One-Shot Learning With Pretrained Models in Siamese Network — ViT vs. CNN

Chun Hua, Lin
cl4335@columbia.edu

1. Introduction

The origin of the term "Siamese twins" can be traced back to the 19th-century conjoined twins, Chang and Eng Bunker, who were born in Siam (now Thailand) in 1811. They were joined at the chest and shared a liver, but each had their own heart and other vital organs. In 1829, they were discovered by a British merchant and entered the circus, performing all over the world. They lived to the age of 63 and had a total of 22 children between them. Their lives were a sensation, and they brought attention to the condition of conjoined twins, which was not well understood at the time. Their legacy lives on, and the term "Siamese twins" has become a well-known synonym for conjoined twins in popular culture.

A Siamese Network is a neural network architecture that consists of two or more identical subnetworks that share the same weights and architecture. These subnetworks are used to generate feature vectors for each input and compare them, typically by computing a distance or similarity metric between the vectors. In other words, two inputs are processed by a network with completely shared parameters, and the absolute difference is used as the input for the linear classifier - this is the necessary structure for a Siamese network. The network is like two identical twins who share the same head, hence the name "Siamese". The following image is a perfect representation of this structure. Siamese Networks can be applied to a variety of use cases. For example, they are commonly used in tasks such as detecting duplicates, finding anomalies, and most commonly used nowadays, face recognition.

In face recognition, Siamese Networks can be trained to compare the facial features of different images and determine whether they belong to the same person. Likewise, to fathom the mechanism of the neural network is asking the model whether the provided images are "similar". Note that similarity can be done with coincidences or unexplainable resemblance of two different images (faces). Overall, Siamese Networks offer a flexible and powerful approach for generating feature vectors and comparing inputs in a wide range of applications.

In addition, the backbone of a Siamese Network is the subnetwork architecture that is used to generate feature vectors for each input. The choice of backbone architecture will depend on the type of input data and the specific task at hand. For example, in image-based applications such as face recognition, a CNN-based backbone may be used to extract features from the input images. On the other hand, in text-based applications, an RNN-based backbone may be used to process sequential input data, such as sentences or paragraphs.

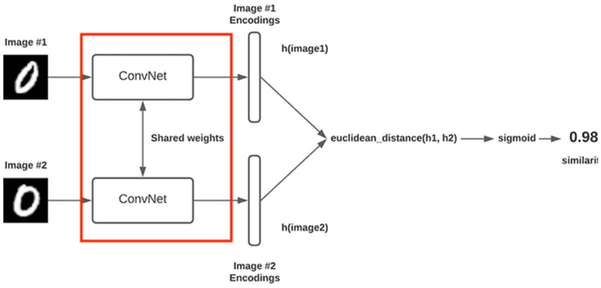


Fig. 1. Basic structure of a Siamese Network, note the red square indicates that the two models shares the same weights. (pyimagesearch.com)

In this study, we investigate the performance of different backbone architectures on Siamese Networks for one-shot learning tasks. One-shot learning is a challenging problem that requires a model to recognize novel objects or patterns with only one example. Our focus is on the use of a vision transformer, which has been shown to outperform CNN-based models in image classification and sees if it still has the high ground on one-shot learning tasks.

2. Background

2.1. Siamese Network

The Siamese Network was originally used to verify whether the signature on a check matches the bank's reserved signature (Bromley et al. 1993[1]), and later to compare the similarity between two inputs. It has since been gradually applied to the field of object tracking (Bertinetto et al. 2016[2]). The Siamese Network takes two inputs, input1, and input2, which respectively enter neural networks network1 and network2. Through the final loss calculation, the similarity of the two vectors generated by the two networks can be evaluated, i.e., the similarity of the two inputs.

