Multimodal Transformer

Multimodal models are characterized by their ability to process multiple modalities, such as text, images, audio, or video, within a single model. The motivation behind these models lies in the idea that models should be able to understand real-world concepts in a way similar to humans. Humans can express the same concept across different modalities, “a cat”, for example, can be represented in text, image, or audio, and regardless of how the concept is expressed, the interpretation and understanding remains the same.

Please note that since our focus is on vision-language models, all further explanations will be based on the multimodality in the context of vision and language.

In the context of Deep Learning, this means that the representations of a concept should be the same (or at least close to each other), no matter if is expressed through text or image, which is also called alignment. However, in most existing models, this is not the case. These models are typically unimodal, meaning they process only one modality, making alignment of multiple modalities impossible. A naive approach would be to pass an image into an image model, and its caption into a text model. Even though the generated representations describe the same concept, they will not be the same, as both models are not related to each other. Each model will have a seperate latent space, as there has been no incentive for the models to learn a representation that is aligned across modalities (Figure 1), resulting in different representations for the same concept. While it is possible to compare the representations of two unimodal models, e.g. through cosine similarity, a similarity close to 1 (the maximum) does not necessarily mean that the concepts expressed in the representations are the same. There simply is no semantic relationship between the representations of the same concept produced by two unimodal models. A proof will be shown in (TODO: cite section where d2v2 image+text is used with retrieval).

To overcome this limitation, we need to develop models that can understand the same concept across different modalities, or input types respectively. They should map the input of different modalities into a common representation space, where the representations of the same concept are aligned, i.e. close to each other.

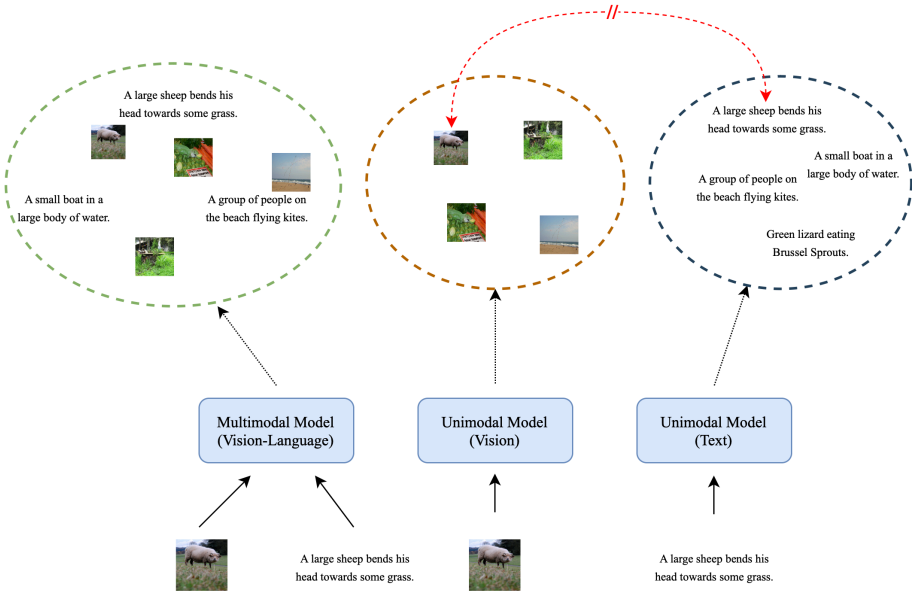

Figure 1: A multimodal model maps multiple modalities into a common representation space, where the representations of the same concept are aligned. In contrast, unimodal models map the input of a single modality into a modality-specific representation space. There is no alignment between the representations of the same concept produced by two unimodal models (indicated by the double slashed [//] arrow). While a comparison between the representations of two unimodal models is numerically possible, e.g. through cosine similarity, the similarity cannot be interpreted in a meaningful way.

Multimodal models consist of both unimodal encoders and multimodal components. Unimodal encoders are needed because of the inherent differences between modalities, e.g. image and text: Images are 2D and composed of pixels, while text is 1D and composed of words. Unimodal encoders encode the input into a modality-specific representation space, so they are normal unimodal models, e.g. a ResNet for images. In this work, all encoders will be based on the Transformer architecture.

Multimodal models require components that enforce a common representation space for the different modalities. There are two options: A multimodal (or shared) encoder, or a loss function (training objective).

The multimodal encoder is responsible for mapping the modality-specific representations into a unified/shared representation space, where representations should be independent of the modality. That means, the representations should not contain any modality-specific information, e.g. pixel information in images or single-word information in text. Only then representations of the same concept can be aligned, or close to each other under some distance metric, respectively.

To actually ensure that the representations of the same concept are aligned, and not only in the same space, a training objective is needed that pushes the representations of the same concept closer together in representation space, while pushing the representations of different concepts further apart. For vision-language models this translates to pushing the representations of an image and its caption closer together, while pushing the representations of an image and an unrelated caption (or vice versa) further apart. To quantify the similarity between two representations, a distance metric is used, e.g. cosine similarity. The loss function is usually the contrastive loss, and its implementation for vision-language models will be introduced in the next section. An illustration of a multimodal model is provided in Figure 2, concrete examples will be introduced in (TODO: cite related work).

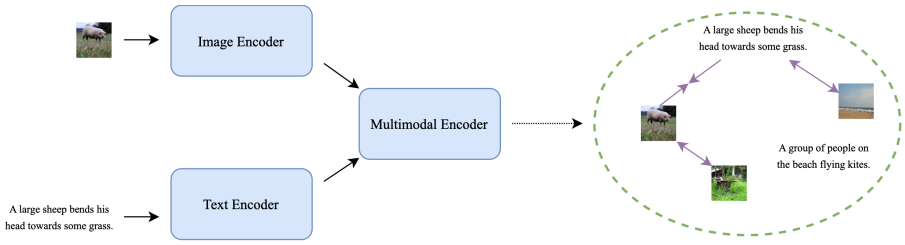

Figure 2: An abstract representation of a vision-language model. Image and text are first passed through unimodal, modality-specific, models (encoders), and then through a multimodal encoder that maps the modality-specific representations into a common representation space. A contrastive loss ensures the alignment and repulsion of similar and dissimilar concepts, respectively. We indicate this through purple arrows.