Facial Expressions Classification with *Ensembles* of Convolutional Neural Networks and Smart Voting

1

Abstract. *Facial Expression is a very important factor in the social interaction of human beings. And technologies that can automatically interpret and respond to stimuli of facial expressions already find a wide variety of applications, from anti-depressant drug testing to fatigue analysis of drivers and pilots. In this context, the following work presents a model for Automatic Classification of Facial Expression using as a training base the dataset Challenges in Representation Learning (FER2013), characterized by examples of spontaneous facial expressions in uncontrolled environments. The presented method is composed by a Convolutional Neural Networks Ensemble architecture, using a non-trivial voting system, based on a smart model, Xtreme Gradient Boosting - XGBoost. As performance criteria for validation of the proposed model, were used K-fold and F1 Score Micro techniques to guarantee robustness and reliability of the results, which are competitive with state-of-the-art works.*

Resumo. *A Expressão Facial é um fator de suma importância na interação social dos seres humanos. E tecnologias que podem interpretar e responder de forma automática a estímulos de expressões faciais já encontram uma grande variedade de aplicações, desde teste de fármacos anti-depressivos, até análise de fadiga de motoristas e pilotos. Neste contexto, o seguinte trabalho apresenta um modelo para Classificação Automática de Expressão Facial utilizando como base de treinamento o dataset Challenges in Representation Learning: Facial Expression Recognition Challenge(FER2013), caracterizado por exemplos de expressões faciais espontâneas em ambientes não controlados. O método apresentado é composto por uma arquitetura Ensemble de Redes Convolucionais Neurais, utilizando um sistema de votação não-trivial, baseado em um modelo inteligente, Xtreme Gradient Boosting - XGBoost. Como critérios de desempenho para validação do modelo proposto foram empregadas técnicas de K-fold e F1 Score Micro, para garantia de robustes e confiança dos resultados, que são competitivos com trabalhos estado-da-arte.*

1. Introdução

As expressões faciais são um componente não-verbal da comunicação humana, compreendendo sorrisos, franzir de sobrancelhas e outras ações, com vistas a transmitir emoções e sentimentos [Ekman and Friesen 1971]. Por meio do reconhecimento e interpretação de expressões faciais é possível obter um *feedback* mais verossímil e imediato da percepção dos indivíduos sobre um determinado contexto ou situação, razão pela qual é adotada, por exemplo, para avaliação do efeito de fármacos, detecção de fadiga de motoristas e pilotos, etc. [Fasel and Luettin 2003]. Dada a importância desta percepção, é natural a demanda pelo desenvolvimento de soluções automáticas que capturem o significado das expressões faciais de indivíduos.

Nesta perspectiva, a classificação automática de expressões faciais consiste em localizar faces humanas em uma cena, extrair características faciais da região detectada,

reconhecer padrões e então categorizar o resultado obtido em uma das sete expressões faciais canônicas: felicidade, tristeza, medo, nojo, surpresa, raiva e neutro [Pantic 2009]. Esta tarefa, entretanto, envolve a superação de diversos desafios, tais como as distintas formas que estas expressões podem ser manifestadas por indivíduos diferentes, a presença de outros elementos corporais, como as mãos, para composição da expressão, e até mesmo as características pessoais dos sujeitos, como presença ou ausência de barba, óculos, etc.

Com o advento dos métodos de *Machine Learning*, houve um progresso na classificação automática de expressões faciais. As técnicas de *Deep Learning*, em particular, têm colaborado para o avanço do estado da arte neste problema, este sucesso está relacionado a alta habilidade de representação das informações de entrada, utilizando várias camadas de neurônios artificiais compondo vários níveis de abstração[LeCun et al. 2015]. Porém, mais esforços precisam ser efetuados para uma melhor eficiência nesta tarefa.

Considerando o contexto apresentado, este trabalho se propõe a apresentar uma abordagem baseada no uso de Redes Neurais Convolucionais organizadas segundo um *ensemble* com votação mediada por um modelo de *Machine Learning* baseado em *Boosting*, para determinação das expressões correspondentes a partir de imagens de faces humanas. Os resultados obtidos com esta abordagem mostram-se animadores, pois obteve-se um desempenho de 71.74% nesta tarefa, equiparável com algumas das contribuições mais recentes da literatura neste problema.

Para apresentar os resultados obtidos, este trabalho está organizado como segue. A Seção 2 apresenta uma visão geral do estado da arte para a classificação automática de expressões faciais humanas com técnicas de *Deep Learning*. Em seguida, na Seção 3, são apresentados os materiais e métodos a serem considerados na elaboração da solução proposta. Os resultados obtidos e a discussão são então apresentados na Seção 4. Por fim, as considerações finais e perspectivas de trabalhos futuros são mostradas na Seção 5.

2. Trabalhos Relacionados

No campo da visão computacional a pesquisa de Classificação Automática de Expressões Faciais tem se mostrado bastante ativa. Muitos trabalhos tem aplicado os modelos de Redes Convolucionais Neurais na tarefa de Classificação de Expressões Faciais, o que pode ser observado pelos trabalhos relacionados, em que todos utilizam de alguma forma uma combinação com o modelo Convolucional, bem como o *Ensemble* destes. Além disto, a construção destes modelos seguem padrão de construção semelhante ao deste trabalho, descrito na Subseção 3.3, diferenciando-se apenas na última camada, Camada de Saída, por não fazerem uso da função *Softmax*, mas sim das especificidades de sua arquitetura.

A utilização de técnicas de pré-processamento bem como o emprego de filtros comumente utilizados na visão computacional para comparação de imagens, tem-se mostrado favoráveis no aumento do desempenho de modelos classificadores de expressões faciais. Exemplo de técnicas de pré-processamento pode ser visto em [Kim et al. 2016], onde é analisado o alinhamento da face contida na imagem, e mostrado que a correção desta condição se apresenta favorável ao desempenho de dectores automáticos. E exemplo de filtros de visão computacional podem ser vistos em [Al-Shabi et al. 2016], onde é analisado o uso combindado de SIFT, um descritor de imagem para correspondência e reconhecimento de pontos de interesse, e sua variação D-SIFT [Lindeberg 2012]. Contudo, [Pramerdorfer and Kampel 2016] apresenta e evidencia que resultados competitivos com o estado-da-arte podem ser obtidos mesmo sem o emprego de pré-processamento e

filtros de computacional sofisticados, mas também que modelos razos combinados ou não, para esta tarefa em específico, apresentam resultados igualmente competitivos. Estas comprovações são ainda reforçadas por [Tang 2013], o ganhador da competição em que a base de dados foi empregada publicamente pela primeira vez. Onde este não utilizou *Ensemble* de modelos de Redes Convolucionais, mas sim, uma única rede que possui na saída um modelo SVM Linear, onde comumente são empregados camadas *Fully Connected*.

Diferentemente das pesquisas anteriores, este trabalho utiliza nas Redes Convolucionais o mesmo padrão de arquitetura, bem como camadas de saída *Fully Connected* e função de ativação *Softmax* na última. Além disto, o *Ensemble* deste é realizado pelo modelo *XGBoost*, que é responsável por analisar e combinar os resultados para então classificar a entrada.

3. Materiais e Métodos

Esta seção descreve os materiais e métodos utilizados para desenvolvimento e análise dos experimentos conduzidos. As subseções abragem os seguintes tópicos: Dados Experimentais, que apresenta e analisa a base de dados utilizada; Descrição da Tarefa de Aprendizado de Máquina, que apresenta a divisão da base dados que foi realizada, para que fosse possível o emprego das métricas de desempenho escolhidas; Proposição de Modelos, que apresentam os tipos de modelos utilizados bem como a sua construção e emprego na Tarefa de Aprendizado de Máquina.g

3.1 Dados Experimentais

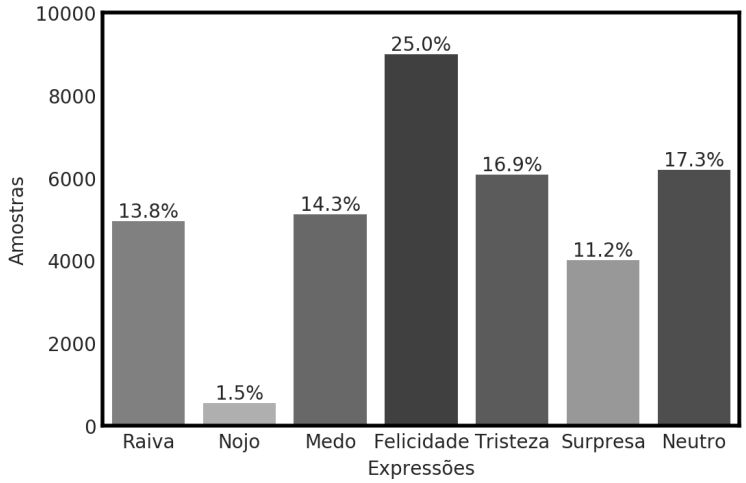
A base de dados de expressões faciais utilizada para o desenvolvimento deste trabalho é denominada *Facial Expression Recognition Challenge* (FER2013). Esta base contém 35.887 imagens faciais em escala de cinza com dimensões de 48×48 pixels, rotuladas de maneira supervisionada segundo uma das sete expressões faciais universais, conforme amostras ilustradas na Figura 1.

Figura 1: Amostras de imagens faciais da base de dados FER2013.

Conforme ilustra a Figura 1, é interessante notar algumas características particulares das imagens do FER2013 que ressaltam a relevância desta base de dados. Observa-se que, embora as faces estejam centralizadas nas imagens, elementos como cortes de cabelo, barba, óculos e até mesmo mãos encontram-se presentes, diminuindo a distância entre os exemplos contidos nesta base de dados e aqueles passíveis de ocorrência em um cenário realístico.

Os exemplos disponíveis na FER2013 se distribuem de maneira heterogênea perante as classes consideradas, conforme ilustra o gráfico da Figura 2. O número de exemplos rotulado com a expressão "nojo", por exemplo, representam apenas 1.5% do total de exemplos disponíveis. Estas características evidenciam o desbalanceamento do conjunto de dados considerado no tocante à quantidade de amostras por classe.

Figura 2: Distribuição de imagens por tipo de expressão facial na base FER2013.



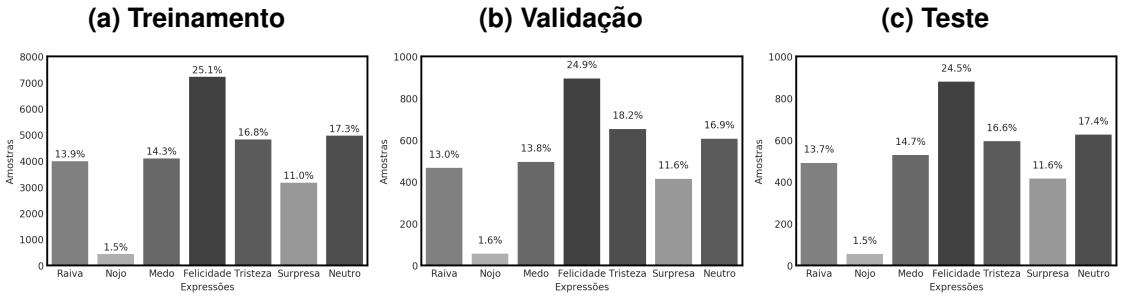
3.2 Descrição da Tarefa de Aprendizado de Máquina

A base de dados em questão será usada para realização de tarefa de classificação multirótulo segundo o paradigma de aprendizado supervisionado. Nesta tarefa, exemplos de expressões faciais e seus respectivos rótulos serão fornecidos previamente aos modelos de Aprendizado de Máquina para escolha e ajuste de parâmetros e realização do treinamento. Posteriormente, expressões faciais ainda não vistas serão apresentadas e o objetivo será avaliar o desempenho do modelo na classificação destes exemplos, isto é, aferir a respectiva capacidade de generalização.

Obedecendo a uma partição do FER2013 previamente considerada em competições de Visão Computacional [Kaggle 2013], esta base de dados será dividida em 3 partes, sendo: 80% dos exemplos para treinamento, 10% dos exemplos para validação e 10% dos exemplos remanescentes para testes, a serem utilizados seguindo uma abordagem de *holdout* de validação cruzada [Brink et al. 2017]. Apenas os exemplos da partição de testes serão utilizados para obtenção das métricas de desempenho e comparação dos modelos. Conforme ilustra a Figura 3, as partições preservam a distribuição de amostras por classe na base de dados original.

Levando em conta os modelos de Aprendizagem de Máquina para a tarefa em questão, é essencial que um número razoável de exemplos esteja disponível para um ajuste apropriado dos parâmetros treináveis, pois com uma base de dados pequena estes ficam propensos a *overfitting*, o que inibe a capacidade de generalização, em dados não vistos[Taylor and Nitschke 2017]. Considerando esta necessidade prática, os exemplos da partição de treinamento passaram por um processo de pseudo-expansão do tipo *data augmentation*, em que novas imagens foram geradas a partir das previamente existentes considerando transformações lineares de rotação, translação, escala e reflexão, colaborando para a posterior regularização dos modelos [Chollet 2017], tornando-os invariantes

Figura 3: Distribuição das classes nas partições adotadas para o conjunto de dados.



aos tipos de operações realizadas[Taylor and Nitschke 2017]. Que consistiram de rotação em módulo de até 10^{o}no sentido horário e anti-horário, translações de até 10 pixels nas direções ortogonais, fator de escala em módulo de 4 e reflexão em relação ao eixo das ordenadas. Ao final desta etapa, o conjunto de treinamento passou a conter 1.33×10^{15} exemplos, um aumento de 3.70×10^{10} vezes em relação ao seu tamanho original, mas preservando a distribuição de exemplos nas classes.

A métrica de desempenho adotada para comparação dos modelos na realização desta tarefa foi o Micro F-*Score*. Embora a acurácia seja uma métrica mais popular, que descreve o percentual de acertos do modelo em relação ao total de previsões efetuadas, não fornece detalhes acerca dos acertos por classe. Para contornar esta dificuldade, o Micro F-*Score* foi preferido, pois contempla a média harmônica entre precisão e revocação por classe ao passo que considera as diferentes frequências nas classes do problema [Kubat 2015]. Esta métrica é especialmente utilizada em problemas de classificação com classes desbalanceadas, ou seja, em situações análogas ao cenário considerado no escopo deste trabalho.

Dentre os modelos a serem avaliados, serão elencados como mais aptos para a tarefa de classificação proposta aqueles que maximizarem a métrica de desempenho Micro F-*Score* para os exemplos pertencentes à partição de testes.

3.3 Proposição de Modelos

Os modelos propostos e seus respectivos parâmetros e hiperparâmetros para a tarefa de classificação de expressões faciais são descritos detalhadamente ao longo desta seção.

Seguindo a abordagem predominantemente adotada pelo estado da arte no tocante ao aprendizado de características em dados de alta dimensionalidade para tarefas de Visão Computacional [Khan et al. 2018], as redes neurais convolucionais foram o modelo de Aprendizado de Máquina adotado na tarefa elencada. Em particular, a arquitetura base foi a da rede neural convolucional canônica VGG-16 [Simonyan and Zisserman 2015], mas com algumas adaptações. Esta arquietura originalmente proposta destacou-se mediante a ideia de que uma rede neural precisa ter uma quantidade razoável de camadas convolucionais profundas para uma representação hierárquica adequada das informações visuais.

As adaptações da VGG-16 levaram em conta diferentes quantidades de repetições

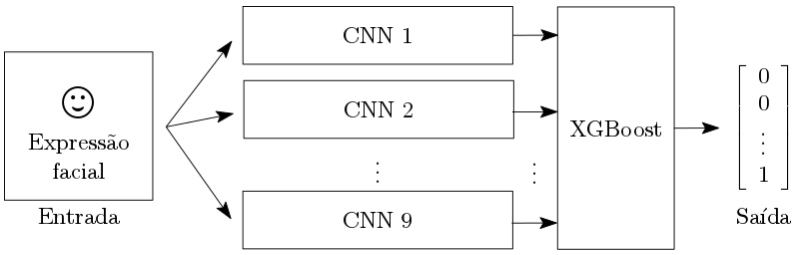
de certas operações, apresentadas de acordo com a seguinte ideia geral:

$$\begin{aligned} \text{Camada de Entrada} \Rightarrow [&(\text{Convolução} \rightarrow \text{Batch Normalization}) \cdot i \\ &\Rightarrow (\text{Pooling} \rightarrow \text{Dropout}) \cdot j] \cdot k \\ &\Rightarrow [\text{Fully Connected} \rightarrow \text{ReLu}] \cdot \ell \\ &\Rightarrow \text{Flatten} \Rightarrow \text{Camada de Saída}, \end{aligned} \tag{1}$$

em que i, j, k, ℓ são números inteiros que denotam a quantidade de repetições da operação associada perante multiplicação. Os valores destes inteiros foram: $i = 2$, $j = \{1, \ldots, 5\}$, $k = \{2, 3, 4\}$ e $\ell = \{1, 2, 3\}$. Considerando esta abordagem de adaptação, foram então propostas 9 CNNs diferentes a serem treinadas e testadas, conforme a abordagem de validação cruzada previamente descrita.

Além da avaliação individual do desempenho das CNNs propostas, considerou-se também a posterior combinação destas redes em um *ensemble*. Para valoração da classificação final, ao invés de considerar as abordagens típicas de votação unânime ou majoritária adotadas por *ensembles*, utilizou-se um modelo baseado em *Boosting*, o XGBoost [Chen and Guestrin 2016]. Este modelo recebe as saídas de todas as redes individuais e, após ter sido treinado com os exemplos da base de dados, decide dentre as classificações individuais qual a classificação final mais apropriada. Observa-se aqui uma modificação não-trivial em *ensembles*: a votação mediada por um modelo inteligente. A Figura 4 ilustra a ideia considerada.

Figura 4: *Ensemble* de CNNs com votação mediada por XGBoost.



4. Resultados e Discussões

Nesta seção são apresentados os resultados obtidos na metodologia proposta em classificar automaticamente expressões faciais, com base na partição de teste. Em seguida, é analisado o desempenho dos componentes do *Ensemble* na medida Micro *F1 Score*, para cada expressão e por fim é feita um comparativo com os resultados de trabalhos relacionados.

4.1 Classificação das Expressões Faciais

Para executar a avaliação do modelo proposto na tarefa de classificação, aplicou-se os exemplos contidos na partição de teste, de forma sequencial e supervisionada. Onde o rótulo fornecido pelo modelo para cada exemplo foi armazendo como resultado predito, para futura comparação com o rótulo verdadeiro, que é o contido na base dados.

Após o processo de rotulação das amostras de teste pelo modelo, foram fornecidas as informações de dados rotulados, juntamente dos rótulos verdadeiros, a medida Micro *F1 Score*, onde obteve-se como resultado o valor de 71.74%. Contudo, os trabalhos relacionados aqui apresentados fazem uso da medida de acurácia como resultados de desempenho, a qual foi obtida pelo mesmo método da medida Micro *F1 Score*, e resultou

em 71.74%. Resultados estes que superam os seres humanos nesta base de dados, que possuem acurácia de $65 \pm 5\%$ [Goodfellow et al. 2013]. Na Figura 5 é mostrada a Matriz de Confusão Normalizada, com os valores resultantes do método de rotulação, onde as linhas representam a classe verdadeira, e as colunas as classes preditas pelo modelo.

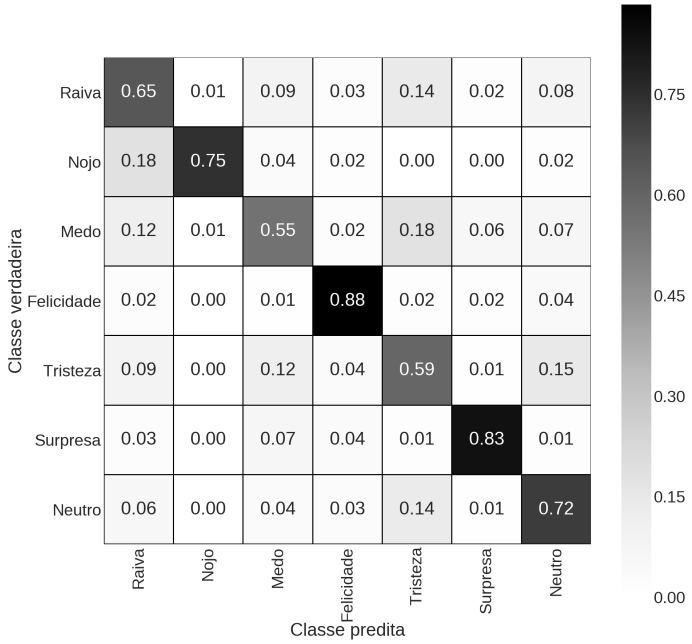


Figura 5: Matriz de Confusão do *Ensemble* (CNN + *XGBoost*)

Na diagonal principal da Matriz de Confusão Normalizada da Figura 5 é possível observar os fundos das células mais escuros, que indica alta densidade de valores nessas células. E este comportamento evidencia a classificação correta de bastantes elementos do conjunto de teste, que apesar do desbalanceamento da base de dados apresentou bons resultados em expressões com poucas quantidades de elementos, como é o caso da expressão de nojo que obteve um dos melhores resultados 1, acompanhadas de surpresa e felicidade, com *F1 Score* de 78%, 83%, e 89% respectivamente. As classificações das outras expressões também apresentaram bons resultados, todas com *F1 Score* maior ou igual a 58%.

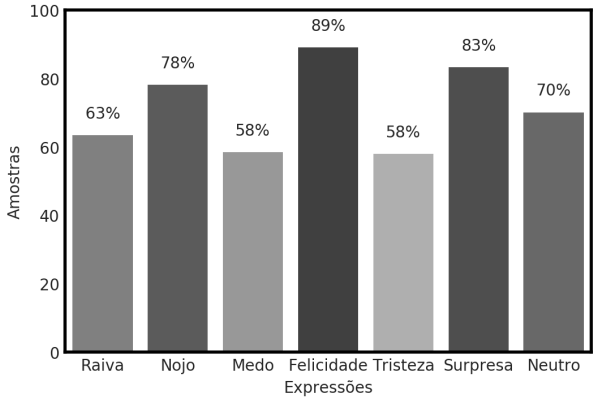


Figura 6: F1 Score por Expressão

Na Tabela 1 é possível observar o resultado geral do modelo proposto, acompa-

nhado dos modelos estado-da-arte. Evidenciando que o modelo proposto está um pouco abaixo do estado-da-arte, mas que ainda se mostra competitivo em relação a outros modelos da litetatura.

Tabela 1: Comparação de resultados com a Literatura

Modelo	Acurácia
[Pramerdorfer and Kampel 2016]	75.20
[Kim et al. 2016]	73.73
[Al-Shabi et al. 2016]	73.40
Proposto	71.74
[Tang 2013]	71.20

4.2 Desempenho dos componentes do *Ensemble*

O método utilizado para obtenção dos valores de Micro *F1 Score* e acurácia para o *Ensemble* foi o mesmo utilizado para os modelos de Rede Convolucional e *Xtreme Gradient Boosting*. Ressaltando que os resultados das medidas de desempenho obtiveram os mesmos valores, sendo este o motivo de serem apresentados somente os valores da medida Micro *F1 Score* na Tabela 2.

Tabela 2: Micro *F1 Score* e Parâmetros Treináveis do modelo proposto

Modelo	Micro F1 Score	Parâmetros Treináveis
1	0,6898857620507105	15.250.887
2	0,6767901922541097	1.778.247
3	0,6606297018668152	2.134.471
4	0,6798551128448036	1.778.247
5	0,6667595430482028	1.189.511
6	0,6781833379771524	1.780.103
7	0,6949010866536640	6.653.599
8	0,6285873502368348	1.186.231
9	0,6244079130677069	923.575
Ensemble	0,7174700473669546	64.000
Total		32.738.871

Como pode ser visto na Tabela 2, todos os modelos obtiveram bons resultados nesta tarefa de classificação, pois, considerando-se um palpite aleatório há 14.28% de chances de acerto de uma classificação correta, levando em consideração que 7 é a quantidade de classes desta tarefa, e que os modelos obtiveram pelo menos 62.44% como valor de Micro *F1 Score*. Apesar disto, alguns modelos se destacam, como são os casos dos modelos 7 e 1, com valores de 69.49% e 68.98% respectivamente, na métrica de desempenho.

Além disto, os modelos 7 e 1 também se destacam devido a serem os mais profundos e por terem as maiores quantidades de parâmetros treináveis, 6.653.599 e 15.250.887 respectivamente. Juntos, estes correspondem a 66.91% do total de parâmetros treináveis do *Ensemble*, que é de $32.738.871$. Observando esta relação, valor de Micro *F1 Score* com

quantidade de parâmetros treináveis, vê-se que os modelos mais profundos, consequentemente com maior número de parâmetros treináveis, foram os que obtiveram melhores desempenho na classificação geral. Enquanto os modelos mais rasos, apesar de obterem desempenho abaixo dos profundos, obtiveram os melhores resultados em classificar expressões específicas.

