

Redes Neurais Convolucionais para Avaliação de Dor Neonatal em Imagens de Face: Uma Análise Quantitativa e Qualitativa

Gabriel de Almeida Sá Coutrin¹, Carlos Eduardo Thomaz¹

¹FEI, Depto.de Engenharia Elétrica, São Bernardo do Campo, SP, Brasil

gcoutrin@outlook.com, cet@fei.edu.br

Abstract. *Pain experience may harm the development of newborns. The analysis of facial expressions is one of the most validated methods for neonatal pain assessment. Thus, this work investigates five CNNs for the classification of neonatal pain: VGG-16, ResNet50, SENet50, Inception-V3 and N-CNN. Our results indicate the superiority of models originally trained with face images, highlighting most relevant differences owing to the explainable information extracted by each model and to the current issue of limited neonatal face images available.*

Resumo. *A experiência da dor pode prejudicar o desenvolvimento de recém-nascidos. A análise da mímica facial é um dos métodos mais validados para a avaliação da dor neonatal. Assim, este trabalho investiga cinco CNNs para a classificação da dor neonatal: VGG-16, ResNet50, SENet50, Inception-V3 e N-CNN. Os resultados indicam a superioridade dos modelos pré-treinados com imagens de face, destacando diferenças relacionadas à informação extraída por cada modelo e a quantidade limitada de imagens de face neonatal disponíveis.*

1. Introdução

A dor experienciada por recém-nascidos (RNs) criticamente doentes está associada a perturbações em sua estabilidade respiratória, cardiovascular e metabólica, sendo estes fatores que elevam os índices de morbidade e mortalidade neonatal [Guinsburg 1999]. A avaliação da dor neonatal é realizada pela observação de indicadores fisiológicos e comportamentais, dentre os quais, a análise da mímica facial é considerada o método mais validado, proeminente e específico à dor [Guinsburg 1999, Heiderich et al. 2015].

Neste cenário, surgem diferentes propostas de sistemas computacionais para avaliação contínua e automática da dor neonatal utilizando expressões faciais [Zamzmi et al. 2019, Carlini et al. 2021, Gkikas and Tsiknakis 2023]. Em geral, estas soluções aplicam o aprendizado por transferência em Redes Neurais Convolucionais (CNNs) pré-treinadas para o reconhecimento de faces ou objetos. [Zamzmi et al. 2019] propuseram a Neonatal Convolutional Neural Network (N-CNN), a primeira arquitetura de CNN desenvolvida especificamente para a avaliação da dor neonatal. Em experimento com uma base de 3026 imagens de RNs, a N-CNN alcançou uma acurácia de 91%, superando, por exemplo, um método LBP+SVM tradicional (acurácia de 85,5%), mostrando não somente o potencial da arquitetura proposta, mas também a superioridade do aprendizado profundo na tarefa em questão.

Apesar destas metodologias implementarem diversas arquiteturas de CNN com alto desempenho de classificação, cada trabalho foi realizado com um procedimento de treinamento/teste diferente, utilizando bases de dados distintas, tornando inviável a

comparação direta do desempenho das arquiteturas. Neste contexto, o objetivo do presente trabalho é realizar uma comparação experimental de diferentes modelos de CNN (VGG-16, ResNet50, SENet50, Inception-V3 e N-CNN) aplicados na classificação automática de expressões faciais de dor em recém-nascidos, considerando resultados quantitativos e qualitativos, utilizando o método Grad-CAM [Selvaraju et al. 2017].

2. Materiais e Métodos

Nesta seção, são apresentados os bancos de imagens de face neonatal utilizados, seguidos pelos modelos de classificação e o protocolo de treinamento/teste. Ao final, argumenta-se sobre o Grad-CAM, o método implementado para a análise qualitativa das CNNs.

2.1. Bancos de Imagens de Face Neonatal

UNIFESP: Desenvolvido por [Heiderich et al. 2015], o Banco de Imagens de Face Neonatal da UNIFESP é composto por 360 imagens de resolução 320x233, pertencentes a 30 RNs. A base é dividida em 164 imagens de “dor” e 196 imagens “sem dor”.

iCOPE: Criada por Brahnem et al. [Brahnem et al. 2006], a base de dados *infant Classification of Pain Expression* (iCOPE) possui 200 imagens de face, com resolução de 3008x2000, de 26 RNs. Apenas as imagens rotuladas como “dor” e “repouso” foram utilizadas neste trabalho, que correspondem a 60 e 63 amostras, respectivamente.