Since the weights are shared in the Siamese Network, it to some extent limits the difference between network1 and network2, so it is usually used to handle problems where the difference between the two inputs is not very large, such as comparing the similarity between two images, two sentences, or two words. For problems with a large difference between inputs, such as the similarity

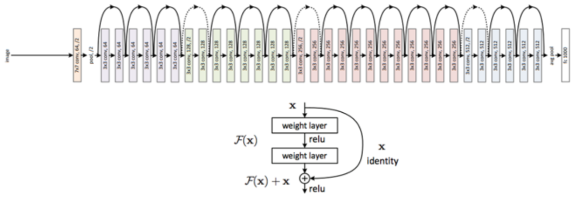

x
weight layer
relu
weight layer
$\mathcal{F}(\mathbf{x})$
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Fig. 2. ResNet50 model overview (He et al. 2015 [4]).

between an image and its corresponding text description or between an article title and its paragraphs, a pseudo-Siamese Network is needed.

2.2. Few-Shot Learning

Few-shot learning is the application of meta-learning in the field of supervised learning. Meta-learning, also known as learning, to learn, decomposes the dataset into different meta-tasks during meta-training to learn the model's generalization ability under category changes. During beta testing, the model can classify new categories without modifying the existing model. The training set in few-shot learning contains many categories, with multiple samples in each category. During training, C categories and K samples per category are randomly selected to construct a meta-task as the support set input of the model. Another batch of samples is then taken from the remaining data in these C categories as the prediction objects of the model. The task is to require the model to learn how to distinguish these C categories from the data, also called a C-way K-shot problem. In this study, we focus on one-shot learning.

During training, different meta-tasks are sampled each time, so the training includes different category combinations. This mechanism enables the model to learn the common parts of different meta-tasks, such as how to extract important features and compare sample similarities and forgets the task-related parts of the meta-task. The model learned through this learning mechanism can perform well in classifying new unseen meta-tasks.

2.3. Vision Transformer

ViT (Vision Transformer) is a model proposed by Google in 2020 that directly applies the Transformer to image classification (Dosovitskiy et al. 2020 [3]). Through the experiments presented in this article, the best model achieved an accuracy of 88.55% on ImageNet1K (after pre-training on Google's in-house JFT dataset), demonstrating that the Transformer is indeed effective in the field of computer vision and the results are quite impressive.

The structure of Vision Transformer looks very similar to the self-attention mechanism at first glance. It mainly consists of three modules:

- Linear Projection (Embedding layer of Patch + Position)
- Transformer Encoder
- MLP Head (classification layer)

In ViT, simply dividing the image into small blocks (patches) is not enough. The Transformer Encoder requires a vector with the shape of [num_token, token_dim]. For image data, a shape of [H, W, C] does not meet the requirements. Therefore, The image data is transformed into tokens through the Embedding layer with the position embedding. Then the vector is passed to the Transformer Encoder, which applies Multi-Head Attention, alongside Layer Normalization and DropOut layers. At last the MLP Block.

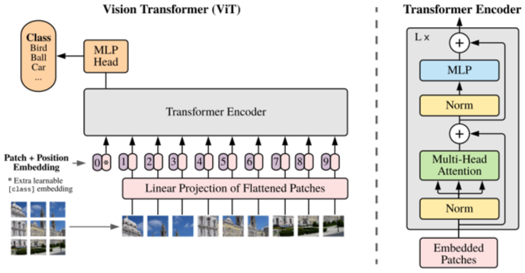


Fig. 3. ViT model overview (Dosovitskiy et al. 2020 [3]).

In the past, CNNs downsampled through convolution and pooling, theoretically allowing the model to increase the receptive field by deepening the model. However, In practice, CNNs respond weakly to edges. This is easy to understand, as pixels closer to the edge are convolved fewer times, so they naturally contribute less during gradient updates. In addition. CNNs can only compute correlations with adjacent pixels. Due to the sliding window convolution operation, pixels outside the local neighborhood cannot be jointly convolved, for example, the pixel in the upper left corner cannot be jointly convolved with the pixel in the lower right corner. This means that some spatial information cannot be utilized. In the Transformer, the correlation between each token is computed, which is very different from CNNs.

