Compressing Self-Supervised Models

Tzu-Quan Lin[1], Chun-Yao Chang[1], Huan Yang[1], Guang-Ming Chen[1], Tzu-Hsun Feng[1], Hao Tang[2], Hung-yi Lee[1]

[1]National Taiwan University, [2]The University of Edinburgh

Universality

Self-supervised models enable semi-supervised learning for various downstream tasks.

Universality

Self-supervised models enable semi-supervised learning for various downstream tasks.

Usage

- Feature extraction
- Fine-tuning

Goal

Can we find small networks (e.g., subnetworks) that enjoy the same universality?

Goal

Can we find small networks (e.g., subnetworks) that enjoy the same universality?

Approach

- Low-rank approximation
- Pruning
  - Weight pruning
  - Head pruning
  - Layer pruning
- Distillation
- Other architectures
- Anytime inference

Progress

- Baseline (Tzu-Quan Lin)
- FLASH (Chun-Yao Chang)
- Structured pruning (Huan Yang)
- Weight pruning (Tzu-Hsun Feng)
- Low-rank approximation (Guang-Ming Chen)

Progress

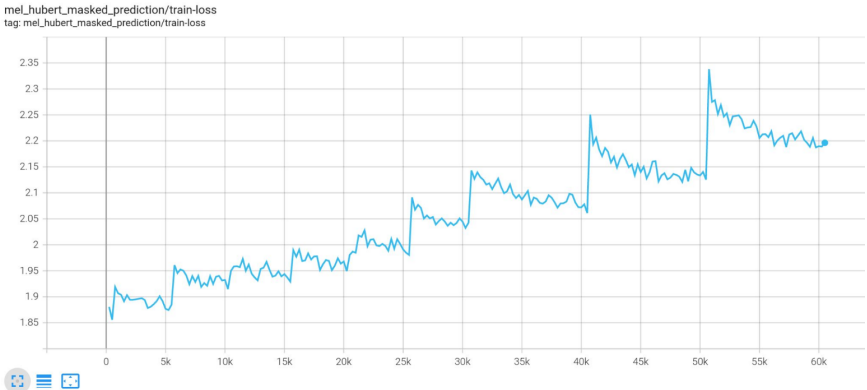
- **Baseline (Tzu-Quan Lin)**
- **FLASH (Chun-Yao Chang)**
- **Structured pruning (Huan Yang)**
- Weight pruning (Tzu-Hsun Feng)
- Low-rank approximation (Guang-Ming Chen)

ℓ_1 norm of the heads

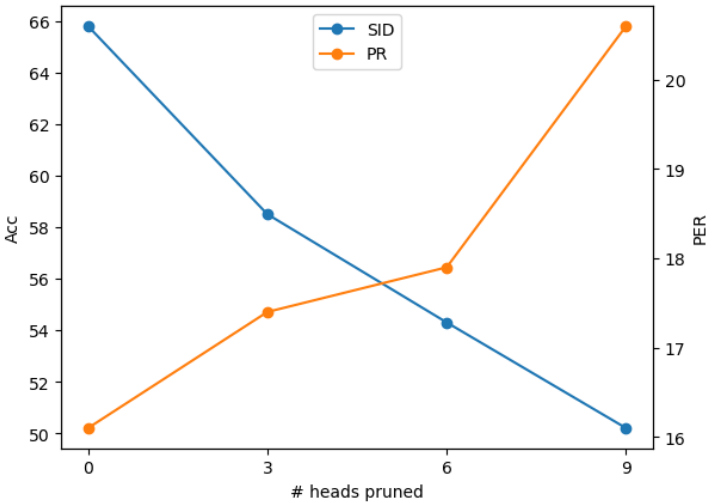
Pruning algorithm

- Prune heads based on ℓ_1
- Fine-tune

Training loss



Downstream tasks



Plan

- Before the workshop
  - MelHuBERT on 360 hours of LibriSpeech
  - Head pruning
  - Low-rank approximation
- During the workshop
  - Weight pruning
  - Scaling up to 960 hours of LibriSpeech
  - Distillation
  - Anytime inference

Scientific questions

- Is there a subnetwork that can match the self-superised loss?
- Does the subnetwork enjoy the same universality?
- Should we prune based on the downstream tasks?