Implicit Multimodal Alignment: On the Generalization of Frozen LLMs to Multimodal Inputs

Anonymous Author(s)
Affiliation
Address
email

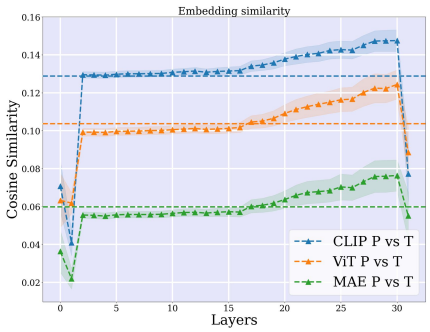

Figure 1: **Comparison of implicit multimodal alignment score across layers for different encoders.** CLIP models produce features that are most aligned to textual tokens across LLM layers. On the other hand, self-supervised encoders (*e.g.* MAE) produce the least text-aligned features. However, the relatively low cosine similarity score (closer to 0), reveals that the modality gap (*e.g.* different mutimodal narrow cones) still exists in LLMs, even for text-aligned encoders.

Table 1: **IMA score across different encoders.** We report the IMA score and the task performance with the ST setup (OPT). A positive correlation exists between IMA score and the performance; the most aligned encdoers (CLIP) have the best accuracy/CIDEr on VQA and captioning tasks.

LLM	Encoder	IMA Score ↑	COCO ↑	VQAv2 ↑	GQA ↑
			CIDEr (test)	Acc (Val)	Acc (Val)
Vicuna-v1.5	CLIP-ViT-L	0.130	127.63	63.05	54.34
Vicuna-v1.5	ViT-L (ImageNet)	0.105	116.76	61.27	51.57
Vicuna-v1.5	MAE-L	0.060	76.40	59.57	52.88

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.