Image Similarity

Munteanu Victor

December 2022

Abstract

In this paper, we compare several Siamese network architectures for image similarity, as well as several pretraining techniques and explain their performances using a variety of saliency methods. Our resulting approach combines the strengths of both adaptation layers and contrastive learning to improve the model's performance and suggests a workflow for obtaining insights into its internal workings.

1 Introduction

The remainder of this paper is organized as follows. In Section 2, we provide a brief overview of the image similarity problem and the relevant background knowledge. In Section 3, we present a comprehensive review of the state-of-the-art approaches to image similarity using deep learning. In Section 4, we evaluate and compare the performance of these approaches on various image similarity benchmarks. In Section 5, we analyze and provide further insights into the model internal workings by using Saliency maps. Finally, in Section 6, we discuss the strengths and limitations of the different approaches and provide some concluding remarks.

2 Background

Image similarity is a crucial problem in various fields such as image retrieval, object recognition, and content-based image classification. The ability to accurately

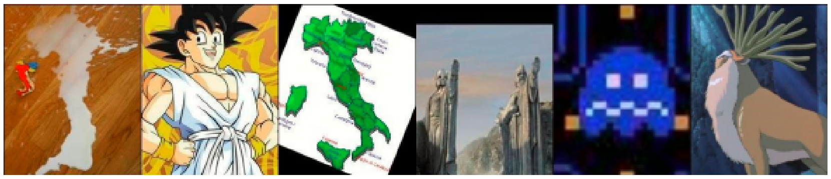

Figure 1: The task: For the given left-most image, find the most visually similar image between the rest of the images.

measure the similarity between two images enables the development of many important applications, such as efficient image search engines and improved image compression algorithms.

In recent years, deep learning has become a popular choice for tackling the image similarity problem due to its ability to learn high-level features from data. Deep neural networks have achieved state-of-the-art performance on various image similarity benchmarks and have become the standard approach for this task.

However, the choice of the specific deep learning approach to use for measuring image similarity is not straightforward. Different approaches have been proposed in the literature, each with its own set of advantages and disadvantages. In this paper, we provide a comprehensive review of the state-of-the-art approaches to image similarity using deep learning. We evaluate and compare the performance of these approaches on various image similarity benchmarks and discuss their strengths and limitations.

3 Related Work

The problem of determining the similarity between two images has been widely studied in the computer vision community. One approach is to use deep learning models, such as convolutional neural networks (CNNs), to extract features from the images and compare them using a distance metric.

One popular method for image similarity is to use a Siamese network, a type of CNN architecture that consists of two identical subnetworks that share weights and are trained to process a pair of input images. The output of the Siamese network is then used to measure the similarity between the two images.

Several approaches have been proposed to improve the performance of Siamese networks for image similarity. One method is to add extra layers, known as adaptation layers, to a pretrained CNN on a large dataset, such as ImageNet, and fine-tune the model on a specific image similarity dataset, such as the Totally Looks Like dataset. This approach has been shown to effectively capture domain-specific features and improve the model's performance.

Another approach is to use contrastive learning, a type of self-supervised learning that maximizes the distance between negative pairs of images and minimizes the distance between positive pairs. This can be achieved using the triplet loss, a type of loss function that compares the distance between a positive pair and a negative pair in a triplet of images.

Other methods for image similarity include using metric learning algorithms, such as the K-nearest neighbors (KNN) algorithm and the maximum margin criterion (MMC), and feature matching techniques, such as the Scale Invariant Feature Transform (SIFT) and the Speeded Up Robust Features (SURF).

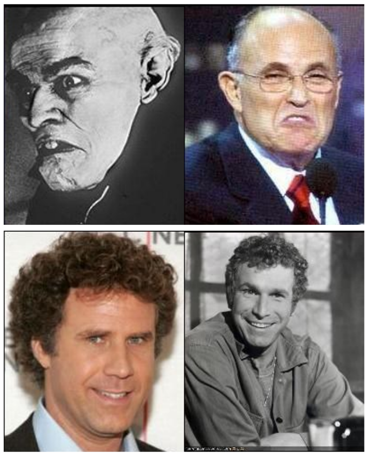

Figure 2: Pairs of images matched either by facial expressions or facial features.

