Data selection

Finding data points which can help the neural network learn and are relevant to the target dataset.

What is my task?

My task is to use efficiently select a subset of data points from a web scale dataset.

- Suppose I have 50,000 hours of dataset. Can I select a 500 hrs of relevant utterances from this dataset.
- I want to train a neural network with this 500 hrs of utterances. I expect this network to perform better than a network trained with 50,000 hours of dataset.

Related Work?

- ICML 2021 had a relevant workshop, it is based on subset selection. https://icml.cc/Conferences/2021/ScheduleMultitrack?event=8351
- In this conference there was a paper with the title: GoldiProx Selection: Faster training by learning what is learnable, not yet learned, and worth learning. They select those data points from the training set which improve the loss value on the development data.

What is their Advantage and Drawback?

- They have a selection approach which directly includes the target dataset.
- The approach is slow and do not involve efficient ranking of the data points.

What is the advantage of our approach:

- we aim to explicitly model relevance of the data points to the target dataset.