Note on Multi-Modal generalization of CLIP

by Paul Thompson, 7/3/2025

Another insanely cool paper by Adriel Saporta and her team generalizes CLIP to any number of data modalities: [6]

Why is this ground-breaking ?

You have probably seen Generative AI make videos or images from a text prompt, explained here

<https://youtube.com/watch?v=fOORfzGjCTA&t=2260s>

For this to work, you need something like CLIP [1,2] which pre-trains an image encoder (like a Vision Transformer) and a text encoder (like BERT) to predict which images are paired with which text captions in a training dataset. With enough training data (400 million images with captions from the internet), you can use contrastive learning to push the feature vectors of matching image-text pairs closer together in the embedding space while pushing the vectors of non-matching pairs further apart. Once trained, you can use CLIP to go and retrieve images that match text and get text that matches images.

Building on CLIP: in a latent diffusion generative AI model (e.g. stable diffusion), you can use CLIP to help generate an image matching a text prompt using cross-attention [3,4] but what if there are more than 2 modalities? Adriel Saporta and her team show that you can extend CLIP to a multilinear form (see the very ingenious diagram) to lock onto features in a 3rd modality that are linked to 2 or more modalities - CLIP only captures the information between modality pairs. They explain this happens by targeting the total correlation between all modalities (i.e., all of their interactions [5]).

This is important, as you can now capture joint information between modalities much better than before, using auxiliary modalities. You can think of many scenarios where 2 modalities of data are needed to make a prediction of a 3rd one (or a whole image, which is also a prediction), e.g. predicting 3D amyloid PET from multimodal MRI and blood markers; their paper has a nice example of whether an ECG and labs collected at admission are predictive of a chest X-ray (CXR) taken shortly thereafter (as ECGs and labs are both safer than CXRs).

Bonus exercise: Can you see how this can be applied to empower imaging genomics (finding genes that affect multimodal imaging)? \*this could be a good PhD written exam question.

References

[1] [Connecting Text and Images, OpenAI blog, 2021](https://openai.com/index/clip/)

[2] [Introduction to CLIP and Multi-Modal Models, Course Module, Marqo](https://www.marqo.ai/course/introduction-to-clip-and-multimodal-models)

[3] [How does Stable Diffusion work? Tutorial on stable-diffusion-art.com, 2024](https://stable-diffusion-art.com/how-stable-diffusion-work/)

[4] [Faster Image2Video Generation: A Closer Look at CLIP Image Embedding’s Impact on Spatio-Temporal Cross-Attentions , Ashkan Taghipour et al, 2024](https://github.com/dimitarpg13/deep_learning_for_image_processing/blob/main/literature/articles/CLIP/Faster_Image2Video_Generation-A_Closer_Look_at_CLIP_Image_Embeddings_Impact_on_Spatio-Temporal_Cross-Attentions_Taghipour_2024.pdf)

[5] [Twitter feed of Adriel Saporta, 2024](https://x.com/ARSaporta/status/1854596579514061148)

[6] [Contrasting with Simile: Simple Model-Agnostic Representation Learning for Unlimited Modalities, A. Saporta et al, 2024](https://github.com/dimitarpg13/multi-modal_models/blob/main/literature/articles/Contrasting_with_Symile-Simple_Model-Agnostic_Representation_Learning_for_Unlimited_Modalities_Saporta_2024.pdf)