2.2. Modelos de Classificação

Para a tarefa de classificação, foram analisadas as seguintes arquiteturas de CNN:

- **VGG-16:** A arquitetura VGG-16 [Simonyan and Zisserman 2014] pré-treinada com a base VGGFace (2,6 milhões de imagens) para o reconhecimento facial;¹
- **ResNet50:** A arquitetura ResNet50 [He et al. 2016] pré-treinada com a base VGGFace2 (3,3 milhões de imagens) para o reconhecimento facial;¹
- **SENet50:** A arquitetura “Squeeze-and-Excitation” ResNet50 [Hu et al. 2018] também pré-treinada com a base VGGFace2 para o reconhecimento facial;¹
- **Inception-V3:** A arquitetura Inception-V3 [Szegedy et al. 2016] pré-treinada com a base ImageNet (1,2 milhões de imagens) para o reconhecimento de objetos;
- **N-CNN:** A primeira CNN desenvolvida especificamente para o reconhecimento da dor neonatal em imagens de face, proposta por [Zamzmi et al. 2019] e implementada e treinada de ponta-a-ponta neste trabalho. O modelo possui 11 camadas e 72593 parâmetros.

Em cada CNN pré-treinada, as camadas convolucionais originais foram preservadas e novas camadas totalmente conectadas foram implementadas para a classificação da dor. Além disso, as últimas camadas convolucionais sofreram um processo de ajuste fino, ou seja, seus pesos também foram atualizados durante o treinamento com imagens neonatais. A Tabela 1 apresenta os modelos adaptados.

2.3. Protocolo de Treinamento e Teste

Inicialmente, as faces das imagens foram recortadas. As imagens recortadas destinadas ao treinamento dos modelos foram submetidas a um processo de *data augmentation*, gerando

¹Modelos pré-treinados disponíveis em <https://github.com/rcmalli/keras-vggface>

20 novas amostras para cada imagem pela aplicação aleatória das seguintes manipulações: rotação (30°), distorção (0,15), deslocamento horizontal e vertical (0,20), zoom (0,70 - 1,5), brilho (0,50 - 1,1) e inversão horizontal. Para cada nova amostra, verificou-se se a face do RN permanecia detectável (isto é, visível e dentro dos limites).

Tabela 1. Modificações dos modelos pré-treinado.

Modelo	VGG-16	ResNet50	SENet50	Inception-V3
Totalmente Conectada 1	512, ReLU	1000, ReLU	1000, ReLU	512, ReLU
Totalmente Conectada 2	512, ReLU	-	-	512, ReLU
Saída	2, Softmax	2, Softmax	2, Softmax	2, Softmax
Total de Parâmetros	27.823.938	25.612.154	28.143.146	23.115.554
Camadas Conv. em Ajuste Fino	6	9	9	18

Obs: depois de cada camada totalmente conectada há um *dropout* de 0,5.

A princípio, o treinamento é realizado somente para a atualização dos pesos das novas camadas totalmente conectadas, exceto para a N-CNN, na qual todas as camadas precisavam de treinamento. Para os modelos pré-treinados, após 5 épocas consecutivas sem redução do erro, inicia-se o ajuste fino das camadas convolucionais, conforme especificado na Tabela 1 apresentada anteriormente. Para os modelos durante o ajuste fino e a N-CNN, o treinamento é encerrado após 10 épocas consecutivas sem redução do erro.

Todas as CNNs foram treinadas e testadas com a união das bases UNIFESP e iCOPE, utilizando o protocolo *leave-sample-subjects-out*: a validação cruzada *k-fold* aplicada aos sujeitos da base, e não às amostras. Para o total de 56 sujeitos (30 da UNIFESP e 26 do iCOPE), 10 *folds* foram criados, cada um contendo as imagens pertencentes a 5 ou 6 sujeitos (sendo 3 necessariamente da UNIFESP para balancear as bases dentro dos *folds*). Esta abordagem evita o vazamento de dados (*data leakage*) e confere uma quantidade maior de imagens para teste (em comparação com o tradicional *leave-one-subject-out*).

2.4. Grad-CAM

O Grad-CAM (*Gradient-weighted Class Activation Mapping*), proposto por [Selvaraju et al. 2017], avalia os gradientes entre determinada camada convolucional e a saída do modelo e gera um mapa de ativação de classe, o qual destaca na imagem original as regiões de maior influência na classificação do modelo. Os autores do método sugerem utilizar a última camada convolucional devido ao melhor compromisso entre informação espacial e semântica de alto nível. O presente trabalho implementou o Grad-CAM conforme a descrição de [Selvaraju et al. 2017].

3. Resultados e Discussão

Nesta seção, é inicialmente feita a análise dos resultados quantitativos de cada modelo de classificação, ou seja, as métricas de avaliação. Em seguida, são apresentados os resultados qualitativos obtidos com o Grad-CAM. Por fim, a discussão dos resultados.

3.1. Análise Quantitativa

A Tabela 2 expõe as médias das métricas de avaliação de cada modelo. Observa-se que as CNNs VGG-16, ResNet50, SENet50 e Inception-V3 apresentaram desempenhos similares e superaram a N-CNN. As quatro redes pré-treinadas alcançaram um nível maior que

80% para todas as métricas, sendo que não há diferença estatística entre seus resultados (para $p = 0,05$). A VGG-16 apresentou a convergência mais rápida, utilizando 34 épocas em média. Para todas as CNNs, foi observado um elevado desvio padrão ($\geq 5\%$) em todas as métricas, sugerindo alta sensibilidade à escolha das amostras de treinamento e teste.

Tabela 2. Métricas de avaliação dos modelos.

Métricas	VGG-16	ResNet50	SENet50	Inception-V3	N-CNN
Acurácia	86,2% ± 7%	85,6% ± 7%	86,1% ± 5%	81,3% ± 5%	77,1% ± 7%
F1	87,7% ± 6%	87,0% ± 7%	87,3% ± 5%	84,1% ± 4%	80,8% ± 6%
AUC	85,4% ± 8%	85,2% ± 7%	85,7% ± 5%	80,6% ± 6%	76,0% ± 7%
Épocas	34 ± 7	44 ± 12	40 ± 10	47 ± 12	33 ± 17

3.2. Análise Qualitativa

A Figura 1 apresenta os resultados para cada CNN, considerando quatro exemplos da correta classificação de imagens de RN.

Base	Classe	Imagem	VGG	ResNet	SENet	IncepV3	N-CNN
iCOPE	Dor						
iCOPE	Sem Dor						
UNIFESP	Dor						
UNIFESP	Sem Dor						

Figura 1. Exemplos de resultados qualitativos gerados com o Grad-CAM.

Todos os modelos aprenderam a detectar regiões faciais para a avaliação da dor. A VGG-16 associa o estado “sem dor” apenas ao nariz e a região entre sobrancelhas. Quando “dor” é detectada, o modelo destaca também os sulcos nasolabiais e a frente do RN. Os modelos ResNet50 e SENet50 aparentam utilizar toda a face no processo de decisão. Porém, não está clara a relação entre regiões da imagem e classes. A Inception-V3 é capaz de reconhecer elementos faciais, mas apresenta elevada sensibilidade à presença de artefatos na imagem, como é o caso do ornamento no segundo exemplo (de cima para baixo). Por fim, a N-CNN aparenta relacionar o estado de “dor” com a boca aberta e o estado “sem dor” com os olhos abertos. Porém, seus mapas de ativação têm caráter esparsos e ruidosos, o qual se intensifica quando estes elementos faciais não estão presentes, dificultando a identificação de algum padrão de análise das expressões faciais.

3.3. Discussão

Os resultados quantitativos demonstraram a superioridade das redes VGG-16, ResNet50, SENet50 e Inception-V3 em relação a N-CNN, que pode ser atribuída ao fato das quatro

primeiras terem sido pré-treinadas: o conhecimento adquirido para a classificação de outras bases favorece o aprendizado de uma nova tarefa em um conjunto de dados reduzido.

Na análise qualitativa, VGG-16, ResNet50 e SENet50 superaram a Inception-V3, pois a última apresentou elevada sensibilidade à presença de artefatos nas imagens. Desta vez, a superioridade pode estar relacionada à natureza dos dados utilizados no pré-treinamento dos modelos. As três primeiras CNNs foram treinadas para o reconhecimento de faces e, conseqüentemente, já haviam aprendido filtros para a extração de características relevantes em imagens de face. A Inception-V3, por outro lado, foi pré-treinada com o banco de objetos variados. A partir dos resultados das três primeiras CNNs, observa-se que a visão holística da face utilizada pelos modelos ResNet50 e SENet50 dificulta a identificação da relação entre elementos faciais e classes. Em contrapartida, esta relação é nítida e plausível para VGG-16, apresentando, inclusive, concordância com escalas de dor empregadas por neonatologistas ao utilizar regiões como sulco nasolabial e a frente do RN [Tamanaka et al. 2022]. Desta forma, VGG-16 se destaca positivamente por também apresentar melhor interpretabilidade das informações discriminantes.

