Effects of Layer Freezing on Transferring a Speech Recognition System to Under-resourced Languages

Onno Eberhard and Torsten Zesch

onno.eberhard@stud.uni-due.de

LANGUAGE TECHNOLOGY LAB

UNIVERSITÄT DUISBURG ESSEN
Offen im Denken

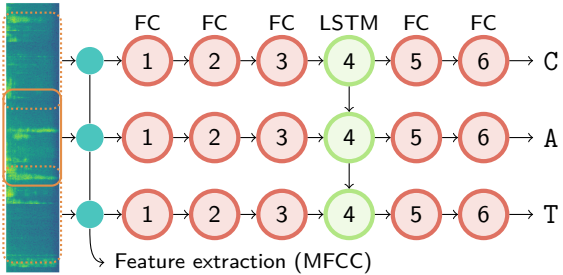
KONVENS · 8 September 2021

Introduction

- Most available automatic speech recognition (ASR) systems are built for English
- What about other languages?
  - Problem: Less available data
  - Large models will overfit
  - Small models usually have bad global optimum
- The solution: transfer learning
  - Train a large model on a related task with more available training data (for ASR: English language)
  - Freeze some parameters, thereby reducing model complexity
  - Not new in this context
- We want to understand how the number of frozen parameters changes the outcome.

ASR Architecture

- Mozilla DeepSpeech version 0.7
- Input features: Mel-frequency cepstral coefficients (MFCC)
- Output: Character probabilities over alphabet (using softmax)
- Loss: Connectionist temporal classification (CTC)
- Language model KenLM for better results



Transfer Learning

- Deep neural networks need lots of data to perform well
- If not enough data is available, transfer learning often helps
- Alternative to using a less flexible model (e.g. fewer layers)
- First optimize weights for a related "pretext" task (pre-training)
  - In Computer Vision: ImageNet or self-supervised
  - Here: Train on English language ASR dataset
- Fine tuning: Use these optimized weights as a starting point to learn the original task
- "Pure" tranfer learning does not prevent overfitting
- The model complexity still needs to be reduced
  - This is normally done by freezing some parameters of the model
  - In deep learning called "Layer Freezing" (freeze complete layers)

Which layers to freeze?

- Assumption: Features learned by early layers are more general than transformations learned by later layers.
  - Has been shown in computer vision, less evidence in ASR
  - We assume this to be true and freeze the earlier layers
- Deep neural network can be thought of as two parts:
  1. Feature extractor: First part, *general*
  2. Classifier: Second part, *specific to task*
- We want to freeze the feature extractor and learn the classifier
- No clear line between the two parts in a deep neural network
- Number of layers to freeze is not clear
  - We experiment with this number.

Datasets

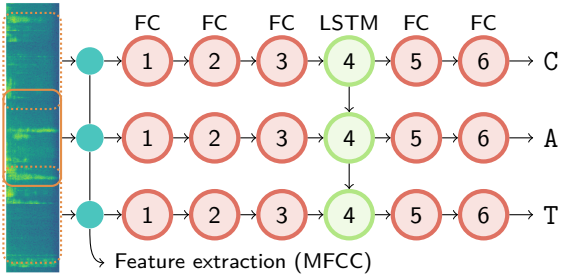
	Dataset	Hours	Speakers
Pre-training	English	>6,500	?
Transfer	German	315	4,823
	Swiss German	70	191

- English dataset for pre-training:
  - Trained by Mozilla on mix of many sources, incl. LibriSpeech and Common Voice English
- German dataset: Common Voice German
  - Uterrances 3 – 5 seconds; collected and reviewed by volunteers
- Swiss German dataset:
  - GermEval 2020 dataset: Parliament speeches; Standard German transcripts
- Language Model (Standard German): Wikipedia, Europarl, crawled sentences.

Training Details

- We do the following experiments:

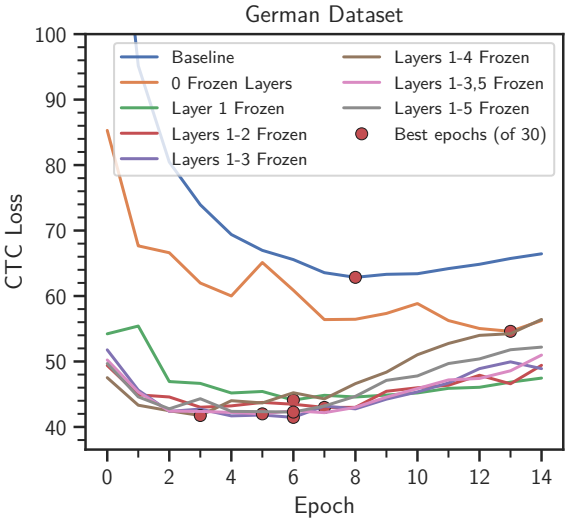
Baseline	Train the complete model from scratch, no transfer.
0 Frozen Layers	Weight initialization using pre-trained model, no layer freezing.
Layers 1-*N* Frozen	Weight initialization using pre-trained model, freeze first *N* layers.
Layers 1-3,5 Frozen	Freeze 5th instead of LSTM layer

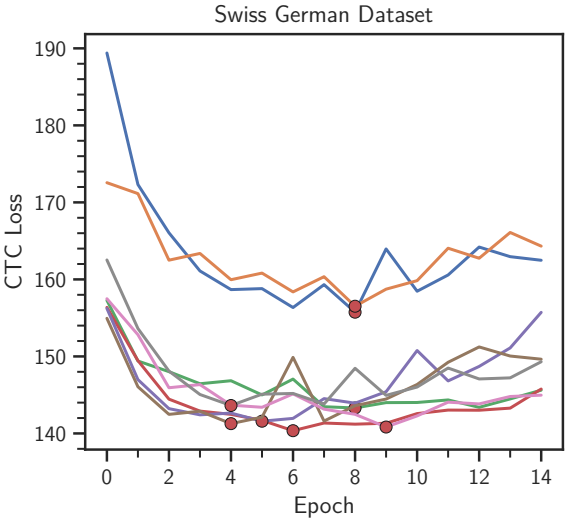


Results

	German		Swiss	
Method	WER	CER	WER	CER
Baseline	.70	.42	.74	.52
0 Frozen Layers	.63	.37	.76	.54
Layer 1 Frozen	.48	.26	.69	.48
Layers 1-2 Frozen	**.44**	**.22**	**.67**	**.45**
Layers 1-3 Frozen	**.44**	**.22**	.68	.47
Layers 1-4 Frozen	.45	.24	.68	.47
Layers 1-3,5 Frozen	.46	.25	.68	.46
Layers 1-5 Frozen	**.44**	.23	.70	.48

Learning Curves





Summary of results

- Very similar results for German and Swiss German
- Best results with $2 - 3$ layers frozen
- Number N of frozen layers seems unimportant, as long as $N \geq 1$
  - Large difference between freezing 0 or 1 layers
  - Even freezing all layers except the last (= linear classifier) yields good results
- No difference between freezing LSTM vs. a normal layer

Interpretation of results

- Seems to indicate the features learned by one language (English) are general enough to apply to other Languages
- *Even general enough to provide good features for a linear classifier*
- Takeaway: Don't worry about the number of frozen layers as a hyperparameter
- Open: Would it also work if we froze the last instead of the first layers? Or any random subset of parameters with the same degrees of freedom?

Limitations:

- We only looked at closely related languages
- Things might look very different when considering dissimilar languages:
  - Tonal languages like Mandarin or Thai
  - Languages heavily using phonemes non-existent in English
  - etc.
- The features learned by an English ASR system might not always be enough