Whisper Fine-Tuning

Whisper model

- Multilingual ASR model
- Transformer based
- Variants
  - Tiny
  - Base
  - Small
  - Medium
  - Large

Model	Layers	Width	Heads	Parameters
Tiny	4	384	6	39M
Base	6	512	8	74M
Small	12	768	12	244M
Medium	24	1024	16	769M
Large	32	1280	20	1550M

Dataset

- Czech Parliament Plenary Hearings - 444 hours
- Common Voice - 256 hours of crowdsourced Czech
- Preprocessing
  - Hugging Face datasets loading script
  - NeMo Punctuation and Capitalization model
    - Finetuned using UD_Czech-CAC and UD_Czech-PDT treebanks
  - Merging datasets

Training

- classical fine-tuning
  - base and small variant of the model
  - batch size = 64
  - learning rate = 1e-5, 10000 steps
- PEFT + LORA + INT8
  - large v2 variant of the model
  - rank=16, scale=16
  - learning rate = 1e-4, 5000 steps
  - batch size=64 vs. 4 without any on A40 48GB
- trained on Galdor cluster at MetaCentrum
  - NVIDIA A40 48GB

Results - Common Voice and Parliament

Model	Cross-entropy loss	WER
OpenAI Base	2.37	75.19
Common Voice and parliament Base (ours)	0.25	23.18
OpenAI Small	3.104	49.3
Common Voice and parliament Small (ours)	0.154	13.8
OpenAI LargeV2	2.39	34.2
Common Voice and parliament LargeV2-LORA-Int8 (ours)	0.17	15.53

Results - VoxPopuli

Model	Cross-entropy loss	WER
OpenAI Base	1.715	64.66
Common Voice and parliament Base (ours)	0.51	35.31
OpenAI Small	1.81	36.2
Common Voice and parliament Small (ours)	0.36	23.7
OpenAI LargeV2	1.32	17.8
Common Voice and parliament LargeV2-LORA-Int8 (ours)	1.21	16.9