3.1 Dataset

To evaluate our model, we use the Totally Looks Like dataset, which contains over 6000 pairs of images collected from a popular entertainment website with the same name. The Totally Looks Like dataset is a good choice for our purposes because it contains a diverse set of images that are visually similar, but not necessarily semantically related. However, it does have some drawbacks.

One drawback of the Totally Looks Like dataset is that it is asymmetric, meaning that it is not known whether the left image was found to be similar to the right, or vice versa. This can make it difficult to accurately evaluate the model's performance, as the direction of the comparison is not known.

Another drawback is that 75 of the dataset needs to be removed since these images were paired only on face expressions and/or face features, while for the image similarity task we need very general visual features. This leaves us with a smaller dataset to work with, which can limit the model's ability to learn.

Despite these drawbacks, the Totally Looks Like dataset is still a useful resource for evaluating image similarity models, and we believe that our proposed Siamese network architecture and saliency analysis methods will provide valuable insights into the model's performance.

As you can see in figure 2, the Totally-Looks-Like dataset contains images paired only by facial features.

4 Experiments

In our experiments, we first pretrained a feature extractor on the ImageNet dataset using the Barlow Twins Loss. This loss function was specifically designed to learn representations of image similarity, and it has been shown to lead to improved performance in image similarity tasks compared to other loss functions.

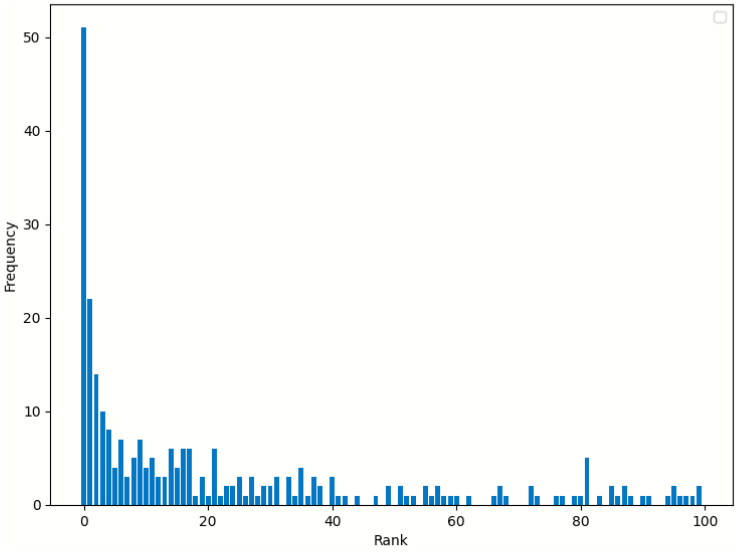

Figure 3: Frequency vector of ranks for the Siamese Network pretrained with Barlow Twins. Each column represents for how many images, their correct match was found at that exact rank.

After pretraining the feature extractor, we integrated it into a Siamese network architecture, adding additional projection layers on top. The Siamese network was then fine-tuned using a contrastive loss function, such as the Triplet Loss. This loss function minimizes the difference between the feature representations of similar images and maximizes the difference between the feature representations of dissimilar images. This fine-tuning step was designed to further improve the performance of the Siamese network for the image similarity task.

For evaluation, we used the top-k metric. It is worth noting that the top-k metric is a commonly used evaluation metric in image similarity tasks. It provides a more comprehensive view of the performance of the model by considering not only the most similar image, but also the top-k most similar images. This metric can be particularly useful in real-world applications, where it may be desirable to retrieve a set of similar images, rather than just a single most similar image. This metric works as follows: given a reference image, we look through all 1200 non-face-like images and choose only the 25 most similar images. If the target image is among these, it is added to the accuracy as correct.

5 Results

As we can see in the plot 3, the Siamese network pretrained using the Barlow Twins Loss achieved a top-1 score of 51/325, compared to the same Siamese network pretrained as classifier which had a score of 41/325, shown in the plot

