Research Questions

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?

Structure

1. Introduction
2. Representation Learning
3. Multimodal Learning
   1. Alignement of input data (Data2Vec?)
4. Methodology
   1. Comparing models (compare with known models, and with single-modal models of the same size+architecture)
5. Resarch/Experiments
   1. Choosing the right architecture (prob. ViT)
   2. Choosing the right training process (ViT Autoencoder, for arithmetic, Teacher student)
   3. Analysis of/Avioding Collapse
   4. Study on different Loss-functions
   5. Suitability as feature extractor
   6. Performance after fine-tuning
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
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