CLIP (helper model of DALL-E)

CLIP stands for Contrastive Language–Image Pre-training.

CLIP is very good at telling if an image and a short
piece of text belong together or not.

But why do we want to know if a given text and image fit well together in the first place?

This task formulation is general enough to derivate many other tasks from it: Like image recognition on ImageNet, for example.
The task is to classify images. And this is what neural networks have been doing so far:
They take all available classes in the dataset up to 1000 other classes
and assign a probability for each class. The highest-scoring class is the classification prediction of the model. But with CLIP it gets even better:
Because of the way it was trained, CLIP is not restricted to the classes from the dataset, but knows virtually all English words, being able to formulate the ImageNet classes into prompts containing more language than just the classes. And here we already see why CLIP has the potential to be so much better than its competitors: Firstly,
it can also model the notion of “photo” or “image” or “satellite”, extending its capabilities far over a restricted set of classes, showing great zero-shot performance on unseen datasets.

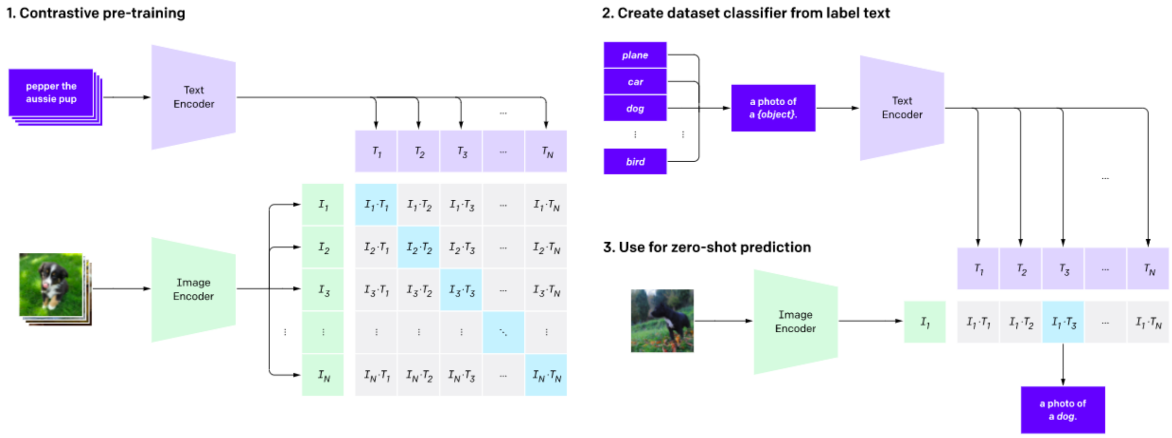
Secondly, CLIP’s training is never restricted to reducing an image to a single concept or word, therefore it never loses or forgets other aspects of the image that are not captured in the class, like the grass or wood in the background. Impressively, but not perfectly, CLIP can solve tasks and datasets it has not explicitly seen during training, like optical character recognition, geo-localization, texture detection, and others like facial emotion recognition, and action recognition (idk too much about so can’t rly explain). It is an impressive zero-shot performance for a model trained only to predict similarity between a text prompt and an image.

Why is CLIP successful?

The zero-shot capabilities of CLIP are extraordinary! CLIP's zero-shot capabilities are also prevalent in GPT-3 trained on huge amounts of data with high variance: CLIP has been trained on 400 million images with paired text descriptions.
This is 100 times more data than ImageNet, and it is very diversified data too
because it was all scraped from the internet. The more data the better and the more diversified the data, the more we can expect the model to be able to perform afterward in zero-shot.

Then on this data, CLIP was trained like this: The authors take a batch of images and a batch of their corresponding textual descriptions. The images are encoded, by either a ResNet or a Transformer and the text is encoded by a transformer too. Then for each image in the batch, the encoder computes an image vector; square similarities are computed between all these images and text vectors.
This means that the diagonal elements should be maximized by CLIP. In a contrastive fashion, the off-diagonal

elements should be minimized.
CLIP is predicting similarities between images and textual descriptions and training in a contrastive fashion (related to backpropagation).
But this idea is not new at all but have been seeing this idea in Multimodal Transformers as a pretraining task, like for example in LXMERT or ViLBERT, VisualBERT, and so on.


1. Contrastive pre-training
pepper the aussie pup
Text Encoder
Image Encoder
T_1 T_2 T_3 ... T_N
I_1 I_2 I_3 ⋮ I_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
$I_2 \cdot T_1$ $I_2 \cdot T_2$ $I_2 \cdot T_3$ ... $I_2 \cdot T_N$
$I_3 \cdot T_1$ $I_3 \cdot T_2$ $I_3 \cdot T_3$ ... $I_3 \cdot T_N$
$I_N \cdot T_1$ $I_N \cdot T_2$ $I_N \cdot T_3$ ... $I_N \cdot T_N$
2. Create dataset classifier from label text
plane
car
dog
bird
a photo of a {object}.
Text Encoder
3. Use for zero-shot prediction
Image Encoder
I_1
T_1 T_2 T_3 ... T_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
a photo of a dog.

CLIP is also computationally efficient (idk too much about this)

CLIP is trained to predict high similarity for fitting image-text pairs and low similarity for random ones.
Image recognition only has one word to describe one image,
CLIP rather likes more words, like a description, so the authors create a prompt, “a photo of a” and then they insert all possible objects, creating as many prompts as there are objects.
Then, they compute zero-shot the similarity between a say, ImageNet image, and this text prompts that differ only in the mentioned object. Then the prompt and object with the highest similarity are chosen as predictions. And the same CLIP model performs really well, not only on the classical ImageNet image recognition challenge but on many other tasks with different kinds of images, or styles because first: it has seen these image styles during training, and secondly, it has seen the words to describe these kinds of images and their details.

Limitations

Zero-shot CLIP is competitive with the baseline of a supervised ResNet-50 model, trained specifically on the training set of that dataset.
And this is impressive because it means that CLIP trained on OpenAIs dataset of 400 million images absolves a little training specifically on tasks and the datasets associated with them.
However, it needs a lot of work before beating the supervised baseline which can become infeasible due to the sheer amount of resources needed.

From paper:
"The performance of CLIP is poor on several types of fine-graind classification such as differentiating models of cars, species of flowers and variants of aircraft." And CLIP has also trouble with "abstract and systematic tasks such as counting the number of objects in an image."

CLIP has also trouble with tasks that are unlikely in the pretraining dataset by saying that CLIP has trouble generalizing to out-of-distribution images. And they illustrate with an example of optical character recognition, where CLIP is really good at recognizing digitally rendered text but has troubles with MNIST, which is a dataset containing handwritten digits. And the reason the authors can identify is that there are a lot of digitally rendered text images in CLIP's training data, but not so many handwritten digits, and CLIP fails to generalize from digitally rendered text to handwritten text.

The next limitation is that CLIP is not a caption generation model. It can only tell you how a given, an existing image and an existing text fit together, but it cannot really compose new text.

Some CLIP Applications

https://ieeexplore.ieee.org/document/9707036/
https://twitter.com/l4rz/status/13526...
https://twitter.com/haltakov/status/1...
https://twitter.com/nagaraj_arvind/st...