Research Questions

(Our contributions:)

(Single multimodal model possible?)

How does it perform compared to specialized models (single-modal)?

Can it be used as a general go-to pre-trained model to fine-tune to different downstream task? (Like EfficientNet for images, but for all tasks?) -> Is it a good transfer learner?

Do the representations match across modalities?

Are they still aligned after corruption/augmentation?

Can we go smaller? -> A step away from billion-sized models as feature extractors

Multi-modal Variational ViT Autoencoder for representation learning and latent-space arithmetic

Structure

1. Introduction
2. Representation Learning
3. Multimodal Learning
   1. Alignement of input data
   2. Data2Vec
   3. VLMo (They do not have any task where they mask parts of the image, why?)
4. Methodology
   1. Comparing models (compare with known models, and with single-modal models of the same size+architecture)
   2. Datasets used
   3. Metrics (Benchmarks)
5. Resarch/Experiments
   1. Choosing the right architecture (prob. ViT)
   2. Teacher-Student ViT
   3. ViT Autoencoder
   4. Analysis of/Avioding Collapse
   5. Study on different Loss-functions
   6. Performance after fine-tuning (benchmarking to compare with e.g. efficient-net?)
   7. Uncurated vs. curated dataset
   8. Effect of different Masking and data augmentation
   9. Only multi-modal examples vs. multi-modal and single-modal examples
   10. Single task vs. multi-task (VLMo)
6. Analysis and resulting behavior (What can the network do? What not? Maybe already in 5.?)
   1. Retrieval/Query similarity (explained below)
   2. Latent Space Arithmetic (inter- and intra-modal)
   3. Transitivity between modalities (synchronization -> <https://arxiv.org/pdf/1706.00932.pdf> end of chapter 2)
   4. Automatic input discrimination (For Teacher-Student and AE Encoder -> Do not tell the model through e.g. flag or similar if it receives text or image)
   5. Suitability as feature extractor (or fine-tuner) (… on multi-modal or single-modal tasks)
      1. Fine-tuning
      2. Low-shot
      3. Few-shot
7. Outlook
8. Conclusion

## Retrieval similarity

Pass a lot of data through the network during inference (text, image (,speech)) and record their representation in higher layers or of last hidden layer.

Pass example text, image or sound through network and find top-n most similar representations recorded before.

Visualize!

Augment same input (image probably best), and do again. What changes in top-n?

(=> Basically the same seen in the Deep Learning lecture!)

-> Do the exact same but with one of the single-modal models!

-> E.g. use image model and multi-modal model, pass a lot of images (exactly the same) through networks and record, then take query images and retrieve most similar outputs. Are the retrieved images similar between both networks?

Use <https://www.researchgate.net/publication/221573542_A_New_Approach_to_Cross-Modal_Multimedia_Retrieval> ? A New Approach to Cross-Modal Multimedia Retrieval