CLIP 2021

Tuesday, July 8, 2025 1:46 AM

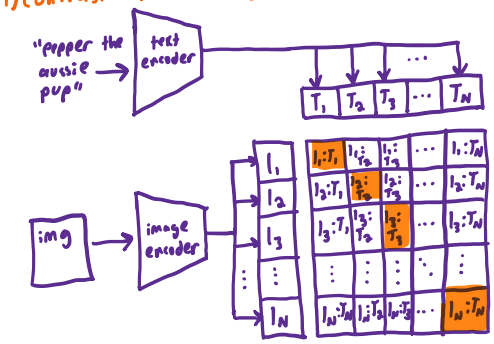
CLIP: Learning Transferable Visual Models From Natural Language Supervision

Abstract

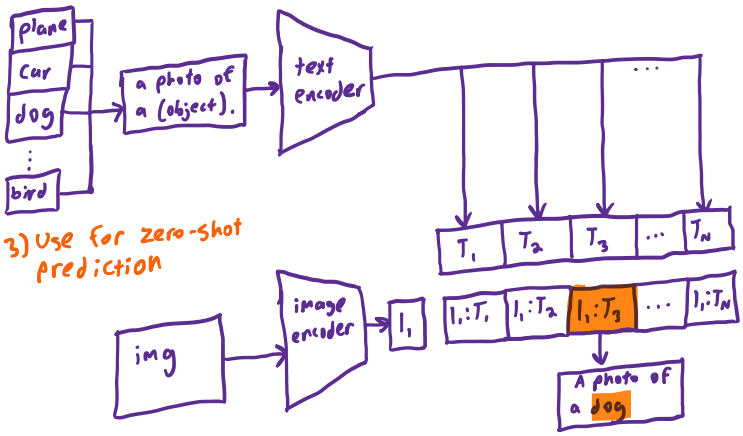
- Introduces SOTA model that learns (text, image) pairs
- performs well on 30+ datasets
- Isn't a generative model; rather, it learns (text, image) pairs

Introduction

1) Contrastive pre-training



2) Create dataset classifier from text label



3) Use for zero-shot prediction

Approach

- 400m (text, image) pairs
- 2 separate encoders:
  1) Image encoder: either Resnet or ViT
  2) text encoder: a Transformer like GPT
- Both map inputs to 512-dimensional vectors
- For N pairs, the model:
  - Embeds N images and N texts into shape (N, D)
- Computes cosine similarity between each image and text (NxN matrix)
- For each row:
  - Max similarity should be at the diagonal entries (correct text-image pair).
  - other entries negative examples

Loss Func: InfoNCE

Training Objective: train the encoders such that correct pairs are close together in shared embedding space - and everything else is pushed apart.

Note: CLIP is just 2 encoders that together learn to generate correct embeddings