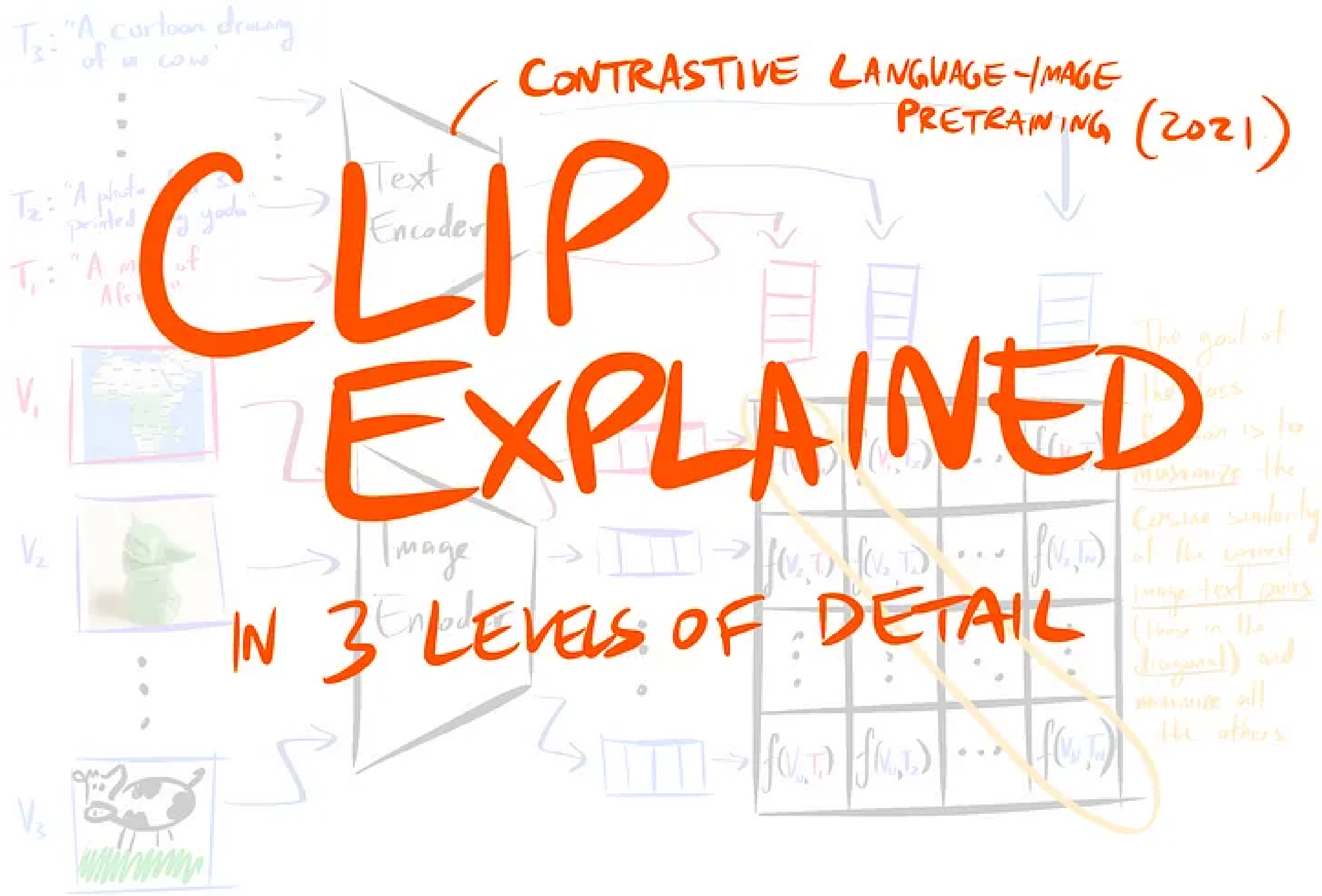
CLIP Paper Explained Easily in 3 Levels of Detail

🤔 And the key points to remember

Jeffrey Boschman · Follow

Published in One Minute Machine Learning · 4 min read · Jul 17, 2023



Can you answer these questions easily?