3. Methodology

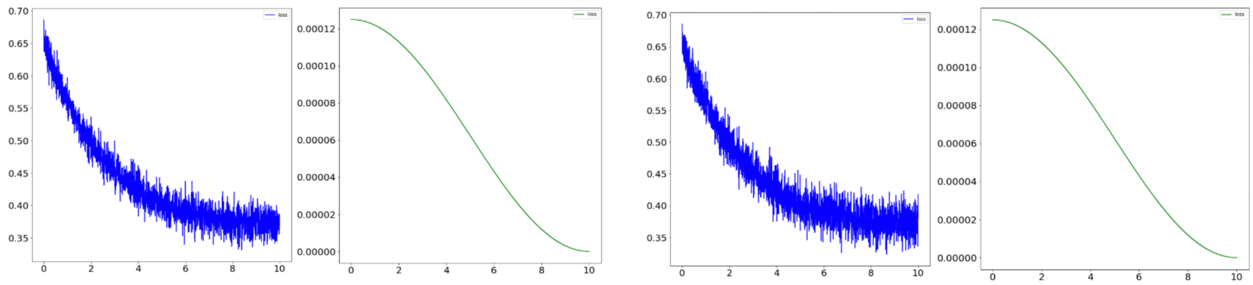
In this work, we focus on the factors including the sister network, pre-trained models, on one-shot learning using Siamese Network.

3.1. Dataset

For the dataset, we used the BIRDS 500 SPECIES dataset on Kaggle, which consists of 500 bird species and includes 80,085 training images, 2,500 test images (5 images per species), and 2,500 validation images (5 images per species). Each image contains only one bird, which occupies at least 50% of the pixels. Therefore, even a moderately complex model can achieve training and test accuracies in the mid-90% range. The images are original and not created by augmentation, and all are 224 x 224 x 3 color images in JPG format. The dataset includes three sets: training, test, and validation, each containing 475 subdirectories (one for each bird species).

3.2. Model

We used the ViT model pre-trained on ImageNet-21k (14 million images, 21,843 classes) at resolution 224x224 (Dosovitskiy et al. 2020 [3]) via Hugging Face. ViT is a transformer encoder model (BERT-like) pretrained on a



(a) ViT Siamese

(b) CNN Siamese

Fig. 5. (a) the loss (left) and the learning rate (right) of the ViT Siamese network. (b) the loss (left) and the learning rate (right) of the CNN Siamese network. Both models were able to converge to an optimal solution.

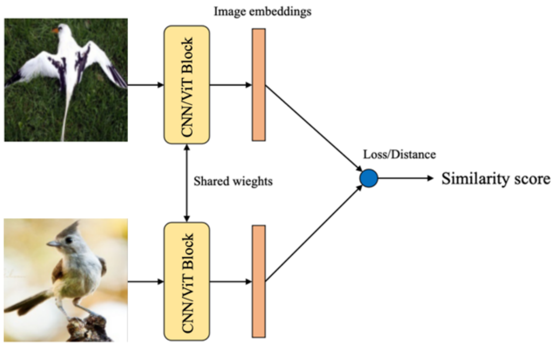


Fig. 4. The network architecture of the Simese Netowrk in this study.

Python VERSION	3.8.10
pyTorch VERSION	1.13.1+cu116
CUDNN VERSION	8302
Number CUDA Devices	1

Table 1. The comfiguration of the emvironemnt (Colab).

large collection of images in a supervised fashion, namely ImageNet-21k, at a resolution of 224x224 pixels. Note that the model we used does not provide any fine-tuned heads. However, the model does include the pre-trained pooler, which was used for downstream tasks, in this case, one-shot learning.

In addition, we used the well known ResNet model pre-trained on ImageNet-1k at resolution 224x224 (He et al. 2015 [4]) for out CNN counterpart. ResNet (Residual Network) is a convolutional neural network that democratized the concepts of residual learning and skip connections. This enables to train much deeper models.

3.3. Setup

We implemented our system in Google Colab. I used the official open-source "google/vit-base-patch16-224-in21k" model and "microsoft/resnet-50" model on Hugging Face. I used the BIRDS 500 SPECIES dataset downloaded from Kaggle and saved it as hdf5 files for later reading. And finally, we built a Siamese Network with two of the models mentioned respectively. Specifically retrieving the absolute distance of the features obtained and passing it to a linear layer.

Model	Accuracy
ViT Siamese	93.28
CNN Siamese	93.26
CUDNN VERSION	93.42

Table 2. The accuracy of the models and benchmark.

4. Result

The metric used is accuracy. Accuracy is usually expressed as a percentage and is calculated by dividing the number of correct predictions made by the model by the total number of predictions made. In other words, it measures the "rate of correct predictions".

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad (1)$$

As seen in Fig. 5., after 10 epochs, both models' losses were able to decrease and achieve a minimum. In addition, we can see that corresponding to our step algorithm, the learning rate decreases accordingly, which helps the model converge to the optimal solution. As the learning rate approaches 0, the model updates its parameters very slowly. This can be beneficial when the model has already converged to a good solution and further updates to the parameters may not be necessary or may even cause the model to diverge.

As seen in Table 2., the ViT Siamese Network was able to achieve 93.28% accuracy, compared to the CNN Siamese Network with 92.26%, which is slightly better. This shows that using ViT as the backbone of the Siamese Network helps with capturing image features that are critical for later similarity comparison. Note that this resonates with past researchs, which showed ViT defeats CNN based model on image classification tasks.

ViT is able to capture global contextual information from an image, whereas CNNs typically focus on local features. This is because ViT uses self-attention mechanisms to learn representations of the entire image, rather than relying on convolutional filters to identify local features. alongside with this study, it shows that one-shot learning as image classifiction are tasks that benefits from global context.

However, we would like to argue that both of the backbones used in this study are pre-trained and their weight was frozen. In other words, only a linear layer is being trained throughout the training process. We showed that by using pertained models (on a large dataset) directly as the feature extractor gives us good enough results, this could imply that image classification is not so different from on-shot learning, both learning to extract features from the image. However, whether this configuration affects the overall result of one-shot learning is left to be studied.

5. Conclusion

we conclude that both models were able to achieve a minimum loss after 10 epochs and that the ViT Siamese Network (slightly) outperformed the CNN Siamese Network in terms of accuracy. This could be attributed to ViT's ability to capture global contextual information from images using self-attention mechanisms, which is particularly useful in tasks such as one-shot learning and image classification that require feature extraction from images.

It is also noteworthy that using pre-trained models as feature extractors, with only a linear layer being trained during training, yields good results. However, it is important to investigate how this configuration may affect the overall outcome of one-shot learning.

In addition, Given the advantages of both Siamese networks and ViT, it is likely that combining these two approaches could lead to further improvements in performance on tasks that require global features. Future studies could explore the use of ViT-based Siamese networks in a wide range of applications, including natural language processing.

Furthermore, By combining the ability of ViT to capture global features with the pairwise similarity measurements of Siamese networks, ViT-based Siamese networks could be effective in tasks that involve complex, multi-modal data, such as medical diagnosis or autonomous driving; developing highly accurate and robust models for a wide range of real-world applications.

References

[1] Bromley, Jane & Bentz, James & Bottou, Leon & Guyon, Isabelle & Lecun, Yann & Moore, Cliff & Sackinger, Eduard & Shah, Rookpak. (1993). Signature Verification using a "Siamese" Time Delay Neural Network. International Journal of Pattern Recognition and Artificial Intelligence. 7. 25. 10.1142/S0218001493000339.

[2] Fully-Convolutional Siamese Networks for Object Tracking arXiv:1606.09549 [cs.CV]

[3] An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929 [cs.CV]

[4] Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs.CV]

[5] Siamese neural networks for one-shot image recognition. ICML deep learning workshop, vol. 2. 2015. 2015