Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)

doi:10.3233/FAIA230665

OPNet: A One-Shot Image Similarity Algorithm for Production Systems

Karl Fabian Svensson[a,1] and Carlos Guerrero-Mosquera[a,2]
[a] *La Salle Ramon Llull University*
ORCiD ID: Carlos Guerrero-Mosquera https://orcid.org/0000-0001-8265-3651

Abstract. The appearance of new trends in the field of cognitive neuroscience, for example object persistence, has paved the way for the evolution of deep CNNs into Siamese Neural Network architectures such as OPNet. These networks allow for image recognition without the need for expensive labelled data. In this work, we apply this technology to a small Spanish tech e-commerce struggling with the production of their customizable products. Our goal was to automatically identify each product's order in the company's internal system by matching photos of the products taken by workers with system-generated images. After testing various architectures, we achieved 91% accuracy with a triplet loss model using deep CNN embedding networks. The algorithm was trained on a dataset of 9696 unique product images captured in the company's production department. The paper details the technical aspects of the Siamese Neural Network architecture, including the triplet loss and SoftMax distance function used to train it. Our results demonstrate the potential of these deep learning models to generate practical benefits for firms, since it reduces human errors, while improving the effectiveness and efficiency of the company's internal processes.

Keywords. SNN, Image-Similarity, Triplet-Loss, Artificial Vision, Image Processing, Artificial Neural Networks, AI Problem Solving

1. Introduction

Despite the success of the neural network when it comes to image classification and object detection, image similarity has traditionally proven to be a more daunting task.

Advanced statistical comparison techniques such as comparing ontology trees or latent semantic indexing perform poorly when the data is not organized or structured in a specific and predictable manner.

Extracting feature vectors is an excellent way of analyzing images, since the information therein lies in the interaction between the pixels. Arguably the most popular machine learning algorithm for extracting feature vectors is called Convolutional Neural Network (CNN). The ConvNets which constitute these are excellent at capturing the spatial dependencies in an image by applying certain filters[3].

[1] Karl Fabian Svensson, karlfabian.svensson@students.salle.url.edu
[2] Carlos Guerrero-Mosquera carlos.guerrero@salle.url.edu

2. Algorithm

SNNs are state-of-the-art in the field of image similarity. They contain two identical subnetworks which perform an identical transformation on some input data. The two subnetworks commonly have a CNN architecture which ends in a fully connected network that constitutes some feature vector or output.

Two major sub-architectural branches can be distinguished within SNNs depending on the loss function employed.

The traditional SNN (Contrastive Loss Function) relies on an activation function, commonly SoftMax, to evaluate whether two images are similar. Du and Fang at Stanford output a SoftMax cross-entropy loss which returns 1 if the inputs are similar and 0 if they are not[2]. DeepFace use a similar approach, exploring different similarity metrics such as the Kai squared formula[4].

The second branch, called Triplet Loss Function, was first introduced by Hoffer et al and trains on three images for each iteration. Its objective is to minimize the distance between a query image and a positive image, and to maximize the distance between the query image and the negative image (see *formula 1*).

$$\| f(x_i^a) - f(x_i^p) \|_2^2 + \propto < \| f(x_i^a) - f(x_i^n) \|_2^2, \sum_i^N [\| f(x_i^a) - f(x_i^p) \|_2^2 - \| f(x_i^a) - f(x_i^n) \|_2^2 + \propto]$$

Formula 1. Triplet Loss Distance Function

2.1. One-Shot Learning

The purpose of few-shot learning is to bring artificial intelligence models closer to human conceptual knowledge.

While humans can learn from just a handful of examples, the highest performing deep learning models are very "data-hungry". One-Shot learning consists of training models to correctly classify elements having seen only one example of its class.

3. Methods

OPNet was proposed to improve the efficiency of the production department of a Spanish company that offers personalized products. Because of the personalization, the dataset used to train and test OPNet consists entirely of unique classes, i.e., the model is both one-shot training and one-shot testing. The company's main product is the die-cut sticker.

In order to attain the best result possible, different types of embedding models were experimented with. In total, 7 variations were created, 3 using a SoftMax activation function and 4 with a Triplet Loss function.

4. Data Setting

OPNet was trained using 9696 black and white images of dimension 320x180. An additional 1250 images were stored independently of the model to be used uniquely for specific validation purposes.

All these images were taken by the production department at OriginalPeople throughout a 3-month period ranging from May 2nd, 2022, to August 5th, 2022.

In order to ensure data consistency, a small wooden structure was built with a camera attached to the top. A piece of black vinyl was then placed on the area directly below the camera to provide a neutral background for all images.