A N-CNN utilizada neste trabalho não possuía treinamento prévio e, certamente, a união das bases iCOPE e UNIFESP não foi suficiente para o treinamento adequado da CNN, visto que esta apresentou as menores métricas de desempenho e seus mapas de ativação de classe possuem caráter ruidoso. Em [Zamzmi et al. 2019], a N-CNN alcançou 91% de acurácia em experimento com um conjunto de 3026 imagens neonatais, o que evidencia o potencial desta arquitetura quando treinada com um conjunto de dados maior.

A falta de exemplos de treinamento é um fator limitante para a aplicação de redes neurais neste problema. Além das políticas de proteção de dados, a natureza da tarefa estudada dificulta a coleta de dados, pois o envolvimento de um procedimento doloroso implica em questões éticas. Outro ponto agravante é a falta de padronização das bases. Este trabalho uniu duas bases distintas, porém, é importante destacar uma possível incompatibilidade dos dados, devido às resoluções e aos métodos de rotulação distintos.

4. Conclusão

Este trabalho apresentou uma comparação experimental, baseada em resultados quantitativos e qualitativos, de cinco CNNs aplicadas à avaliação da dor neonatal. Nas condições da presente investigação, a VGG-16 proposta é o modelo mais adequado para a tarefa, uma vez que combina alto desempenho de classificação com melhor interpretabilidade e menor esforço computacional de treinamento.

Como trabalhos futuros, vislumbra-se a aplicação do Grad-CAM para geração de explicações contrafactuais [Selvaraju et al. 2017] para visualizar as áreas da imagem que reduziram a confiabilidade do modelo sobre a decisão realizada, reforçando as observações referentes à relação entre regiões e classes. Adicionalmente, vislumbra-se também a aplicação de outros métodos de Interpretação de Inteligência Artificial (IIA) para a compreensão das características extraídas pelos classificadores.

Os resultados deste estudo mostram potencial para aprimorar um aplicativo móvel capaz de reconhecer a dor neonatal de forma automática [Carlini et al. 2021]. Além da determinação da CNN com melhor desempenho para ser integrada à solução, o presente trabalho evidenciou que métodos de IIA podem permitir que o profissional de saúde

interprete a decisão do programa, conferindo maior segurança e melhor treinamento à aplicação do sistema na prática clínica em questão.

Agradecimentos

Ao apoio da FAPESP (2018/13076-9), CAPES, FEI e UNIFESP.

Referências

- Brahnam, S., Chuang, C.-F., Shih, F. Y., and Slack, M. R. (2006). Machine recognition and representation of neonatal facial displays of acute pain. *Artificial intelligence in medicine*, 36(3):211–222.
- Carlini, L. P., Ferreira, L. A., Coutrin, G. A. S., Varoto, V. V., Heiderich, T. M., Balda, R. C. X., Barros, M. C. M., Guinsburg, R., and Thomaz, C. E. (2021). A convolutional neural network-based mobile application to bedside neonatal pain assessment. In *2021 34th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 394–401.
- Gkikas, S. and Tsiknakis, M. (2023). Automatic assessment of pain based on deep learning methods: A systematic review. *Computer Methods and Programs in Biomedicine*, 231:107365.
- Guinsburg, R. (1999). Avaliação e tratamento da dor no recém-nascido. *J Pediatr (Rio J)*, 75(3):149–60.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- Heiderich, T. M., Leslie, A. T. F. S., and Guinsburg, R. (2015). Neonatal procedural pain can be assessed by computer software that has good sensitivity and specificity to detect facial movements. *Acta Paediatrica*, 104(2):e63–e69.
- Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141.
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.
- Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.
- Tamanaka, F. G., Carlini, L. P., Heiderich, T. M., Balda, R. C. X., Barros, M. C. M., Guinsburg, R., and Thomaz, C. E. (2022). Neonatal pain assessment: A kendall analysis between clinical and visually perceived facial features. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 0(0):1–10.
- Zamzmi, G., Paul, R., Goldgof, D., Kasturi, R., and Sun, Y. (2019). Pain assessment from facial expression: Neonatal convolutional neural network (n-cnn). In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE.