Um modelo de rede neural siamesa para re-identificação de pessoas em imagens, utilizando rede neural convolucional e *autoencoder*

Fábia Isabella Pires Enembreck [1], **Erikson Freitas de Morais** [1]

[1]Departamento de Informática – Universidade Técnológica Federal do Paraná (UTFPR)
Ponta Grossa – PR – Brasil

enembreck@alunos.utfpr.edu.br, emorais@utfpr.edu.br

Abstract. *The problem of determining whether a person being watched by a camera has ever been present in an environment is called person re-identification. This problem is considered challenging since the images obtained by cameras are subject to suffer variations, such as illumination, and partial occlusions. In this work, a method was developed for the problem, the method consists of using a siamese neural network architecture, composed of two identical subnets. Each subnet is formed by a convolutional neural network and a denoising autoencoder, responsible for rebuilding the vectors produced by the convolutional neural network, maintaining the most important features for the re-identification.*

Resumo. *O problema de determinar se uma pessoa que esta sendo observada por uma câmera já esteve presente em um ambiente é chamado de re-identificação da pessoas. Esse problema é considerado desafiador, pois as imagens estão sujeitas a sofrer variações, como iluminação, além de oclusões parciais. Este trabalho propõe um método para o problema, que consiste em utilizar uma rede neural siamesa, composta por duas sub-redes idênticas. Cada sub-rede é formada por uma rede neural convolucional e um denoizing autoencoder, responsável pela reconstrução dos vetores produzidos pela rede neural convolucional, mantendo as características mais importantes para a re-identificação.*

Keywords: Person re-identification; Deep Learning; Artificial neural networks

Palavras-chave: Re-identificação de pessoas; Aprendizagem profunda; Redes neurais artificiais

1. Introdução

A re-identificação de pessoas consiste em identificar se uma pessoa já esteve em um determinado ambiente antes, de forma que seja atribuído o mesmo rotulo a ela em todas as suas aparições [Ahmed et al. 2015], definida como o processo em que se pretende estabelecer correspondência entre diferentes imagens de uma mesma pessoa [Bedagkar-Gala and Shah 2014]. O problema de re-identificar pessoas em imagens é considerado muito complexo, uma vez que as imagens analisadas estão sujeitas a inúmeras variações de iluminação e pontos de vista, além de oclusões parciais e baixas resoluções,

dentre outros problemas. Dessa forma, duas imagens de uma mesma pessoa tiradas em uma mesma cena podem estar muito diferentes entre si, fazendo com que um re-identificador reconheça as imagens como sendo de duas pessoas distintas.

Vários métodos foram desenvolvidos para tratar esse problema, entre eles destacam-se métodos que utilizam a combinação de histogramas de duas imagens de pessoas, como em [Prosser et al. 2010] e [Gray and Tao 2008]. Além disso, aprendizagem de máquina é uma área que vem sendo utilizada para a re-identificação de pessoas, como em [Yi et al. 2014] e [Ahmed et al. 2015] que implementam variações de uma rede neural siamesa. No entanto, o problema ainda segue sem uma solução definitiva, devido as variações apresentadas pelas imagens.

Este trabalho propõe um método utilizando uma Rede Neural Siamesa para re-identificar pessoas em imagens. A rede e formada por duas sub-redes idênticas compostas por uma Rede Neural Convolucional (CNN) e um *Autoencoder* (AE). Cada sub-rede produz um vetor de características de uma imagem e no final a Rede Neural Siamesa estima a similaridade entre os vetores para verificar se pertencem a uma mesma pessoa ou não. Para isso, a CNN produz um vetor de características gerais da imagem, que passam pelo AE que produz a saída de cada sub-rede, mantendo as características mais relevantes para a re-identificação, de forma que sejam amenizadas possíveis variações entre as imagens.

O objetivo deste trabalho e o desenvolvimento de um método para re-identificar pessoas em imagens utilizando técnicas de aprendizagem profunda. Para isso, foram determinados os seguintes objetivos específicos: 1) Identificar possíveis técnicas de aprendizagem profundas que possam ser aplicáveis ao problema; 2) Propor e implementar um modelo para re-identificação de pessoas; 3) Selecionar *datasets* públicos para testar o modelo proposto; 4) Realizar experimentos para validação do modelo.

2. Metodologia

Esta seção apresenta a metodologia proposta neste trabalho para a re-identificação de pessoas em imagens, obtidas por diferentes câmeras e pontos de vista. A Figura 1 exibe a Rede Neural Siamesa utilizada, que consiste em duas sub-redes, sendo que cada uma delas recebe uma imagem de entrada.

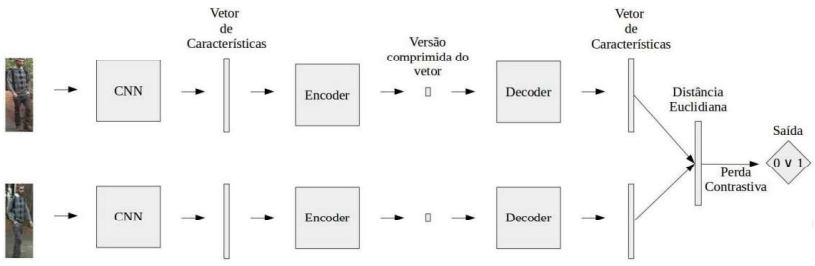

Figura 1. Modelo de Rede Neural Siamesa proposto

2.1. Rede Neural Siamesa

Uma Rede Neural Siamesa consiste em duas sub-redes idênticas que compartilham os mesmos parâmetros e são unidas em suas saídas. Cada sub-rede recebe uma entrada

diferente que é mapeada para um descritor de características. Com isso, a rede obtém dois descritores diferentes que são comparados para estimar a similaridade entre eles, resultando na saída da rede [Bromley et al. 1994].

Na Figura 1 é apresentado o modelo de Rede Neural Siamesa proposto. A entrada da rede e composta por um par de imagens e um valor binário Y que identifica se o par de imagens e de uma mesma pessoa ou não. Com isso, se $Y = 0$ então as imagens são da mesma pessoa e se $Y = 1$ as imagens são de pessoas diferentes.

Durante o treinamento, a rede modifica seus parâmetros para que possa diminuir a distância entre um par de imagens de uma mesma pessoa e aumentar entre um par de imagens de pessoas diferentes, de acordo com o rótulo Y recebido. Considerando que X_1 e X_2 representam o par de imagens de entrada e G_W é um conjunto de funções de rede, a função $E_W = |G_W(X_1) - G_W(X_2)|$ é utilizada para calcular a distância e passa por uma Função de *Constrative Loss* L.

A Função de *Contrastive Loss*, dada pela Equação 1, é utilizada para estimar a capacidade da rede de encontrar semelhanças entre as imagens. Para isso, a função aprende parâmetros, de forma que exemplos mais semelhantes se aproximem e os diferentes sejam separados.

$$L(W, Y, X_1, X_2) = (1 - Y)\frac{1}{2}(E_W)^2 + Y\frac{1}{2}\{max(0, m - E_W)\}^2 \quad (1)$$

onde m é a margem, se $m > 0$ e (X_1, X_2) é um par de imagens positivo e (X_1', X_2') é um par de imagens negativo, então $E_W(X_1, X_2) + m < E_W(X_1', X_2'))$ [Bromley et al. 1994].

2.2. Rede Neural Convolucional - CNN

A primeira parte de cada sub-rede da rede proposta consiste em uma CNN, responsável por produzir um vetor de características da imagem de entrada. Esta CNN á formada por 4 camadas de convolução, alternadas entre uma camada de *max-pooling* e outra de normalização.

Uma CNN é caracterizada pela utilização de operações de convolução, em pelo menos uma de suas camadas, para aprender padrões de um determinado conjunto de dados. No caso de um conjunto de dados formado por imagens, a camada de convolução é formada por um conjunto de filtros, produzindo na sua saída mapas de características de uma imagem [LeCun et al. 2015].

Uma camada de agrupamento *pooling* é utilizada com objetivo de reduzir o tamanho dos mapas de características produzidos, de forma que se diminua também o número de parâmetros e custo computacional. O tipo de agrupamento utilizado foi o *max-pooling*, que, para uma região 2x2 dos mapas de características, substitui no *pixel* correspondente do mapa de característica da saída o maior valor dessa região [LeCun et al. 2015].

2.3. *Denoising Autoencoder* - DAE

Uma rede neural do tipo *autoencoder* (AE) é formada por pelo menos 3 camadas: entrada, saída e camada intermediária, sendo que a saída deve ter o mesmo tamanho da entrada,

uma vez que o AE procura reproduzir a entrada da rede na sua saída. Isso é possível através de uma função codificadora, que extrai características da entrada e produz um vetor de características comprimido. A descompressão deste vetor é realizada através de uma função decodificadora que traduz as informações o mais próximo possível da entrada, priorizando as propriedades mais importantes dos dados [Goodfellow et al. 2016].

Um tipo de AE é o *Autoencoder* de Denonização ou *Denoising Autoencoder* (DAE). Este modelo recebe dados corrompidos como entrada e, durante o treinamento, procura prever os dados de entrada não corrompidos em sua saída [Goodfellow et al. 2016]. Com isso, um DAE foi utilizado na rede para, a partir do vetor gerado pela CNN, produzir uma representação de forma que sejam reconstituídas as informações mais relevantes para a re-identificação durante o treinamento da rede.

3. Experimentos e Resultados

Esta seção apresenta os resultados obtidos com o treinamento e teste da rede neural siamesa proposta. A rede foi implementada utilizado *Python*, através das bibliotecas para redes neurais *Keras* e *Tensorflow*.

3.1. VIPeR

O *dataset ViewPoint Invariant Pedestrian Recognition* - (ViPER) contém um total de 1264 imagens de 632 pedestres, assim, para cada pedestre, existem duas imagens capturadas por diferentes câmeras, contendo variação entre os pontos de vista e também mudanças na iluminação e nas poses dos indivíduos [Gray and Tao 2008].

Com objetivo de aumentar o numero de imagens de treinamento e reduzir o *overfitting*, foi aplicada a técnica de *data augmentation*. Assim, para cada imagem, foi aplicada 11 transformações, resultando em um total de 15144 imagens. Os resultados obtidos após o treinamento e teste podem ser observados na Tabela 1 e na Figura 2. Neste teste, a maior acurácia obtida foi com 1200 épocas de treinamento, atingindo uma porcentagem de 87.93%. O treinamento com 600 épocas obteve o pior resultado, com uma acurácia de 65.2%.

Tabela 1. Acurácia em relação ao N° de épocas de treinamento, utilizando o *dataset* VIPeR

Épocas	Acurácia (%)	Épocas	Acurácia (%)
100	84,47	800	80,23
200	82,92	1000	86,59
400	82,92	1200	87,93
600	65,2		

3.2. iLIDS-VID

O *dataset iLIDS Video Re-Identification* (iLIDS-VID) contém imagens de 300 pessoas diferentes obtidas através de 2 câmeras não sobrepostas. Esse *dataset* disponibiliza um maior número de imagens do que o VIPeR, uma vez que as imagens foram obtidas

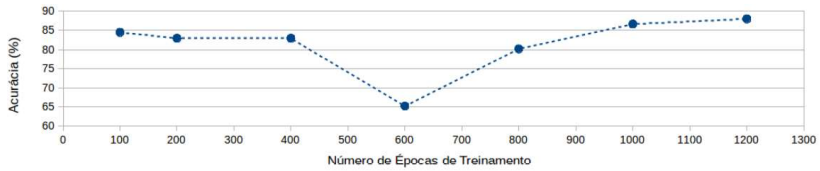

Figura 2. Acurácia em relação ao N° de épocas de treinamento, utilizando o *dataset* VIPeR

por meio de rastreamento de pedestres, contendo de 23 a 192 *frames* para cada pessoa. As imagens tem diversas variações, como iluminação, pontos de vista e oclusões [Wang et al. 2014].

No primeiro experimento com o iLIDS-VID, foram utilizadas 21969 imagens de 319 pessoas diferentes obtidas por duas câmeras, sendo que foram destinadas 14751 imagens para treinamento e 7218 imagens para testes. Os resultados podem ser vistos na Tabela 2 e na Figura 3, na linha representada por "CNN+AE". A melhor acurácia foi 94.26% para 600 épocas de treinamento e o pior foi 89.37% para 1400 épocas de treinamento.

Tabela 2. Acurácia em relação ao N° de épocas de treinamento, utilizando o *dataset* iLIDS-VID

Épocas	Acurácia (%)	Épocas	Acurácia (%)	Épocas	Acurácia (%)
100	93,04	1000	92,4	2000	93,22
200	93,49	1200	89,82	2200	92,06
400	91,2	1400	89,37	2400	92,45
600	94,26	1600	94,09	2600	93,46
800	91,39	1800	90,93	2800	93,82

Para comparação, também foi realizado um experimento com sub-redes compostas apenas pela CNN. A Tabela 3 e a Figura 3 na linha representada por "CNN"apresentam os resultados obtidos. A melhor acurácia foi alcançada com 2000 épocas de treinamento, 87, 78% e a pior foi igual a 87, 19% para 2600 épocas de treinamento.

Tabela 3. Acurácia em relação ao N° de épocas de treinamento, para sub-redes formadas apenas pela CNN, utilizando o *dataset* iLIDS-VID

Épocas	Acurácia (%)	Épocas	Acurácia (%)	Épocas	Acurácia (%)
100	87,42	1000	87,30	2000	87,78
200	87,48	1200	87,43	2200	87,54
400	87,42	1400	87,25	2400	87,27
600	87,73	1600	87,33	2600	87,19
800	87,56	1800	87,51	2800	87,54

4. Conclusão e Trabalhos Futuros

Um modelo de aprendizagem profunda foi proposto para o problema de re-identificação de pessoas em imagens, que consiste em uma rede neural siamesa, formada por duas sub-

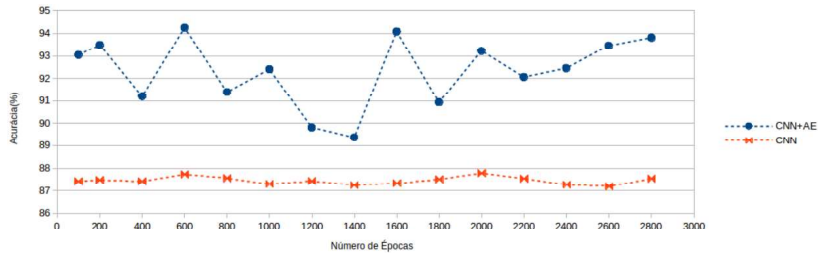

Figura 3. Comparação das acurácias em relação ao N° de épocas de treinamento, usando o *dataset* iLIDS-VID para as duas redes implementadas.

redes idênticas, para estimar a similaridade em duas imagens. Cada sub-rede é composta por uma CNN e um DAE. Através de experimentos realizados, verificou-se que a rede siamesa proposta possui grande potencial na re-identificação de pessoas, visto que obteve resultados melhores do que a rede com apenas uma CNN, muito utilizada em outros trabalhos da área, em todos os casos. Como trabalhos futuros, pretende-se realizar mais testes usando outros *datasets* e verificar o desempenho da rede para diferentes qualidades de imagens, condições de iluminação e oclusões parciais.

Referências

Ahmed, E., Jones, M., and Marks, T. K. (2015). An improved deep learning architecture for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3908–3916.

Bedagkar-Gala, A. and Shah, S. K. (2014). A survey of approaches and trends in person re-identification. *Image and Vision Computing*, 32(4):270–286.

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., and Shah, R. (1994). Signature verification using a"siamese"time delay neural network. In *Advances in neural information processing systems*, pages 737–744.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

Gray, D. and Tao, H. (2008). Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European conference on computer vision*, pages 262–275. Springer.

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

Prosser, B. J., Zheng, W.-S., Gong, S., Xiang, T., and Mary, Q. (2010). Person re-identification by support vector ranking. In *BMVC*, volume 2, page 6.

Wang, T., Gong, S., Zhu, X., and Wang, S. (2014). Person re-identification by video ranking. In *European conference on computer vision*, pages 688–703. Springer.

Yi, D., Lei, Z., Liao, S., and Li, S. Z. (2014). Deep metric learning for person re-identification. In *2014 22nd International Conference on Pattern Recognition*, pages 34–39. IEEE.