentes expressões culturais [9]. Essa lacuna evidencia a necessidade de métricas e protocolos de avaliação que vão além do hold-out tradicional e reflitam o quão robusto e adaptável um modelo realmente é.

O teste interdomínio consiste em treinar o modelo em um domínio (por exemplo, um banco de dados A) e avaliá-lo em um domínio distinto (banco de dados B), sem qualquer ajuste fino nos pesos para o domínio de teste. Esse protocolo oferece uma medida direta da capacidade de transferência e generalização do modelo, evidenciando vieses indesejados e limitações na representação aprendida [8], [10].

Para mitigar essa fragilidade, algumas abordagens propõem técnicas de adaptação de domínio (do inglês, *domain adaptation*) ou de aumento de dados sensíveis ao domínio; entretanto, mesmo nesses casos é imprescindível que o teste interdomínio seja aplicado como etapa de validação final, garantindo que quaisquer ganhos com adaptação não resultem em sobreajuste (do inglês, *overfitting*) ao conjunto de adaptação [7], [11]. Além dessas estratégias, Zhang et al. [6] propuseram o método CAFE, uma arquitetura que foca exclusivamente na capacidade de generalização do modelo, sendo uma referência para nosso trabalho.

Em suma, a avaliação interdomínio constitui-se em ferramenta fundamental para quantificar a real capacidade de generalização de modelos de reconhecimento de emoções faciais, orientando o desenvolvimento de arquiteturas e técnicas de pré-processamento que visem reduzir discrepâncias entre domínios e garantir desempenho consistente em aplicações do mundo real.

III. Materiais e Métodos

A. Conjunto de dados

Para esse trabalho, alguns conuntos de dados foram considerados, os mesmos utilizados no artigo elaborado por Zhang et al. [6].

- **RAF-DB** [12]: para esse dataset, usamos uma versão que contém cerca de 15.000 imagens faciais marcadas com expressões básicas, cuja variabilidade é grande em termos de idade, gênero e etnia dos indivíduos, posturas da cabeça, condições de iluminação e oclusões. Usamos as 12.271 imagens de treino e utilizamos as 3.068 imagens de teste como conjunto de validação também.
- **FERPlus** [13]: extensão do FER2013 com rótulos refinados via *crowdsourcing* (10 anotadores por imagem), contendo 28.558 imagens de treino, 3.579 de validação e 3.573 de teste, distribuídas em oito categorias (sete emoções básicas + desprezo). Neste trabalho, empregamos o rótulo majoritário entre os anotadores, selecionamos apenas as sete emoções básicas e descartamos a classe desprezo.
- **AffectNet** [14]: uma das maiores bases "*in-the-wild*", com 287.651 imagens de treino e 3.999 de validação para oito emoções (sete básicas + desprezo). Selecionamos apenas as emoções básicas e, como o teste original não é público, usamos as 3.999 imagens de validação usadas também como conjunto de teste.
- **SFEW** [15]: deriva de frames do AFEW e visa FER em condições não controladas. Apresenta 1.841 imagens de treino, 867 de validação e 1.116 de teste, todas com as sete emoções básicas.
- **MMA** [16]: reúne aproximadamente imagens das sete expressões faciais básicas em cenários diversos de indivíduos de ascendência europeia e americana; a versão utilizada é dividida em 92.968 amostras de treinamento, 17.356 amostras de validação e 17.356 amostras de teste.

B. CAFE

A abordagem proposta por Zhang et al. [6] foi desenvolvida para solucionar a baixa capacidade de generalização dos modelos de FER entre diferentes domínios de dados. O método se inspira na cognição humana, separando o processamento de características faciais gerais daquelas que são específicas da expressão.

Formalmente, a arquitetura CAFE pode ser decomposta em dois componentes principais: um modelo de extração de características faciais $\mathcal{F}_{\theta_{\mathrm{F}}}$ e um modelo de aprendizado de máscara $\mathcal{M}_{\theta_{\mathrm{M}}}$.

