Ablation Study: Removing ITC

- we have problems with extending ITC
- approach require more complex components, like momentum encoder, memory bank, or more and better hardware, or more complex operations, requiring too much memory, and therefore also hardware
- ITC is a problem
- best approach would be to not use contrastive learning -> all the work we did before was because contrastive learning requires large number of negative examples or because we want fine-grained representations
- but can we remove it?
- BEiT-3 give good reason to discard it
- generally, self-supervised learning pre-training allows alignment of modalities even without ITC
- retrieval application then truly becomes zero-shot, and authors report high performance
- [1] mentions that without ITC, model outputs are not aligned, and therefore not suited for retrieval
- ablation study in (TODO cite) shows: TODO
- for current approach representations will still be aligned, because we are regressing the `[I_CLS]` token of BEiT-2, not a probability distribution over the classes
- so we are regressing actual features -> loss of 0 would mean representations are exactly the same and therefore aligned

Bibliography

[1] Y. Aytar, C. Vondrick, and A. Torralba, "See, Hear, and Read: Deep Aligned Representations," *arXiv preprint arXiv:1706.00932*, 2017, [Online]. Available: https://arxiv.org/abs/1706.00932