Data and Preparation

General:

- data we need to collect has to be both unimodal and multimodal
  - multimodal obvious -> needed to align modalities, as described in e.g. section about "(See, Hear, and Read:) Deep Aligned Representations"
  - multimodal means in this case dataset of image-text pairs
  - unimodal -> needed for first tests and poc of distillation process
  - unimodal data also needed for evaluation and comparison of unimodal models from research papers, like Data2Vec, as well as comparison between unimodal and multimodal distilled models (models of this thesis)
  - unimodal data also needed for stage-wise knowledge distillation in section about "Mixing Positional Encodings" (will be elaborated on in the respective section)

Data Selection and Collection:

- starting with unimodal:
- collecting unimodal data is not a problem, many highly curated and large datasets available
- for image data, we select imagenet
  - build for image classification and object detection
  - each image corresponds to one of 1k classes
  - very popular, high quality, high variety -> 1000 classes, by standards of SOTA models it is a medium sized dataset (ca. 1.2M train images)
    - why medium sized? -> papers used in this thesis have been trained on much larger data -> VLMo around 14 million image(-text) examples,

    BEiT on more than 35M, and FLAVA (only mentioned a couple of times) even 70M
    - models also much larger than models build here, so we do not need as much, nor is it feasible for us to train on that much data
  - Data2Vec Image model exclusively trained on imagenet
  - Data2Vec, BEiT, VLMo, and FLAVA all use imagenet for evaluation -> we should use it as well
  - we use the full dataset of the 2012 version, with 1.2M images for training and 50k for validation
  - during pretraining, i.e. the Knowledge-Distillation, we apply the same data augmentation as in Data2Vec2
    - as we also use Data2Vec2 as the teacher in many experiments, this allows us a close comparison between the models, as they are trained on the same data
    - data augmentation includes random resized crop, followed by a random horizontal flip, followed by normalization each channel seperatly using channel wise mean and standard deviation computed from the imagenet training set -> standard procedure for image preprocessing, and used for all images throughout this work
    - random resized crop: crop a random part of the image, then resize it to the desired size (224x224)
      - crop size is hyperparameter, but we just use the same as in Data2Vec2, which is 0.08 to 1.0 of the original image size
      - 8% as lower bound seems to be a very low value, a lot of information is lost, but it is a common value in the literature, so we do

    the same

        - random horizontal flip: randomly flip the image horizontally (self-explanatory)
    - for validation, we resize each image to the same size as in training (224x224) and normalize it using the same procedure as in training
    - generally ALL images in this thesis will be scaled to the same size: 224x224
    - we access the data from Huggingface's dataset hub
- for text which dataset we use is less relevant -> no benchmarks published for nlp that rely on one specific dataset, like imagenet in vision
- also often a mix of mulitple text datasets used
- benchmarks published on downstream tasks/dataset like GLUE, more on that later
- for text data (pretraining, which is the Knowledge-Distillation that we do) we select openwebtext
    - dataset build to reproduce datasets used to train GPT-2
    - publicly available, used by e.g. BEiT3
- we access the data from Huggingface's dataset hub
    - published as slices and without any split
    - we take subsets 0-4 for training and 5 for validation, which is about 25% of the data