Análise de Desempenho de Redes Neurais Convolucionais Aplicadas ao Reconhecimento de Emoções*

Leandro P. Heck[1], **Cristiano A. Künas**[1], **Edson L. Padoin**[1]

[1]Universidade Regional do Noroeste do Estado do Rio Grande do Sul (UNIJUI)
Santa Rosa – RS – Brasil

{leandro.h,cristiano.kunas}@sou.unijui.edu.br,
padoin@unijui.edu.br

Resumo. *Considerando o crescente interesse no campo da interação humano-computador e que essa interação vem se tornando algo cada vez mais natural e social, juntamente com o aumento da capacidade computacional proporcionada por GPUs e TPUs, áreas como o reconhecimento de emoções têm se mostrado ser de grande interesse e relevância pela comunidade científica. Porém, mesmo com diversos trabalhos realizados, detectar e reconhecer emoções computacionalmente com a mesma facilidade que humanos reconhecem, ainda é um problema relevante a ser explorado. Para tal, buscando explorar esse tema, este trabalho adotou a utilização de Redes Neurais Convolucionais (CNN) na realização do reconhecimento das emoções em humanos a partir de expressões faciais. Os resultados demonstraram que, com o treinamento da CNN em GPUs, foi possível reduzir o tempo computacional em até 94% e aumentar a acurácia para 66%.*

1. Introdução

Nos últimos anos, o reconhecimento das emoções se tornou um dos principais tópicos no campo da *Machine Learning* (ML) e da Inteligência Artificial (IA). Com o crescente aumento no desenvolvimento de tecnologias cada vez mais sofisticadas de interação homem-computador, impulsionou ainda mais o processo neste campo. As ações faciais transmitem as emoções que, por sua vez, transmitem a personalidade, o humor e as intenções de uma pessoa.

Existem inúmeras maneiras de expressar as emoções humanas, e estas vêm sendo estudadas ao longo dos anos, e diversas fontes de dados têm sido exploradas, como textos, envio de *emoticons*, voz e expressões faciais. Porém, as fontes que mais são utilizadas para realizar o reconhecimento de emoções de um indivíduo são as características faciais juntamente com a voz. No entanto, também existem outras características fisiológicas, que devem ser levadas em consideração, como características sociais, físicas do corpo, entre outras. Cada vez mais trabalhos são realizados para reconhecer as emoções com uma maior precisão e confiabilidade. A IA está revolucionando o campo da interação homem-computador, fornecendo várias técnicas de aprendizagem para o reconhecimento de padrões.

Existem diversas técnicas de ML para reconhecer a emoção, mas este trabalho se concentrará principalmente no reconhecimento de emoções baseado em imagem/vídeo. Reconhecendo as 6 emoções consideradas básicas por Ekman (1973), que são felicidade, tristeza, medo, surpresa, raiva, nojo, mais a emoção neutra (caso não reconheça nem uma das demais emoções).

*Trabalho desenvolvido com recursos do edital MCTIC/CNPq - Universal 28/2018 sob número 436339/2018-8 e do edital da VRPGPE bolsa PIBIC/UNIJUI.

Para a extração das características, será utilizada uma Rede Neural Convolucional (CNN - *Convolutional Neural Network*). O principal objetivo é realizar a análise e comparação dos tempos da CNN nas diferentes arquiteturas de CPU, GPU e TPU da plataforma do Google Colab. Além de observar o desempenho obtido pela CNN nas diferentes arquiteturas.

O restante do trabalho está organizado da seguinte forma. A Seção 2 discute os trabalhos relacionados. Na Seção 3 é apresentada a metodologia que será utilizada na implementação e o ambiente de execução para realização dos testes. Na Seção 4 são apresentados os resultados obtidos pela CNN, seguidos das Conclusões e Trabalhos Futuros.

2. Trabalhos Relacionados

No trabalho de Bartlett *et al.* (2003) é apresentado um estudo com o objetivo de localizar automaticamente faces em um fluxo de vídeo e codificar a expressão visual de maneira dinâmica. Os autores discutem que a comunicação face a face é uma operação em tempo real e com uma escala de tempo em 40 milissegundos. O sistema é capaz de detectar as emoções, possuindo um diferencial de outros trabalhos pois opera em tempo real. Este sistema foi treinado e testado utilizando a base de dados *CohnKanade AU-Coded Expression Database* [Kanade and Cohn 2005]. Esta base de dados contém o registro facial de 210 adultos na faixa etária entre 18 e 50 anos de idade. Os experimentos realizados compararam o desempenho do reconhecimento da abordagem de detecção automática com a abordagem de detecção manual, não encontrando nenhuma diferença significativa entre elas. O sistema apresentou um nível de precisão de 93% de reconhecimento, na seleção de uma das 7 opções de expressões faciais.

O trabalho desenvolvido por Tang and Huang (2008), tem como objetivo reconhecer as seis emoções básicas e universais através de expressões faciais utilizando da geometria 3D. Esta abordagem extrai características que são invariantes sob efeito de iluminação ou postura, características que os autores consideram como obstáculos para o reconhecimento facial em imagem 2D. Neste trabalho foi utilizado, como base de treinamento e teste, a base de dados *BU-3DFE* [Yin et al. 2006]. Esta base de dados é composta por 100 indivíduos, sendo que 60% do sexo feminino e 40% do sexo masculino. Para este trabalho foi constatado que a abordagem produziu um aumento absoluto de 3,5% na taxa média de reconhecimento. Esta abordagem obteve uma precisão média de 87,1%, sendo que a maior taxa foi de 99,2% para o reconhecimento da expressão facial da surpresa.

No trabalho desenvolvido por Amin *et al.* (2017), elaborou-se uma RNA que possui a finalidade de reconhecer emoções através de expressões faciais utilizando a técnica de aprendizado profundo. Segundo o autor, a utilização de redes neurais convolucionais, na abordagem de identificação de emoções, atinge uma precisão média de 60%. Para o desenvolvimento do trabalho foi utilizado a base de dados *Facial Expression Recognition 2013 (FER-2013)* [Carrier et al. 2013]. O trabalho apresentou bons resultados, alcançando uma precisão média de 61,05% para a classificação das emoções. Analisando os resultados, os autores constataram que a emoção de alegria possui a maior taxa de reconhecimento.

3. Metodologia

Para a implementação da CNN proposta utilizou-se da linguagem de programação *Python*. As principais bibliotecas utilizadas são o *TensorFlow 2.0* e o *Keras*, utilizadas na aprendizagem profunda da CNN. Também utilizou-se das ferramentas: *OpenCV*, utilizado para realizar o processamento das imagens; *Scikit-Learn*, utilizado para a análise dos dados; *Numpy* e *Pandas*,

utilizadas para a manipulação e análise dos dados; *Matplotlib*, utilizado para criação de gráficos e visualização dos dados em geral. A base de dados utilizada neste trabalho foi *FER-2013* (*Facial Expression Recognition 2013*) [Carrier et al. 2013].

Para a criação do modelo da RNA, é definido um número total de filtros (*num_features*) como 50, a divisão do trabalho, no caso, para realizar os ajustes dos pesos da CNN (*batch_size*) como 25, e um total de 100 épocas, para realizar o treinamento. Também é definida uma métrica de parada (*EarlyStopping()*), no período de 15 épocas. O modelo da CNN possui uma sequência de quatro camadas de convolução, em cada uma é utilizando a função *Conv2D*, que recebe o total de filtros. Recebendo um *kernel_size* de tamanho 3x3, que percorre toda a imagem captando os seus traços mais relevantes, criando o mapa de características (*feature map*).

Na primeira camada convolucional é utilizada a função de ativação ReLU, já na segunda a ELU, na terceira novamente a ReLU e na quarta a ELU. Optou-se pelo uso da função ReLU por não ativar todos os neurônios ao mesmo tempo, e na próxima camada utiliza-se da função ELU, que resolve o problema das unidades 'mortas' apresentado pelas ReLUs. Também é usado a função de *MaxPolling2D*, que recebe um *kernel_size* de tamanho 2x2, que retorna o maior número da unidade, e passa este valor como saída. Cada camada convolucional possui um *Dropout* de 20%. Depois de realizado os processos anteriores é chamada a função de *Flatten()*. A classe *Model* da API funcional da biblioteca *Keras* é utilizada no desenvolvimento do modelo de dados. A compilação do modelo configura o processo de aprendizado. Sendo definido o otimizador (*adam*), a função de perda (*binary_crossentropy*) e as métricas (*accuracy*).

O ambiente de execução utilizado foi a plataforma do *Google Colab*, um serviço de nuvem gratuito, hospedado pelo *Google*, por possuir aceleradores de GPU grátis e bibliotecas já pré-instaladas. A GPU utilizada foi a NVIDIA Tesla T4, que possui arquitetura NVIDIA Pascal, possuindo 2.560 CUDA *Cores*, com 5.5 TFLOPS de Desempenho de Precisão Única

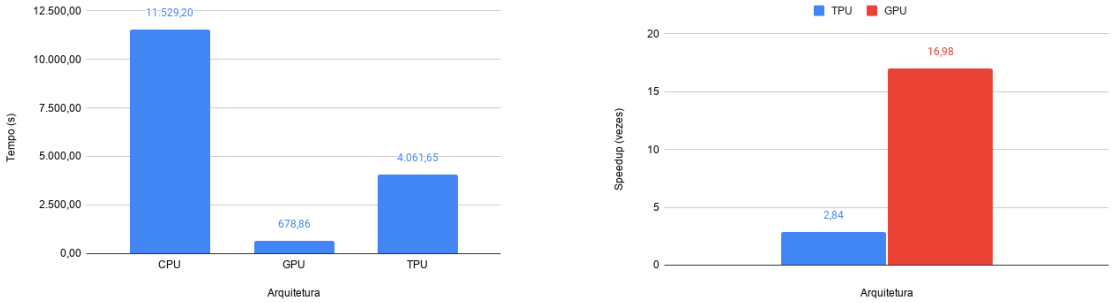
4. Resultados

Realizou-se 10 execuções em cada uma das arquiteturas, assim calculando uma média dos tempos de cada execução para obter o *Speedup*. Os resultados das médias dos tempos de execução, são apresentados na 1(a). O *Speedup* alcançado pela CNN comparado a execução em CPU, é apresentado na Figura 1(b). O *Speedup* do algoritmo executado em GPU apresentou um ganho de $16,98$ vezes sobre a CPU. Reduzindo o tempo de execução de $11529,20$ segundos para $678,86$ segundos, apresentando um ganho de $94,11\%$, com um desvio padrão de $34,24$ segundos. Já o algoritmo executado em TPU apresentou um ganho de $2,84$ vezes sobre a CPU. Diminuindo o tempo de execução de $11529,20$ segundos para $4061,65$ segundos, obtendo um ganho de $64,77\%$, com um desvio padrão de $65,50$ segundos.

A CNN obteve uma acurácia de $66,06\%$ de precisão. Apesar do valor não ser tão alto os testes foram bem precisos, a CNN acertou a maior parte das emoções. Nos testes a CNN apresentou uma confusão na identificação da emoção do medo e surpresa. O fato de ocorrer a confusão é que ambas as emoções possuem características faciais semelhantes, o que pode produzir a confusão na hora de realizar a identificação. As semelhanças entre as duas emoções são sobrancelhas levantadas e maxilar aberto.

5. Conclusões e trabalhos futuros

Este trabalho abordou o desenvolvimento de uma rede neural convolucional para o reconhecimento de emoções através de expressões faciais, realizando inúmeros testes para analisar



(a) tempo alcançado em cada arquitetura

(b) *Speedup* alcançado (vezes)

Figura 1. Comparativo dos tempos e *Speedup* alcançados.

e avaliar o desempenho da aplicação executada em arquiteturas CPU, GPU e TPU. Com a nossa implementação foi possível aumentar a acurácia da RNA alcançando uma precisão de até $66,06\%$. Com relação ao tempo computacional a versão desenvolvida em GPU, obteve-se excelentes resultados, reduzindo o tempo de execução em até $16,98$ vezes se comparado com a execução na CPU.

Como trabalhos futuros pretende-se aplicar a solução desenvolvida em outras bases de dados para validar seu comportamento. Também modificar o algoritmo da Rede Neural Artificial para que você possa usar o ambiente de execução da TPU Google Cloud, que possui TPUs que possuem um melhor arquivamento do que a TPU gratuita fornecida no Google Colab, e também analisar a influência de outros hiperparâmetros, como taxa de aprendizagem, Dropout e Função de Ativação.

Referências

Amin, D., Chase, P., and Sinha, K. (2017). Touchy feely: An emotion recognition challenge. *Palo alto: Stanford*.

Bartlett, M. S., Littlewort, G., Fasel, I., and Movellan, J. R. (2003). Real time face detection and facial expression recognition: development and applications to human computer interaction. In *2003 Conference on computer vision and pattern recognition workshop*, volume 5, pages 53–53. IEEE.

Carrier, P.-L., Courville, A., Goodfellow, I. J., Mirza, M., and Bengio, Y. (2013). Fer-2013 face database. *Universit de Montral*.

Ekman, P. (1973). Cross-cultural studies of facial expression. *Darwin and facial expression: A century of research in review*, 169222(1).

Kanade, T. and Cohn, J. (2005). Au-coded facial expression database.

Tang, H. and Huang, T. S. (2008). 3d facial expression recognition based on properties of line segments connecting facial feature points. In *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 1–6. IEEE.

Yin, L., Wei, X., Sun, Y., Wang, J., and Rosato, M. J. (2006). A 3d facial expression database for facial behavior research. In *7th international conference on automatic face and gesture recognition (FGR06)*, pages 211–216. IEEE.