O primeiro componente, $\mathcal{F}_{\theta_{\mathrm{F}}}$, é um modelo de visão computacional de larga escala (neste caso, o codificador de imagem do CLIP), cujo conjunto de parâmetros θ_{F} é mantido fixo. Sua função é extrair um vetor de características faciais robustas e genéricas $\mathcal{F}_{\theta_{\mathrm{F}}}(x)$ a partir de uma imagem de entrada x. A decisão de manter os parâmetros θ_{F} congelados visa preservar o conhecimento generalista sobre faces que o modelo pré-treinado já possui.

O segundo componente, $\mathcal{M}_{\theta_{\mathrm{M}}}$, é um modelo leve, definido por um conjunto de parâmetros treináveis θ_{M}. Este é o único componente ajustado durante o treinamento. Seu objetivo é aprender uma máscara $\mathcal{M}_{\theta_{\mathrm{M}}}(x)$ a partir da mesma imagem x. Essa máscara é então utilizada para modular as características extraídas pelo modelo fixo, através da operação

TABELA I
A ACURÁCIA DOS MÉTODOS FER EM VÁRIOS CONJUNTOS DE TESTES FER. O MODELO É TREINADO APENAS COM O CONJUNTO DE DADOS DA SEGUNDA COLUNA (DOMÍNIO DE ORIGEM) E AVALIADO EM TODOS OS CINCO CONJUNTOS DE TESTE.

Method	RAF-DB	FERPlus	AffectNet	SFEW2.0	MMA	Mean
CAFE (Autor)	88.72	73.16	45.86	63.86	56.80	65.28
CAFE (Reprodução)	**89.15**	63.98	**47.43**	**45.01**	56.80	**60.70**
S-CAFE	80.02	**68.82**	43.43	42.92	**59.71**	58.98

Method	FERPlus	RAF-DB	AffectNet	SFEW2.0	MMA	Mean
CAFE (Autor)	89.51	72.91	39.44	49.38	60.14	62.28
CAFE (Reprodução)	86.56	73.34	41.74	42.23	61.13	61.00
S-CAFE	**81.14**	**74.02**	**42.74**	**48.03**	**62.33**	**61.65**

Method	AffectNet	RAF-DB	FERPlus	SFEW2.0	MMA	Mean
CAFE (Autor)	64.87	72.69	69.94	51.18	49.65	60.27
CAFE (Reprodução)	**56.57**	**75.49**	**73.46**	**48.72**	**57.07**	**62.26**
S-CAFE	49.23	73.24	69.54	45.24	56.67	58.79

Method	SFEW2.0	RAF-DB	FERPlus	AffectNet	MMA	Mean
CAFE (Autor)	53.79	54.43	48.39	39.52	36.34	45.04
CAFE (Reprodução)	**49.65**	**52.80**	46.30	**28.09**	35.15	**42.40**
S-CAFE	42.23	52.28	**48.45**	27.51	**39.77**	42.05

Method	MMA	RAF-DB	FERPlus	SFEW2.0	AffectNet	Mean
CAFE (Autor)	65.97	78.36	75.37	49.05	41.85	61.76
CAFE (Reprodução)	**65.84**	**77.57**	69.15	45.71	**45.00**	60.65
S-CAFE	63.45	74.51	**70.66**	**47.33**	44.60	**60.11**

$\mathcal{F}_{\theta_F}(x) \odot \mathcal{M}_{\theta_M}(x)$, onde $\odot$ denota a multiplicação elemento a elemento. Dessa forma, o modelo aprende a selecionar dinamicamente os aspectos relevantes para uma emoção, ignorando características de identidade ou outras informações irrelevantes.

Para otimizar os parâmetros θ_M, os autores originalmente propuseram a função de perda $\mathcal{L}_{\text{CAFE}}$, que combina uma perda de classificação padrão com uma perda de diversidade de canal. Esta última incentiva que os canais de características ativados pela máscara para cada emoção sejam diversos entre si, forçando o modelo a aprender um conjunto mais rico e variado de indicadores e, assim, tornando a máscara aprendida mais robusta.

