Differences to Unimodal Knowledge Distillation

- for multimodal Knowledge Distillation, we also need a teacher model
- question is, which model should be the teacher model, or rather, in which modality should the teacher have been (pre-)trained?
- in our case, should the teacher be an image model or a text model?
- why not both?, so why not have a teacher text model and a teacher image model, so two teachers?
- because then the model would have to regress two representations, one for the image and one for the text
- would mean for an image-text pair, the student would regress two representations, one for the image and one for the text
- because both teachers are not related, the representations of the image and text, we would like to learn, are not aligned and related in any way
- recall that a multimodal model always has at least one shared block at the end of the architecture
- constrains the model that the same representation has to be produced for an image and its corresponding text/caption
- only then we can align the representations of the image and text
- if we would now have two teachers, one for the image and one for the text, and the student would regress both representations, then we would have two targets for one image-text pair, but we can only predict one representation, which should be the same for the image and text
- also, with two targets the model would have to learn two different representations for the image and text, and would most likely not learn

anything meaningful, as it is not possible, and not desired, to learn two different representations at the same time, i.e for an image and its corresponding text

- previously, in unimodal knowledge distillation, we were able to regress all time steps of the teacher model with the student model
- means the representation of each patch or text token, respectively
  - included the CLS/BOS token
- for multimodal Knowledge Distillation, we can't do this
  - we have to regress the whole image/text representation
  - and not the representation of each patch or text token
- has two reasons

1.

- image and te