EXTRAÇÃO DE POSES DO CORPO HUMANO UTILIZANDO REDES NEURAIS PROFUNDAS PARA TRADUÇÃO DA LÍNGUA BRASILEIRA DE SINAIS

VITOR S. DE SOUZA[1], LUCAS VENEZIAN POVOA[2]

[1] Graduando em Tecnologia em Análise e Desenvolvimento de Sistemas, Bolsista PIBIFSP, Instituto Federal de Educação Ciência e Tecnologia de São Paulo, Câmpus Caraguatatuba, pliavi.contato@gmail.com.
[2] Professor, Instituto Federal de Educação Ciência e Tecnologia de São Paulo, Câmpus Caraguatatuba, venezian@ifsp.edu.br.

Área de conhecimento (Tabela CNPq): Sistemas de Informação – 1.03.03.04-9

RESUMO: Línguas gestuais-visuais tornam possível a comunicação entre surdos e ouvintes. No entanto, apesar da existência das línguas de sinais e de no Brasil a Língua Brasileira de Sinais ser oficialmente a segunda língua do país, o número de pessoas que utilizam línguas orais-escritas, como o Português, e que detêm o conhecimento de uma língua gestual-visual ainda é muito pequeno. Um agente inteligente capaz de traduzir uma língua de sinais para uma língua oral-escrita pode ser um meio facilitador para a comunicação entre surdos e ouvintes. Desenvolver tal agente ainda é um grande desafio, mas o avanço da inteligência artificial, principalmente com o surgimento das redes neurais profundas, permite que tais desafios possam ser atacados de forma satisfatória. Um dos primeiros obstáculos para a tradução de uma língua gestual-visual é a extração das coordenadas dos juntas do corpo humano, termo simplificado como pose do corpo humano, a partir de uma imagem. Este projeto de Iniciação Científica tem como finalidade implementar uma rede neural profunda capaz de extrair poses do corpo humano utilizando tecnologias de código e padrão abertos. A rede terá como entrada imagens coloridas e de profundidade e a saída será as coordenadas x, y e z de quinze pontos de corpo humano. Esta IC pertence ao escopo do projeto de pesquisa DeepLIBRAS.

PALAVRAS-CHAVE: Línguas de sinais; Extração de poses; Rede neural.

1 INTRODUÇÃO

A extração de poses de um corpo humano é um tópico de pesquisa bem conhecido na visão computacional, pois é utilizado em diversas áreas, como segurança, jogos, análise de comportamento e interação humano-computador. Alguns algoritmos que se utilizam de imagens de profundidade para esta extração já podem ser vistos em prática (KNOOP, 2006; SHOTTON, 2011). Este tipo de estudo é de grande valia para traduções de línguas classificadas como gestual-visual, como a Língua Brasileira de Sinais (LIBRAS), pois tendem a se utilizar dos três pontos — X, Y e Z — do sistema de coordenadas cartesiano para cada parte do corpo humano a fim de possibilitar a leitura de cada movimento necessário. Porém, a utilização de imagens de profundidade para este tipo de tradução se torna inviável, pois é necessário um hardware específico para captura desse tipo de imagem, sendo caro e de difícil implantação ou utilização recorrente.

Uma alternativa é o emprego de redes neurais profundas, permitindo que modelos computacionais aprendam representações de dados em diversos níveis de abstração, sendo nos últimos anos o estado da arte no quesito reconhecimento de voz, reconhecimento de objetos e diversos outros domínios de classificação e regressão de padrões naturais, como descoberta de drogas e genomas (LECUN, 2015). Já existem exemplos da utilização de redes neurais para reconhecimento de membros do corpo humano e estimação de poses com acurácias consideráveis já se utilizando de imagens RGB (SHOTTON, 2011; LI, 2014).

Este trabalho foi iniciado com o foco no estudo da utilização de redes neurais profundas para extração de poses a partir de imagens de entrada a serem utilizadas, neste caso, na tradução da LIBRAS pela equipe do projeto de pesquisa DeepLIBRAS (DEEPLIBRAS, 2016).

2 MATERIAIS E MÉTODOS

Este trabalho é baseado na utilização de uma rede neural convolucional (*ConvNet*), cuja estrutura está descrita na Figura 1. Tal rede filtra e responde a grupos de dados de entrada a fim de encontrar padrões em imagens a serem utilizados em uma rede *multilayer perceptron* (*MLP*) para que de acordo com o dados da rede convolucional possam ser gerados valores relevantes (DESHPANDE, 2016). Ambas as redes trabalham com diversas camadas que guardam pesos para que esses resultados sejam alcançados.

Um algoritmo denominado *backpropagation* é utilizada para que os pesos sejam atualizados com a finalidade de minimizar o erro gerado pela rede neural (LECUN, 2015).

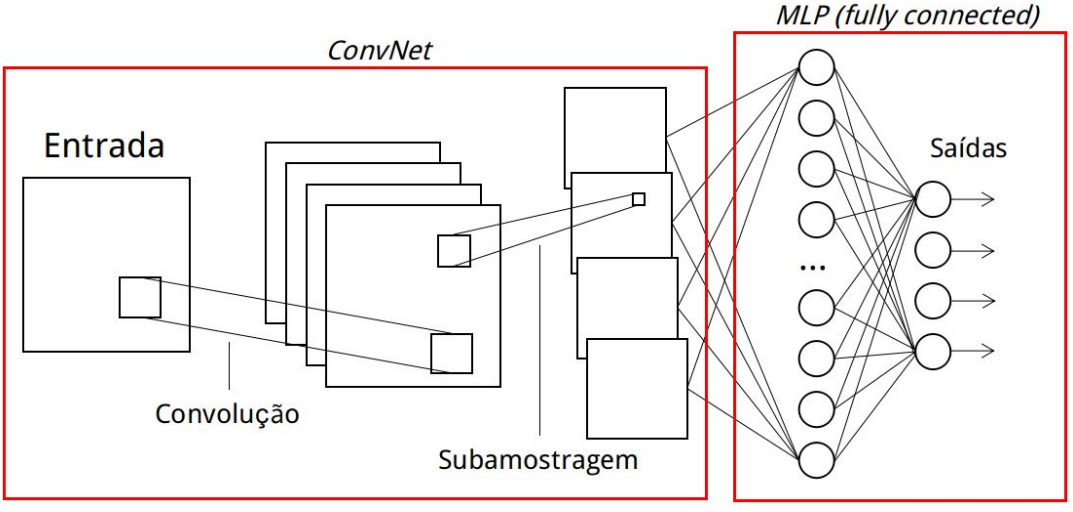

Figura 1: Estrutura de uma rede convolucional

O objetivo deste projeto é a conversão direta de uma imagem, onde um corpo humano pode ser encontrado, em uma matriz de 15 posições contendo em cada índice a posição X, Y e Z de cada membro do corpo pré-definidos para utilização na definição de uma pose, sendo eles: Cabeça, peito, ombros, cotovelos, mãos, tronco, quadris, joelhos e pés.

2.1 Redes Neurais Convolucionais

Redes neurais convolucionais são projetadas para processar dados em forma de matrizes. Existem quatro ideais chaves por trás de uma ConvNet que se aproveitam dos sinais naturais, sendo elas, conexões locais (cada área de acordo com o filtro gera uma nova *feature*), pesos compartilhados, *pooling* (subamostragem) e o uso de diversas camadas que quanto maior seu número, maior o detalhamento na busca de padrões (DESHPANDE, 2016).

3 RESULTADOS E DISCUSSÃO

Foram criadas duas variações baseadas no modelo do artigo *Heterogeneous Multi-task Learning for Human Pose Estimation with Deep Convolutional Neural Network* (LI, 2014), a primeira segue o mesmo modelo do artigo, porém, sem o detector de corpo. Todas as 5400 imagens utilizadas como entrada são do MPII Human Pose Dataset (ANDRILUKA, 2014) e contêm apenas um corpo como foco e apenas a posição dos membros superiores foram esperados como saída.

É possível perceber que após aproximadamente a iteração de número 6 mil, os valores que garantem o funcionamento correto da rede pouco se alteram, e isso em valores não satisfatórios, a taxa de acerto em imagens não conhecidas pela rede teve seu pico em 21%, e seu menor valor de erro foi 1560. As Figuras 2 e 3 apresentam a minimização das redes durante o processo de treinamento.

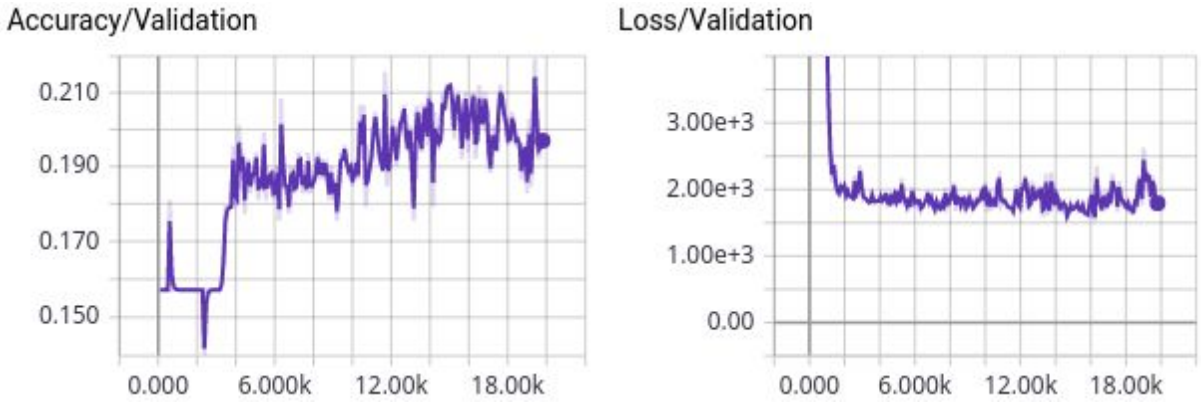

Figura 2: Resultados da primeira variação em 18 mil iterações. *Accuracy* (quanto mais próximo de 1, melhor), *loss* (quanto mais próximo de 0, melhor)

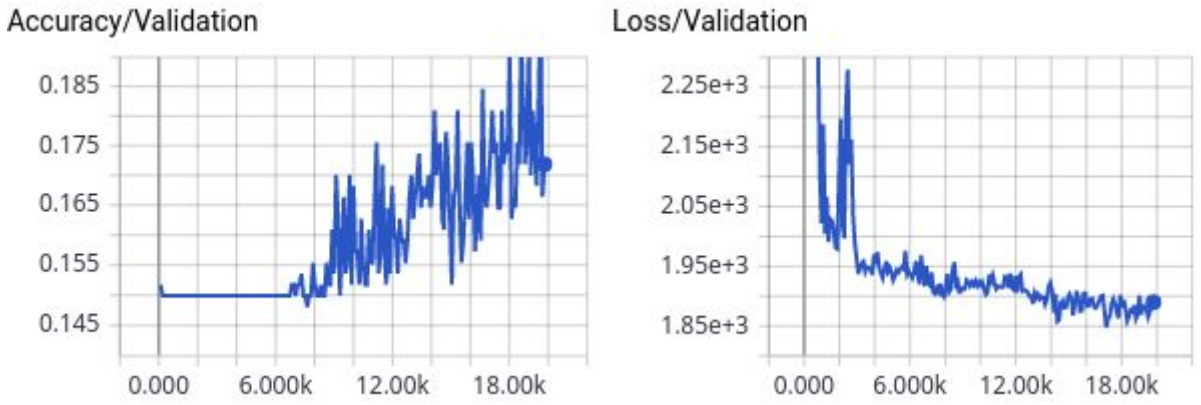

Figura 3: Resultados da segunda variação em 18 mil iterações. *Accuracy* (quanto mais próximo de 1, melhor), *loss* (quanto mais próximo de 0, melhor)

A segunda variação, utilizando *Leaky ReLU*[1] a fim de evitar neurônios que nunca se ativam por conta da *ReLU,* acabou por ter um resultado menos estável e menos satisfatórios que o da variação anterior.

A Figura 4 apresenta um dos resultados gerados pela primeira variação da rede.

Figura 4: à esquerda o resultado gerado pela rede, à direita o resultado esperado. Dataset: MPII Human Pose Dataset

4 CONSIDERAÇÕES FINAIS

Os resultados obtidos ainda não foram suficientes para resolver o problema de gerar as coordenadas dos pontos do corpo em imagens RGB, foi visto que após algumas iterações os valores passavam a se estagnar com erros muito altos. O motivo dessa estagnação e valores baixos serão estudados a fim de mudar esta situação, um estudo maior sobre as saídas e sobre a utilização de normalização e *data augmentation*[2] será feito.

Foi possível perceber que as coordenadas dadas pela rede se mantém sempre próximas ao centro da imagem, sendo necessário em estudo posterior descobrir a causa deste problema, não sendo possível utilizar este projeto no estado atual.

REFERÊNCIAS

KNOOP, S.;VACEK, S.;DILLMANN, R. **Sensor fusion for 3D human body tracking with an articulated 3D body model**. Orlando: IEEE, 2006. ISBN: 0-7803-9505-0.

SHOTTON, J. et al. **Real-Time Human Pose Recognition in Parts from Single Depth Images**. IEEE: Computer Vision and Pattern Recognition, 2011.

[1] Variação da ReLU que permite um limiar de valores negativos.
[2] Técnica que cria distúrbios e ruídos nos dados a fim de resultar em maior abstração dos dados pela rede

LI, S.; LIU, Z. Q.; CHAN, A. B. **Heterogeneous Multi-task Learning for Human Pose Estimation with Deep Convolutional Neural Network**. International Journal of Computer Vision, 2014, vol. 113, p. 19-36. ISSN: 0920-5691.

DEEPLIBRAS. **DeepLIBRAS: Tradução da Língua Brasileira de Sinais com Técnicas de Aprendizagem de Máquina**. 2016. Disponível em: <http://deeplibras.github.io>. Acesso em: 06 de julho de 2017.

LECUN, Y.; BENGIO, Y.; HINTON, G. **Deep learning**. New York: Macmillian, 2015. p. 436-444. DOI:10.1038/nature14539.

HAYKIN, S. **Neural Networks and Learning Machines. 3ª Edição**. Ontario: Prentice Hall, 2009. p. 1-5, 10-14. ISBN-10: 0-13-147139-2.

ANDRILUKA, M. et. al. **2D Human Pose Estimation: New Benchmark and State of the Art Analysis**. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

DESHPANDE, A. **A Beginner's Guide To Understanding Convolutional Neural Networks**. 2016. Disponível em: <https://adeshpande3.github.io/A-Beginner's-Guide-To-Understanding-Convolutional-Neural-Networks>. Acesso em: 06 de julho de 2017.