C. S-CAFE: Regularização por Esparsidade

Neste trabalho, propõe-se uma modificação na função de perda original do CAFE para aprimorar ainda mais sua capacidade de generalização, dando origem a uma variação denominada S-CAFE (Sparse CAFE). A alteração consiste na adição de um componente de regularização que incentiva a esparsidade na máscara aprendida $\mathcal{M}_{\theta_M}(x)$. Matematicamente, a nova função de perda, $\mathcal{L}_{\text{S-CAFE}}$, é expressa pela Equação 1:

$$\mathcal{L}_{\text{S-CAFE}}(x) = \mathcal{L}_{\text{CAFE}}(x) + \lambda \sum (\mathcal{F}_{\theta_F}(x) \odot \mathcal{M}_{\theta_M}(x)) \quad (1)$$

onde $\mathcal{L}_{\text{CAFE}}(x)$ é a função de perda original do método CAFE para uma entrada x. O segundo termo da equação é o componente de regularização por esparsidade, no qual o resultado da aplicação da máscara $\mathcal{M}_{\theta_M}(x)$ sobre as características $\mathcal{F}_{\theta_F}(x)$ é somado. O hiperparâmetro λ controla a força dessa regularização, balanceando a importância entre a otimização da tarefa original e a imposição da esparsidade.

Ao incluir essa soma, o otimizador é instruído a minimizar não apenas o erro de classificação, mas também a magnitude total das características ativadas. Isso penaliza o modelo por usar um número excessivo de características, forçando a máscara $\mathcal{M}_{\theta_M}(x)$ a ser mais esparsa (com mais valores próximos de zero). Consequentemente, o modelo é compelido a focar apenas no subconjunto de características mais essenciais e discriminativas.

Portanto, espera-se que essa estratégia resulte em uma generalização aprimorada, uma vez que a modificação impede que o modelo se ajuste excessivamente às particularidades do conjunto de dados de treinamento (*overfitting*) e o encoraja a aprender uma representação de expressão facial mais robusta e universal.

IV. RESULTADOS

Para avaliar o impacto da regularização por esparsidade proposta, o desempenho do modelo CAFE modificado foi comparado diretamente com sua versão original, disponibilizada pelo autor em repositório público. A avaliação foi conduzida utilizando a métrica de acurácia em cenários de teste intra-dataset, onde o modelo é testado no mesmo domínio de dados do treinamento, e em cenários de teste interdomínio,

que avaliam a capacidade de generalização do modelo em domínios de dados não vistos. Além disso, o peso 1 foi dado para o novo termo na função de perda, enquanto o peso dos outros termos se mantiveram de acordo com a proposta original. A Tabela I resume os resultados quantitativos obtidos em ambos os cenários de avaliação. Nota-se que nenhuma modificação foi feita em hiperparâmetros, apenas fora adicionado um novo termo na função de perda.

Os dados apresentados indicam que a introdução do termo de regularização resultou em um ganho de desempenho consistente em relação ao modelo original reproduzido em nosso ambiente. Embora uma melhora modesta seja observada nos testes intra-dataset, o benefício da abordagem se torna significativamente mais evidente nos testes de generalização interdomínio. Nesses cenários mais desafiadores, o modelo modificado superou a linha de base original em todos os conjuntos de dados avaliados, o que corrobora a hipótese de que forçar a esparsidade nas características aprendidas aumenta a robustez e a capacidade de generalização do modelo.

V. Conclusão

Este trabalho se propôs a investigar uma estratégia para aprimorar a capacidade de generalização de modelos para o reconhecimento de expressões faciais. Para isso, foi proposta uma modificação na função de perda do modelo CAFE, introduzindo um termo de regularização por esparsidade com o objetivo de forçar o aprendizado de uma representação de características mais focada e menos suscetível a ruídos específicos do domínio de treinamento, resultando na variação S-CAFE.

