





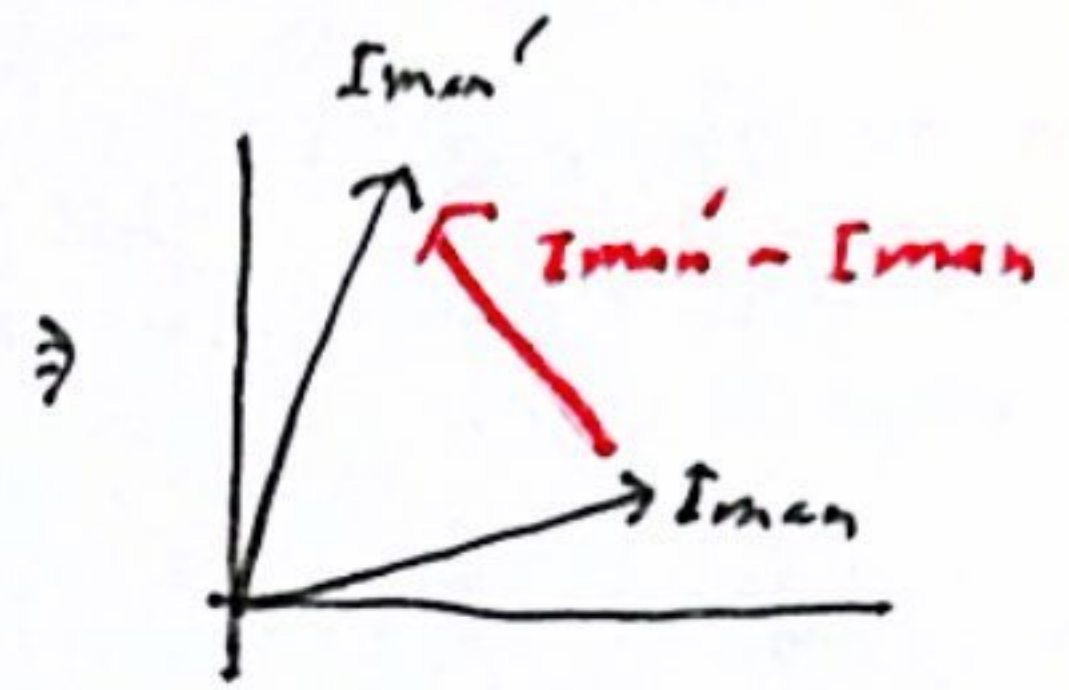
Shared Embedding Space (gemini notes)

- This space talked about in the last page is where the vectors live and it has some key properties

- Ex: if I take 2 imgs one of person wearing a hat and one of a person with no hat then pass it in the img encoder I get 2 vectors in this space

 →  → $I_{man} = [-0.13, -0.10 \dots -0.56]$

 →  → $I_{man'} = [-0.13, -0.10 \dots -0.50]$



- Now if we subtract the hat person's vector with the person with no hat vector we get a new vector in our space

⇒ What text does this new vector correspond to

- Mathematically we took the difference of the two. We can search for corresponding text by passing a lot of different words in the text encoder and getting cos similarity between the new vector and text vector

word	cos similarity
Melancholy	0.05
⋮	⋮
Top matches	
hat	0.165 ←
cap	0.113 ←
helmet	0.106 ←
angry	0.061

- This means if we take out the man if subtract man with hat from man we get vectors corresponding to a hat, cap etc. This means ideas like whether there is a hat in our img is translated into a distance between vectors in our space. This means CLIP can be a very good img classifier by passing a img into the img encoder and computing cos similarity (comparing) it to a set of possible captions (passed in text encoder) the top result was context of img and its label was the img label. So CLIP gives us a shared representation of img and text. BUT we can only map img/text to vectors not backwards.