Multimodal Embeddings

Jan Kels
Heinrich Heine Universität
Düsseldorf, Germany
Jan.Kels@hhu.de
and Omar Hassan
Heinrich Heine Universität
Düsseldorf, Germany
Omar.Hassan@hhu.de

Abstract

we present the concept of multimodal embedding and highlight the CLIP model architecture.

1 Introduction

multimodal embeddings combine different modalities such as text(language), audio and images.

2 CLIP

The CLIP (Radford et al., 2021) architecutre uses contrastive loss to learn a common embedding space for image, text pairs. It's based on consine similarity and can be used with a variety of text encoder as well as image encoders. The main task for CLIP models is zero shot image embeddings, where you construct a classifier by crafting a prompt. A single pretrained model can outperform task specific fine tuned models on various benchmarks. Not all. The blog post provides a great overview.

2.1 Applications

CLIP is the backbone for many text to image generation models such as DALL-E 2 (Ramesh et al., 2022) (”unclip”). Academic use such in the Visual Word Sense Disambiguation (Raganato et al., 2023) task.

2.2 following research

BLIP, BLIP2, cross attention from prompt2prompt paper, dino robustness?

References

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Alessandro Raganato, Iacer Calixto, Asahi Ushio, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2023. SemEval-2023 Task 1: Visual Word Sense Disambiguation. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, Toronto, Canada. Association for Computational Linguistics.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents.