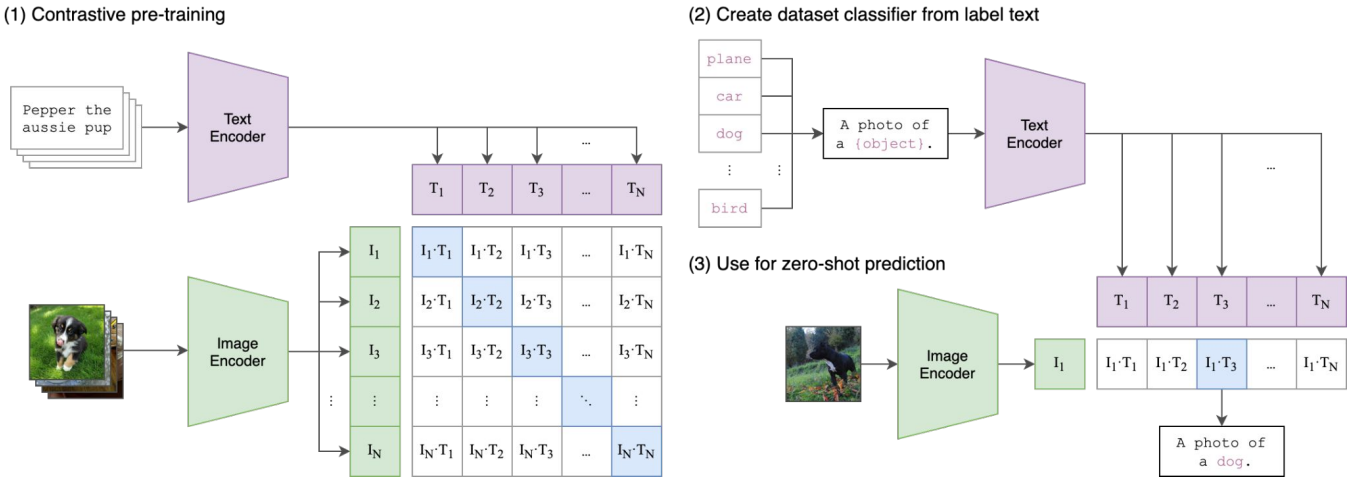
What is CLIP?

- CLIP is a neural network trained on a variety of (image, text) pairs
- It can be instructed in natural language to predict the most relevant text snippet, given an image (and vice versa), without directly optimizing for the task
- CLIP is thus similar to the zero-shot capabilities of GPT-2 and 3
- CLIP matches the performance of the original ResNet50 on ImageNet “zero-shot” without using any of the original 1.28M labeled examples

CLIP Architecture



Contrastive pre-training is a type of self-supervised learning technique to learn representations of data that are useful for downstream tasks, such as image classification or natural language processing.