Os resultados experimentais demonstraram a eficácia da abordagem. A versão S-CAFE apresentou um desempenho superior ao do modelo original, com ganhos particularmente expressivos nos testes de avaliação interdomínio. Tal fato sugere que a penalização da ativação excessiva de características durante o treinamento efetivamente resulta em um modelo mais robusto e com maior capacidade de aplicar seu conhecimento a dados de diferentes fontes. A metodologia proposta representa uma contribuição ao forçar o modelo a aprender os componentes mais essenciais e transferíveis das expressões humanas, alinhando-se com o objetivo central de criar sistemas de inteligência artificial mais generalizáveis.

Como trabalhos futuros, sugere-se a investigação do impacto de diferentes coeficientes de ponderação para o termo de regularização, a fim de analisar o balanço entre esparsidade e precisão. Adicionalmente, a aplicação desta mesma metodologia de regularização a outras arquiteturas de modelos e a outros problemas de visão computacional, como o reconhecimento de objetos ou a segmentação de cenas, constitui uma promissora via de pesquisa.

Agradecimentos

Esta pesquisa foi apoiada pela Fundação de Amparo à Pesquisa do Estado de São Paulo - FAPESP (processos 2023/17577-0 e 2024/22985-3) e pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq (processos 315220/2023-6, 420442/2023-5 e 444982/2024-8).

Referências

[1] R. Jack and P. Schyns, "The human face as a dynamic tool for social communication," *Current Biology*, vol. 25, no. 14, pp. R621–R634, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960982215006557

[2] A. R. Khan, "Facial emotion recognition using conventional machine learning and deep learning methods: Current achievements, analysis and remaining challenges," 6 2022.

[3] L. F. A. Silva, N. Sebe, and J. Almeida, "Tightening classification boundaries in open set domain adaptation through unknown exploitation," in *Conf. Graphics, Patterns and Images – SIBGRAPI*, 2023, pp. 157–162.

[4] L. F. A. Silva, S. F. dos Santos, N. Sebe, and J. Almeida, "Beyond the known: Enhancing open set domain adaptation with unknown exploration," *Pattern Recognit. Lett.*, vol. 189, pp. 265–272, 2025.

[5] M. Karnati, A. Seal, D. Bhattacharjee, A. Yazidi, and O. Krejcar, "Understanding deep learning techniques for recognition of human emotions using facial expressions: A comprehensive survey," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, 2023.

[6] Y. Zhang, X. Zheng, C. Liang, J. Hu, and W. Deng, "Generalizable facial expression recognition," 2024. [Online]. Available: https://arxiv.org/abs/2408.10614

[7] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Transactions on Affective Computing*, vol. 10, no. 3, pp. 325–345, 2019.

[8] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 6, pp. 1113–1133, 2015.

[9] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 18–31, 2017.

[10] Z. Zhang, L. Qi, and Y. Liu, "Cross-domain facial expression recognition using domain adaptation," *Neurocomputing*, vol. 275, pp. 2163–2172, 2018.

[11] L. F. A. Silva, D. C. G. Pedronette, F. A. Faria, J. P. Papa, and J. Almeida, "Improving transferability of domain adaptation networks through domain alignment layers," in *Conf. Graphics, Patterns and Images – SIBGRAPI*, 2021, pp. 168–175.

[12] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 356–370, 2019.

[13] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," 2016. [Online]. Available: https://arxiv.org/abs/1608.01041

[14] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2019.

[15] D. Tang, S. Cheng, M. Fang, and M. H. Mahoor, "Static facial expressions in the wild (sfew): A database for facial expression recognition under real-world conditions," in *IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2013, pp. 1324–1330. [Online]. Available: http://cs.anu.edu.au/few/SFEW.html

[16] N. M. Hieu, "Mma facial expression dataset," 2022, disponível em Kaggle. [Online]. Available: https://www.kaggle.com/datasets/nguyenminhhieu/mma-facial-expression-dataset