One difficulty was choosing the height at which to place the camera. The customizable stickers vary in size from 15x25mm to 575x140mm.

Ultimately, to achieve the best trade-off between resolution and not cutting off too much of large stickers, a height of 48cm was chosen which covers an area 400 mm wide. The images produced are 1280x960 pixels and weigh around 70 – 80 KB.

4.1. Preprocessing

In order to avoid the overhead that comes with preprocessing images at execution time, all images are preprocessed when taken and stored both in their original format and their manipulated format in the OPNet database. Refer to *figure 1*.

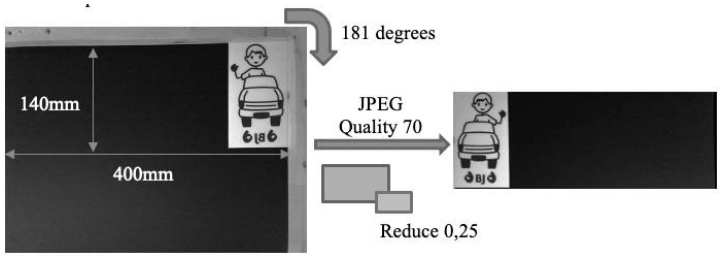

Figure 1. Preprocessing of images.

The resulting images have a dimension of 320x118 pixels and weigh around 1-2 KB. Both the original and the preprocessed images were stored to allow for future experimentation with pre-processing.

4.2. Sampling

Similar images are stored adjacent in memory to make it less costly to sample triplets. OPNet is trained using curriculum learning, which means that a sub-batch of semi-hard triplets are purposefully selected. Triplets are considered semi-hard if the distance between the negative and the query image is very similar to the distance between the latter and the positive image.

Based on tests performed, and in accordance with research by Appalaraju et al.[1], this speeds up training by about 20-30% and results in accuracy levels of a few percentage points higher for the same number of epochs, compared to if only randomly sampled triplets are used.

5. Results and Analysis

All models were evaluated using accuracy, by taking large batches of pairs and triplets and returning a percentage. Other visual functions were implemented to get a better understanding of why an error occurs.

Variations in complexity, depth and feature vector dimension were experimented with. See the summary of the results in *table 1* below:

Table 1. Results of all 7 models.

Model	Loss	Accuracy
SNN Iceberg	0,226	83,95%
SNN Gregory	0,35	88,15%
SNN Flowers	0,76	87,7%
Triplet Dense 1	-	77,6%
Triplet Dense 2	-	70,5%
Triplet CNN 1	-	90,7%
Triplet CNN 2	-	88,4%

5.1. Best Performing model: Triplet CNN 1

This model uses a CNN embedding network, same as the SNN models. The accuracy achieved was 90,7% and the AUC was 90% (see *figure 2)*.

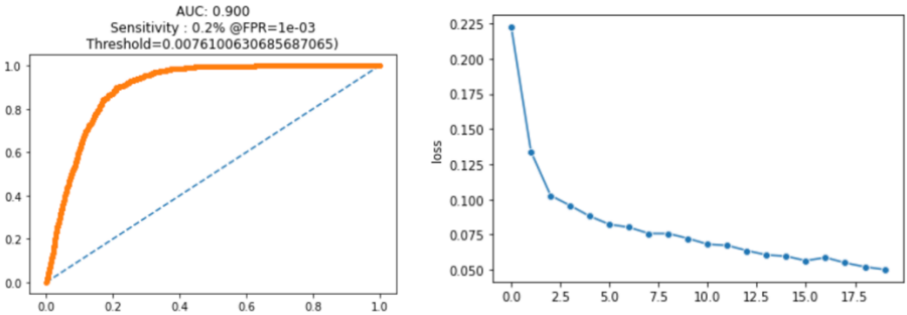

Figure 2: Triplet Model 3 – ROC and training of Triplet CNN 1

This is the best performing model so far, and the training does not converge after 20 epochs, meaning that better results should be attainable through further training.

6. References

[1] Appalaraju, S., Chaoji, V. (2017) *Image similarity using Deep CNN and Curriculum Learning*. ArXiv, abs/1709.08761.

[2] Du W, et al (2017) Siamese convolutional neural networks for authorship verification. Tech. rep. Stanford University

[3] Saha, S., 2018. *A Comprehensive Guide to Convolutional Neural Networks—the ELI5 way*. [online] Medium. Available at: https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53 [Accessed 01 May 2022].

[4] Taigman Y, et al (2014) *DeepFace: closing the gap to human-level performance in face verification*. In: Proceedings of CVPR 2014 – the IEEE conference on computer vision and pattern recognition, pp 1701–1708