Adressing Unimodal Teacher

- while our approach reaches reasonable performance compared to SOTA models, considering the model is less than 50% of the size of the SOTA models, and uses approx. 10% of the training data, we identify the self-supervised nature of the teacher as a problem
- comparing current approach which supervised teacher, where KL-Div is used, we reach lower performance, e.g. on ImageNet-1K zero-shot
- happens, even though supervised teacher used was ResNet-50, which has just 25 million parameters
- compared to that, our self-supervised teacher, BEIT v2, has approx. 86 million parameters
- if we compare the training loss of MSE between image representation output (cls token) of our student and image representation

output of theteacher, and the that between the text representation output (cls token) of our student and image representation output of the teacher, we can see that loss between image representations significantly lower, and also does not saturate as quickly as the text

- makes sense
- BEIT v2 is image-only (unimodal) model, output representations of that model are in image space
- even though representation of cls token gets more abstract the deeper we go into the model, there has been no incentive for the model to learn a representation that is modal-invariant -> so indepedent of the image modality
- so obviously, trying the regress the image representation of the teacher with an image representation of the student

works better than trying to regress the image representation of the teacher with a text representation of the student

- so using MSE to make the representations exactly the same is not a good idea, even if the results on image-text retrieval

and zero-shot classification are significantly better than random chance, and even close to popular models like CLIP or FLAVA

clustering in theory possible, but representations 768-dimensional, even with large amount of data, hard to cluster