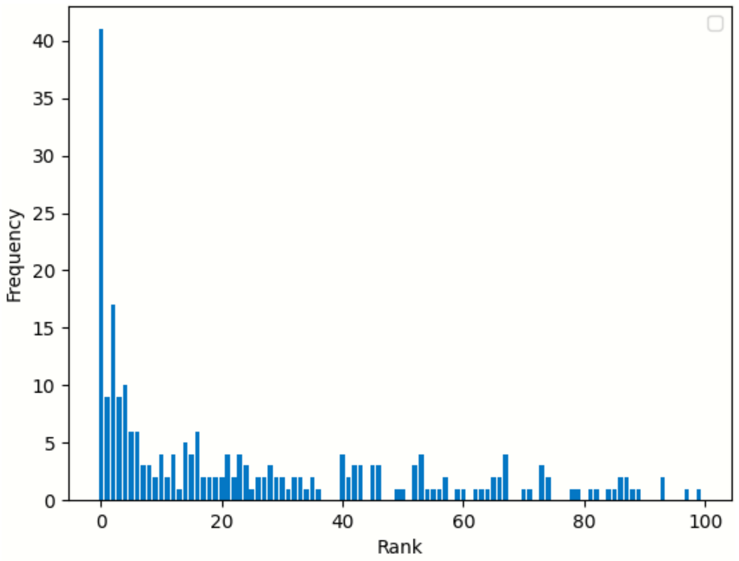

Figure 4: Frequency vector of ranks for the Siamese Network pretrained for classification.

at 4. This result suggests that pretraining using the Barlow Twins Loss leads to a more effective representation of image similarity.

The same pattern can be seen when we also compute the sum of the first 5 frequencies in the plots, we obtain the top-5 metric. The top-5 metric for the network pretrained using the Barlow Twins Loss was 105/325, compared to 86/325 for the network pretrained as classifier.

The results of our experiments indicate that the Siamese network that was pretrained on ImageNet using the Barlow Twins Loss performed better than the one that was pretrained for image classification.

In our experiments, we have also tested how well the feature vector obtained from a pretrained ResNet on ImageNet is for measuring the similarity between images. We found that adding two linear layers as adaptive layers and fine-tuning them on the Totally Looks Like dataset did not give valuable results. However, our preliminary evaluations show that only 20 out of 1200 images had the match found in the 25 most similar images. This indicates that the feature vector obtained from the pretrained ResNet is not sufficient for accurately measuring image similarity on the Totally Looks Like dataset.

5.1 Analisys

To gather insights of the trained model, we used Saliency maps methods by backpropagating the gradients of the MSE Loss through the output embeddings of the 'similar' images up to the 2 input images and then plotted their gradients.

As you can see in figure 5, while deciding that the pair of images is 'similar',

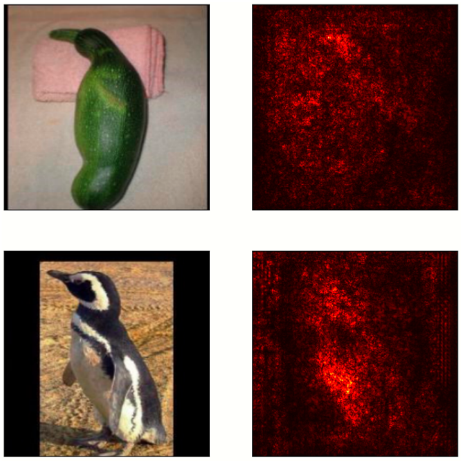

Figure 5: Saliency map of a pair of images that is a match.

the model seems to have considered the shape of the head the most, along with the belly of the penguin in the bottom image.

6 Conclusion

Our results suggest that pretraining Siamese networks using contrastive learning, such as the Barlow Twins Loss, may be a promising approach for image similarity tasks.

References

[1] O. Risser-Maroix (2022) *Learning an adaptation function to assess image visual similarities*, LIPADE, Université de Paris – Paris, France

[2] Karen Simonyan & Andrea Vedaldi & Andrew Zisserman (2014) *Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps.*, Visual Geometry Group, University of Oxford

[3] Richard Zhang & Phillip Isola & Alexei A. Efros (2018) *The Unreasonable Effectiveness of Deep Features as a Perceptual Metric.*, UC Berkeley @OpenAI and Adobe Research

[4] Jure Zbontar & Li Jing et al. (2021) *Barlow Twins: Self-Supervised Learning via Redundancy Reduction.*

[5] Amir Rosenfeld & Markus D. Solbach & John K. Tsotsos (2018) *Totally Looks Like - How Humans Compare, Compared to Machines.*, York University Toronto, ON, Canada

[6] Ting Chen & Simon Kornblith & Mohammad Norouzi & Geoffrey Hinton (2020) *A Simple Framework for Contrastive Learning of Visual Representations .*