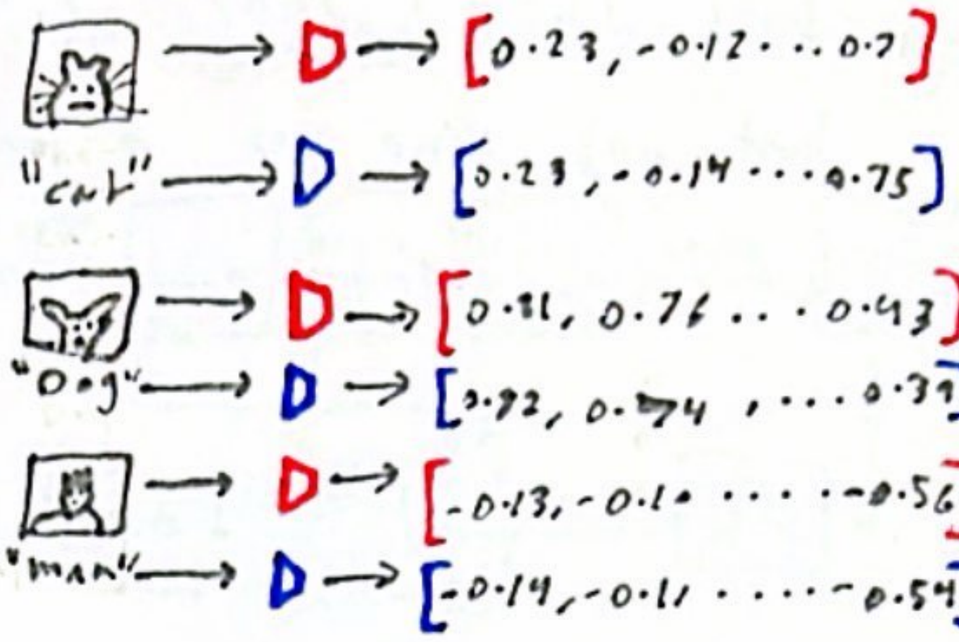
(Contrastive Language Image pre-training)

CLIP (gen AI notes)

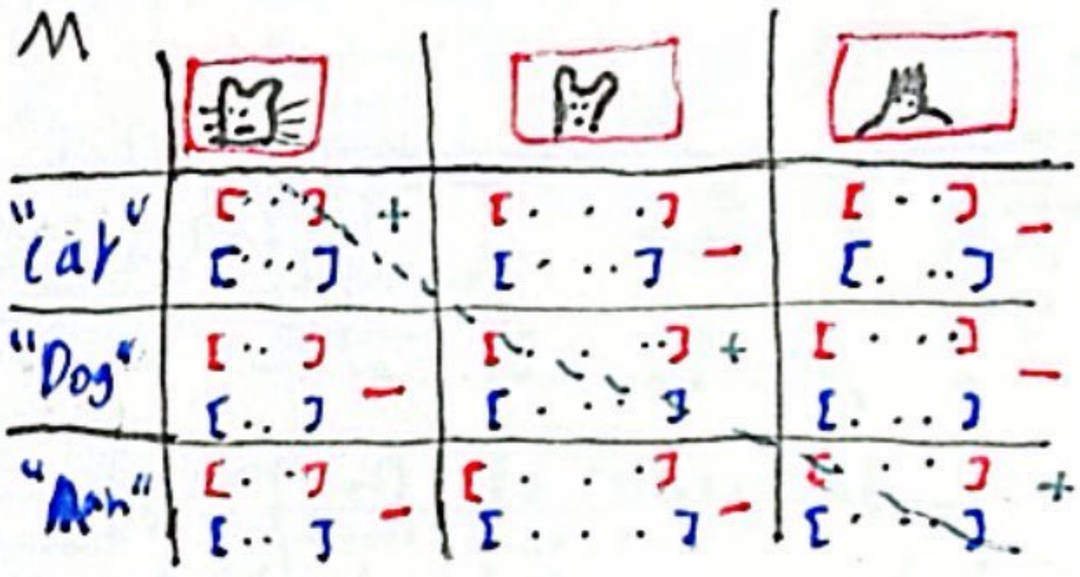
- in Feb 2021 @Open AI released a new model architecture called CLIP trained on 400 million Images and caption pairs from the net.

- CLIP is composed of two models one that processes text and one that processes images. The output of each is a 512 vector. The idea is that both vectors should be similar.

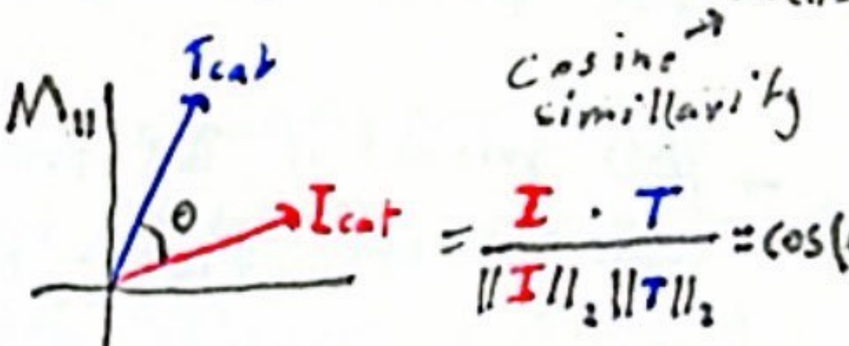


- from here the idea is to make use of the similarity not just between the corresponding images and captions but all image caption pairs in the batch when training the model.

- If we arrange our img vectors as a col and text vector as rows in a matrix the pairs of vectors along the diagonal = ... correspond to matching img and caption all other pairs are non matching images and captions. The CLIP training objective is to maximize the similarity between maching pairs + while minimizing the similarity between non matching pairs = −



- The algorithm measures similarity between M vectors using cosine similarity as the metric The vectors point in some high dim space. Cos Sim measures the cosine of the angle between our vectors "θ" in this space so if the angles between the vectors is 0 the cos sim score will be the max = 1 "↗" so the img and text models that make up clip are trained to maximize the alignment of related img caption pairs while minimizing alignment for non related pairs.



- This shared vector space is 78 called the embedding space →