Outro fator interessante de ser observado em relação a quantidade de parâmetros é a ordem de grandeza do modelo responsável pelo sistema de votação. Que possui valor na ordem de algumas dezenas de milhares, enquanto os modelos de extração de características possuem valores na ordem de algumas dezenas de milhões. Este comportamento se mostra coerente, levando em consideração que os tipos das tarefas realizadas por cada modelo possuem complexidades bastante distintas, os modelos convolucionais responsáveis pela extração das várias características que compoem uma expressão facial humana, enquanto o modelo da votação combina da melhor forma possível, dados os parâmetros, as respostas dos modelos convolucionais.

5. Considerações Finais

O *Ensemble* de Redes Convolucionais utilizando o modelo *XGBoost* como sistema de votação apresentou bons resultados na tarefa de classificação, mostrando-se competitivo com modelos deste campo de pesquisa. Ressaltando que este resultado está atrelado exclusivamente ao desempenho na tarefa de classificação, e não quanto ao tempo processamento da resposta pelo modelo. Pois, este tempo é variado de acordo com as características do *hardware* utilizado, contudo, é evidente o alto uso de recursos computacionais pelo modelo proposto, visto que este possui em torno de 32 milhões de parâmetros nos qual a entrada deve ser processada para sua possível classificação. Visto isso, o emprego deste modelo em aplicações deve ser avaliado de forma cuidadosa.

Contudo, o uso deste modelo em aplicações com acesso a recurso de placa gráfica não deve enfrentar problemas com desempenho, visto que somente as camadas convolucionais possuem altas densidades de cálculo, enquanto as camadas de saída são razas, favorecendo o desempenho em relação a tempo de processamento. Além disto, acredita-se que seja possível melhorar a acurácia do modelo proposto, pois, não foi realizada busca exaustiva para escolha dos parâmetros, mas sim métodos empíricos para a definição destes.

Referências

[Al-Shabi et al. 2016] Al-Shabi, M., Cheah, W. P., and Connie, T. (2016). Facial expression recognition using a hybrid cnn-sift aggregator. *arXiv preprint arXiv:1608.02833*.

[Brink et al. 2017] Brink, H., Richards, J. W., and Fetherolf, M. (2017). *Real-World Machine Learning*. Manning Publications, Estados Unidos.

[Chen and Guestrin 2016] Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD'16, pages 785–794, New York, NY, USA. ACM.

[Chollet 2017] Chollet, F. (2017). *Deep Learning with Python*. Manning Publications, Shelter Island, New York, 1 edition.

[Ekman and Friesen 1971] Ekman, P. and Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2):124–129.

[Fasel and Luettin 2003] Fasel, B. and Luettin, J. (2003). Automatic facial expression analysis: a survey. *Pattern Recognition*, 36(1):259–275.

[Goodfellow et al. 2013] Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., et al. (2013). Challenges in representation learning: A report on three machine learning contests. In *International Conference on Neural Information Processing*, pages 117–124. Springer.

[Kaggle 2013] Kaggle (2013). Challenges in representation learning: Facial expression recognition challenge.

[Khan et al. 2018] Khan, S., Rahmani, H., Shah, S. A. A., and Bennamoun, M. (2018). *A Guide to Convolutional Neural Networks for Computer Vision*. Morgan and Claypool.

[Kim et al. 2016] Kim, B.-K., Dong, S.-Y., Roh, J., min Kim, G., and Lee, S.-Y. (2016). Fusing aligned and non-aligned face information for automatic affect recognition in the wild: A deep learning approach. *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1499–1508.

[Kubat 2015] Kubat, M. (2015). *An Introduction to Machine Learning*. Springer, Estados Unidos.

[LeCun et al. 2015] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436.

[Lindeberg 2012] Lindeberg, T. (2012). Scale invariant feature transform.

[Pantic 2009] Pantic, M. (2009). *Facial Expression Analysis*, volume 6, pages 400–406.

[Pramerdorfer and Kampel 2016] Pramerdorfer, C. and Kampel, M. (2016). Facial expression recognition using convolutional neural networks: State of the art. *CoRR*, abs/1612.02903.

[Simonyan and Zisserman 2015] Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations (ICLR 2015)*, San Diego, EUA.

[Tang 2013] Tang, Y. (2013). Deep learning using support vector machines. *CoRR*, abs/1306.0239.

[Taylor and Nitschke 2017] Taylor, L. and Nitschke, G. (2017). Improving deep learning using generic data augmentation. *CoRR*, abs/1708.06020.