1. CLIP is a pre-trained model for telling you how well a given ___ and a given ___ fit together.

2. In training CLIP, the similarity scores of the correct image-text pairs are found on the ___ of the similarity score matrix for the current batch.

If not, keep reading! (see answers at the bottom of this post)

One Sentence Summary

CLIP is a *pretrained* model for telling you how well a given *image* and a given *text* fit together.

One Minute Summary

CLIP, which stands for Contrastive Language-Image Pre-training, is a model for telling you how well a given image and a given text caption fit together.

In *training*, it tries to *maximize* the "cosine similarity" between *correct* image-caption vector pairs, and *minimize* the similarity scores between all *incorrect* pairs.

In *inference*, it calculates the similarity scores between the vector of a *single image* with a *bunch of possible caption* vectors, and picks the caption with the highest similarity.

Note that CLIP is *not* a *caption generation* model, it can only tell you if some existing text caption fits well with an existing image or not.

Five Minute Summary

CLIP is a pre-trained model for telling you how well a given image and a given text caption fit together, introduced by the paper "Learning Transferrable Visual Models from Natural Language Supervision" (2021) from OpenAI. It was ***trained contrastively*** on a huge amount (400 million) of web scraped data of image-caption pairs (one of the first models to do this). It is useful because this pre-trained model can be used for a lot of downstream tasks; instead of just associating an image with a class label out of a set of class labels, it can ***associate an image with*** a text caption containing ***any words from the English language***.

