0.1 Multimodal Models

- are characterized by their ability to process multiple modalities, such as text, images, audio, or video, in a single model
- motivation is that models should be able to understand real-world concepts in a similar way to humans
- we can express concepts in different modalities, e.g., “a cat” in text, image, or audio -> no matter how we express it, the interpretation and the understanding of the concept remains the same
- in the world of AI models, that means the representation of a concept should be the same, no matter if it is expressed in text, image, or audio
- in most models, not the case -> are unimodal models, so just process one modality
- if we have a caption of an image, feed the caption into a text model, and the image into an image model, the representations of the same concept will be different
- we need to teach one model to understand the same concept expressed in different modalities -> multimodal models
- even though it is a multimodal model, so the same concept expressed in different modalities should have the same representation, the model still first needs to process the different modalities separately -> they are still different, e.g. image 2d and consisting of pixels, and text is 1d and consisting of words
- seperate processing is, as mentioned before, done by modality-specific encoders -> they encode the input into a modality-specific representation space -> just like normal unimodal models
- especially for the modality-specific encoders, any architecture of unimodal models can be used, e.g. ResNet for images, or BERT for text
- however, we will use Transformers
- multimodal models need a component that enforces a common representation space of the different modalities
- responsible for mapping the modality-specific representations into a common representation space
- we classify the component for alignment into three categories: alignment by space, alignment by representation, and alignment by loss
- introduced in the following sections
- since we use vision-language models only, all further explanations will be based on the alignment of vision and language

0.1.1 Alignment by Space

- image and text representations, created by image encoder and text encoder respectively, are mapped into a common representation space

-> the same space for all modalities

- this is enforced by a shared encoder, in which both the image and text representations are passed through, but seperately
- different to representation space of image and text encoder => image and text representation space are not related to each other
- shared encoder can be a normal Transformer layer

- since image and text are encoded as a sequence of tokens, the cls token output of the shared encoder is used as the representation

-> captures global information of the input -> single time step not useful for alignment, as a single time step in images represents a patch of pixels, and in text a word or subword

- however, just because the representations are in the same space, does not necessarily mean they are aligned -> need to enforce alignment

0.1.2 Alignment by Loss

- alignment is enforced explicitly by a loss function
- pushes the representations of the same concept, expressed in positive image-text pairs (so an image and its caption), closer together,

while pushing the representations of different concepts, expressed in negative image-text pairs (so an image and an unrelated caption or vice versa), further apart

- needs a distance metric to quantify the similarity between the representations -> cosine similarity is usually used

- loss function is usually contrastive loss, introduced in (TODO: cite contrastive loss section)

- often combined with alignment by space -> shared encoder maps the image and text representations into the same space, and the contrastive

loss enforces the alignment in this space

- but shared encoder not necessary for alignment if contrastive loss is used -> how is described in (TODO: cite clip section), where CLIP

is introduced

0.1.3 Alignment by Representation

- alignment is enforced implicitly by the representations themselves
- image and text representations, created by image encoder and text encoder respectively, are combined into a single representation

and passed through a shared encoder

- embedding itself is shared -> a single image-text representation is created
- in Transformers, text sequence and image sequence can simply be concatenated, and passed through a shared

Transformer block (vision-language block)

- Self-Attention captures cross-modal interactions, cls token of text can be used as image-text representation (shared representation)

Bibliography