Method

Model Architecture

- Before we build a first true multimodal model, we repeat the same procedure as with image before
- Normal unimodal distillation of a text model
- We use the same approach as with the image model
- ViT-B/16 architecture as teacher
- difference is the modality encoder, which is now a text encoder
  - consists of an embedding table and a learned positional encoding
- as with our image approach, this text encoder is frozen and shared between student and teacher

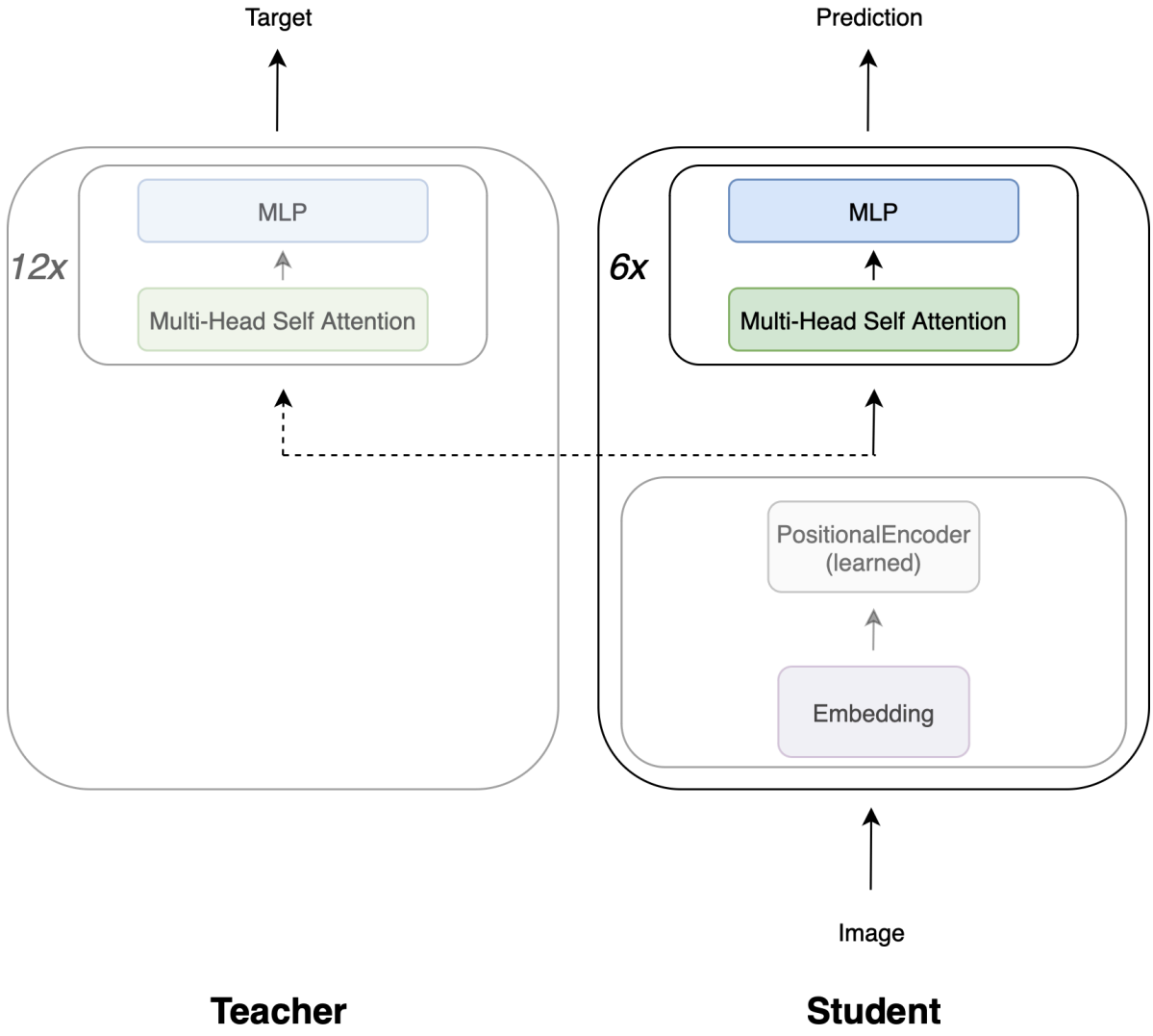

Figure 1: Test

Validation

- For validation we use a subset of the GLUE benchmark datasets Quora Question Pairs (QQP) and

Microsoft Research Paraphrase Corpus (MRPC), and perform zero-shot retrieval

- QQP is a dataset of question pairs with binary target -> questions are duplicates (so same semantic meaning)

or not

- MRPC is a dataset of sentence pairs with binary target -> sentences are paraphrases (again, same semantic meaning)

or not
- in this case we do not do classical classification as for the image model, which was also based on retrieval,

but we split both datasets into two parts, the first contains question1/sentence1 and the second question2/sentence2
- the task in now to find the matching sentence/question in the second part
- so for QQP, for a given question1, we want to find its duplicate in the second part (question2)
- for MRPC, for a given sentence1, we want to find its paraphrase in the second part (sentence2)
- in both datasets question/sentence pairs can also not be duplicates/paraphrases, so we have a negative class
- this is why we take, for both datasets, only the positive class, so duplicated questions and paraphrased, respectively
- we do this procedure for the train and dev set of both datasets seperately, meaning we have 4 datasets in total for which

we do zero-shot retrieval
- we also limit the size of the datasets to 25,000 sample each, which is enought to give insights into the model's ablitity

to create semantic rich embeddings, and make the validation process faster

- this constraint only affects QQP, as its train dataset is a lot larger (363,846 question pairs), as is the dev set (40,430 question pairs)
- therefore, each datasets has a maximum of 25,000 positive pairs
- we always take the first 25,000 positive pairs, if that many are available
  - QQP train: 25000
  - QQP dev: 14885
  - MRPC train: 2753
  - MRPC dev: 1147
- QQP and MRPC datasets, as well as their splits are taken from the glue benchmark website
- we now have a binary classification task
  - for a given question/sentence, we either find its duplicate/paraphrase (correct prediction) or not (wrong prediction)
  - we therefore use recall as our evaluation metric
- as done with multimodal retrieval task on e.g. coco in BEiT, FLAVA and VLMo, we not only use the normal recall, called

recall@1, but also recall@5, meaning we have a correct prediction if the correct question/sentence is in the top 5 matches
- to measure similarity of embeddings we again use cosine similarity
- as in image zero-shot, we not only calcualte both metrics for one dataset once, but twice, by taking the inverse
- so if for e.g. QQP train all question1s are the memory bank and all question2s are the query, we also do the opposite
- combined we have 4 datasets, 2 variants for each dataset, and for each of those 2 metrics, recall@1 and recall@5

- we therefore have 16 metrics in total, from which we take the mean as our final validation metric

First Test

Initialization

- With the though that we later want to combine text and image into one model, and we already showed that initializing the student

with teacher weights is beneficial, we also do this for the text model

- Leads to improvement...
- We want to utlize this improvement also later for the multimodal model
- Question is, how do we initialize the multimodal model?
- there are two parts we need to think about in the trainsformer blocks
  - the mlp blocks
  - the self-attention blocks
- for the mlp blocks, we can either choose to have one mlp block for text and image together (shared)
  - will keep params the same, but might degrade performance, but block now has more data to learn from
- ... or we can have separate mlp blocks for text and image
  - will increase params, but might increase performance, as each modality has its own mlp block
  - this is MoME used in VLMo and BEiT
  - has the advantage that we can initialize the mlp blocks with pretrained weights from the unimodal models
  - source unimodal models could either be the distilled models or the pretrained models
  - only the last layers (last two or one) would definetly have shared mlp -> vision-languange experts
    - so of the mlp layers, only those would have to be trained
- which to do ultimately has to be tested
- same thing for the self-attention blocks, can be shared or separate
  - usually it is shared, as done in VLMo and BEiT
  - VLMo even showed that for text pretraining, initializing self-attention blocks with pretrained weights learned from image pretraining

  can be used for text pretraining, without finetuning them, i.e. they are frozen
- Therefore, we want to have shared self-attention blocks for text and image
  - saves parameters
- We test if the same approach of VLMo also works here, by also doing KD with a student model that has self-attention blocks initialized with pretrained weights

from the image d2v model, which are also frozen

Misalignment of Positional Embeddings

Problems with unused (not pretrained) eos token (EOS token for sequence boundary)