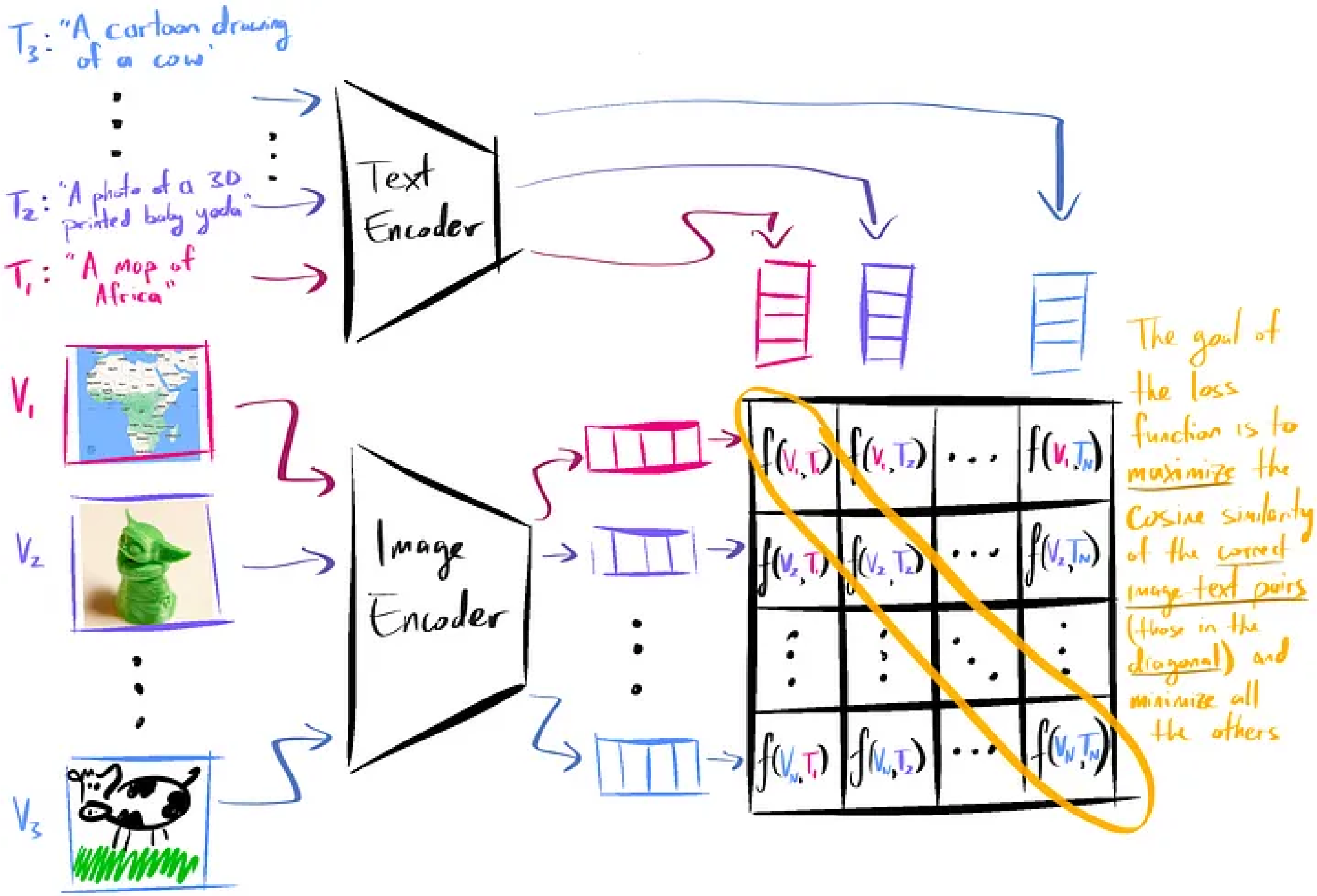
🏋️ Training

Input: A bunch of image-caption(text) pairs (all encoded to be vectors).

Output: The "cosine similarity" scores between *all* the image vector and caption vector combinations.

Objective function: A contrastive function that will modify the weights of the model such that correct image-caption pairs get a high similarity score, and incorrect pairs get low similarity scores.

Note — during training, the model requires that a huge batch of image-text pairs is fed at once (e.g., 20,000 pairs). That way, each batch contains 20,000*20,000 = 400,000,000 possible pairs, with only 20,000 being correct pairs. For efficient processing, the similarity scores of all possible pairs are computed at once to yield a 20,000 by 20,000 matrix, with the values in the ***diagonal*** being the similarity scores for the ***correct image-text pairs***. That way, the objective function can have the goal to maximize the scores in the diagonal and minimize all the scores not in the diagonal. See the figure below.



Overview of how CLIP works during training.

🤔 Inference

Inputs: the vector for a single image, and the vectors for a bunch of different possible text captions.

Output: the similarity scores of the single image to all the different text captions.

Goal: select the text which has the highest similarity with the image.

Note — the authors used this strategy to use CLIP for ***classification*** inference on the ImageNet dataset. They turned the label of each ImageNet class into a sentence, and used these sentences as the possible captions for a given image. For example, instead of using the ImageNet label "cat", they created a sentence like "A photo of a cat" because this is the type of text that CLIP is used to. Then they compared a given ImageNet image with the set of sentences that correspond to the different classes, and picked the sentence

The paper says they have ***high zero-shot performance*** — this is because even though the model might not have been trained on any examples of the classes in the ImageNet dataset, it still performs well because it could kind of figure out what the words of the classes mean and associate that with the images.

Other Notes

Although CLIP itself is not a caption generator model, the pre-trained CLIP model can be used to calculate similarity scores between images and captions, which could therefore be useful ***as part of*** caption generator models.

Answers to questions at the beginning

1. *Image, text (or caption)*
2. *Diagonal*

Links

- https://www.youtube.com/watch?v=dh8Rxhf7cLU — this YouTube video does a great job explaining the main points of CLIP visually.
- https://openai.com/research/clip — This is the official post about CLIP from OpenAI, which contains links to the paper and code.

Hey, have you become a Medium member yet?

If not, please consider signing up with my referral link to get access to ALL my articles, plus the articles from other smart and interesting people on the platform.

