Retaining Pretrained Knowledge

In the previous section (TODO: cite) adding a TTE before the pretrained encoders lead to a decrease in performance.

- driving factor was destroyed features -> made it difficult for the model
- based on the danger of destroyed features from pretrained models, we investigate whether other components of

our setup also have this effect

- we study the learning rate, which is currently the same between the pretrained encoders and the shared Transformer block,

with 2e-4 (base lr of 1e-4 was scaled according the the linear scaling rule [1])

- while this lr is already low, an empirical study, called METER (TODO: cite), showed significant improvement of

vision-language models with pretrained modality-specific encoders, when the learning rate of the pretrained encoders was significantly lower than that of the shared encoder, or Transformer blocks respectively

- we also experiment with retraining all features learned by the image and text encoder, and freeze all but the last layer
- this also allowes us to add additional shared Transformer blocks on top of the model, as no gradients are required

for most of the modality-specific encoders

- we do not freeze all layers of the image/text encoder, as we do not take all layers of the Data2Vec2 models, but only half of it
- the raw features returned by this subset model of the original might be too low level, in order for the shared blocks

to achieve alignment, which is why we do not freeze the last layer of both image and text encoder

Bibliography

[1] P. Goyal *et al.*, "Accurate, large minibatch SGD: Training ImageNet in 1 hour," *arXiv preprint arXiv:1706.02677*, 2017.