Methodology

Tools

Software:

- for all implementations we use pytorch lightning
- provides high-level functionalities on top of pytorch
  - like checkpointing, logging, distributed training, etc.
- we do not need to implement them in pytorch manually (to some extend)
  - pytorch already provides a high-level API but it is more prone to errors
- we save time and can focus on the actual implementation
- errors in vanilla pytorch are likely and hard to debug
- research will enevitably involve a lot of trial and error (experimentation)
- to keep track of all experiments, we use the experiment tracking tool Weights & Biases

Hardware:

- it is not possible to train the models, used in this work, on the CPU

-> GPUs are a requirement

- should be relatively new -> should be able to handle models upwards of 50 million parameters, but should not be too expensive
- we are severely limited by financial constraints, as there is not external funding for this work
- GPUs rented in the cloud
- we do not use popular cloud services like AWS or GCP -> too expensive
- instead, we use the smaller provider runpod.io
- has a high variety of consumer-grade, and enterprise-grade GPUs, much more affordable
- we opt for the (consumer-grade) NVIDIA RTX 4090
  - has one of the highest speeds (TODO: cite?) but lacks high VRAM (only 24GB)
    - is a problem we will address later
  - at the time of this research (June 2024), comes in at around 0.75 USD per hour
  - higher VRAM GPUs, like the A100, are available for 1.89 USD per hour
    - too expensive in the long run

Experimental Approach

- we will start as simple as possible
- always build on the results and knowledge of the previous steps
- to first validate if Knowledge-Distillation, the approach we will use throughout this work, even works for us, we will first test KD of unimodal models (e.g. distilling a ResNet-50 from a ResNet-101 on ImageNet), an area which has already been researched extensively
- from this, we will advance to the actual goal of this work: Multimodal Knowledge-Distillation
- as this is increasingly more difficult than distilling a unimodal model from another unimodal model of the same architecture, we will start with a supervised teacher
  - means, the teacher model has been trained on labeled data, and provides us with logits, and therefore a probabilty distribution, to regress
    - is basically a reproduction of SHRe
    - has been proven to work with this paper as a proof-of-concept
- if this approach works likewise for us, we will advance to a self-supervised teacher

- recall that goal was build a model/procedure for multimodal KD completly unreliant on labeled data
  - also means teacher, or any pretrained module that might be used, can’t be trained on labeled data
  - goal of this work is to check if this is possible
  - as mentioned before, VLMo for example use a BEiT module pretrained on labeled data as part of their model
    - this is not end-to-end self-supervised