Deep Learning NLP Computer Vision Machine Learning AI

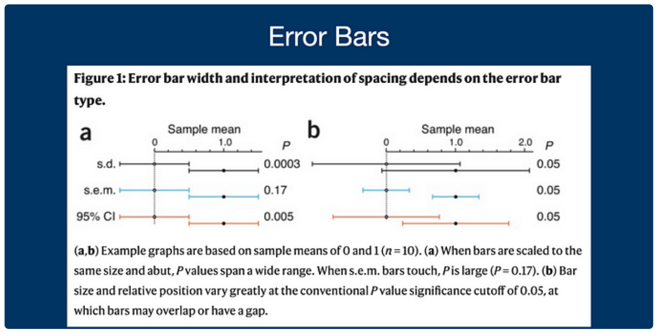
Written by Jeffrey Boschman

Follow

64 Followers · Editor for One Minute Machine Learning

An endlessly curious grad student trying to build and share knowledge.

More from Jeffrey Boschman and One Minute Machine Learning



Jeffrey Boschman in One Minute Machine Learning

How to Properly Interpret Error Bars

“The meaning of error bars is often misinterpreted, as is the statistical…

3 min read · Sep 8, 2021

186

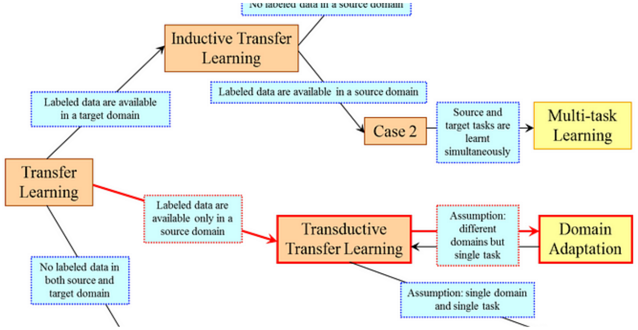
Jeffrey Boschman in One Minute Machine Learning

SimCLR Explained in Simple Terms

SimCLR explained simply in 1 sentence, 1 minute, and 5 minutes

9 min read · Dec 22, 2022

58



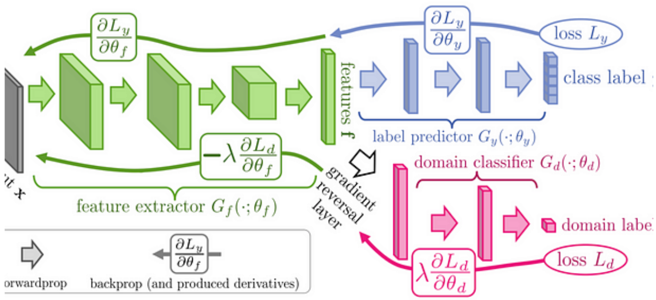
Jeffrey Boschman in One Minute Machine Learning

Transfer Learning vs. Domain Adaptation | one minute…

Are the terms transferable?

1 min read · May 19, 2021

61



Jeffrey Boschman in One Minute Machine Learning

Domain-Adversarial Training of Neural Networks (2016) | one…

DANN, that’s good.

2 min read · May 21, 2021

58

See all from Jeffrey Boschman

See all from One Minute Machine Learning

Recommended from Medium

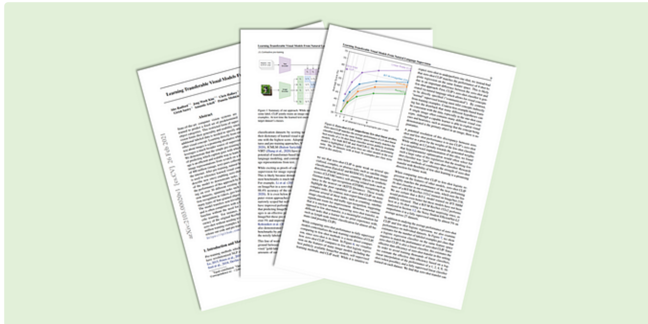
Enrico Randellini

Image and text features extraction with BLIP and BLIP-2: how to buil...

Connect images and text with the power of ViT and LLM to perform the image-text...

11 min read · Sep 26, 2023

39 1



Sascha Kirch in Towards Data Science

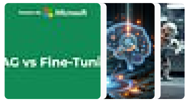
The CLIP Foundation Model

Paper Summary— Learning Transferable Visual Models From Natural Language...

8 min read · Aug 26, 2023

74

Lists

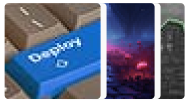


Natural Language Processing

1117 stories · 589 saves

The New Chatbots: ChatGPT, Bard, and Beyond

12 stories · 276 saves

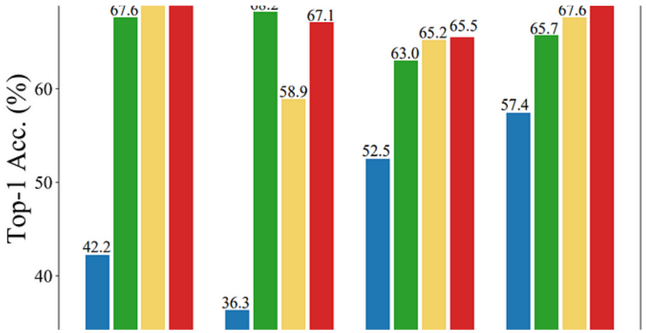


Predictive Modeling w/ Python

20 stories · 824 saves

Practical Guides to Machine Learning

10 stories · 953 saves



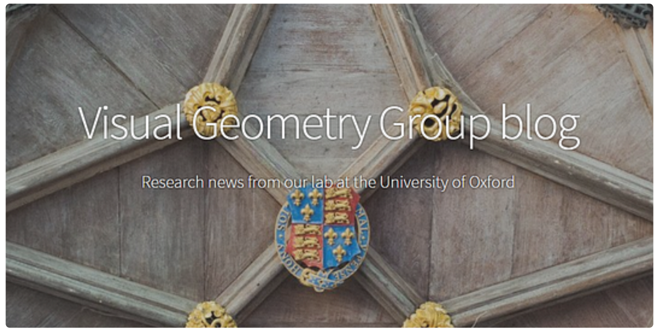
Juneta Tao

Self-Supervised Distillation

Self-Supervised Learning (SSL) methods perform much worse than the supervised...

3 min read · Aug 29, 2023

7



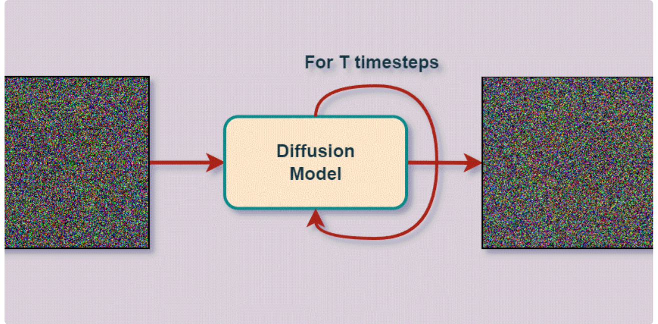
Sik-Ho Tsang

Brief Review — SeLa: Self-Labelling via Simultaneous Clustering and...

SeLa, Using Sinkhorn Distance for Self-Supervised Learning

6 min read · Aug 23, 2023

1



Gabriel Mongaras in Better Programming

Diffusion Models — DDPMs, DDIMs, and Classifier Free Guidance

A guide to the evolution of diffusion models from DDPMs to Classifier Free guidance

28 min read · Mar 13, 2023

616 8

Anirban Sen

Beginner's guide to one of the best Vision model — CLIP (Contrastive...

What is CLIP? Contrastive Language-Image Pre-training (CLIP for short) is a state-of-the...

6 min read · Aug 19, 2023

42 1

See more recommendations