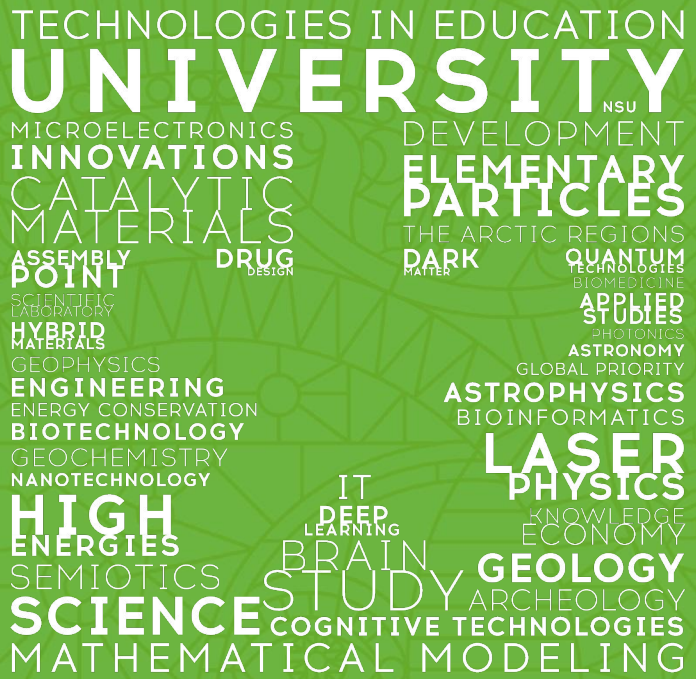


Novosibirsk State University
*THE REAL SCIENCE

Size Matters: About Optimal Amount of Speech Data for Student Hyperparameter Tuning in ASR Knowledge Distillation

Daniil Grebenkin and Ivan Bondarenko

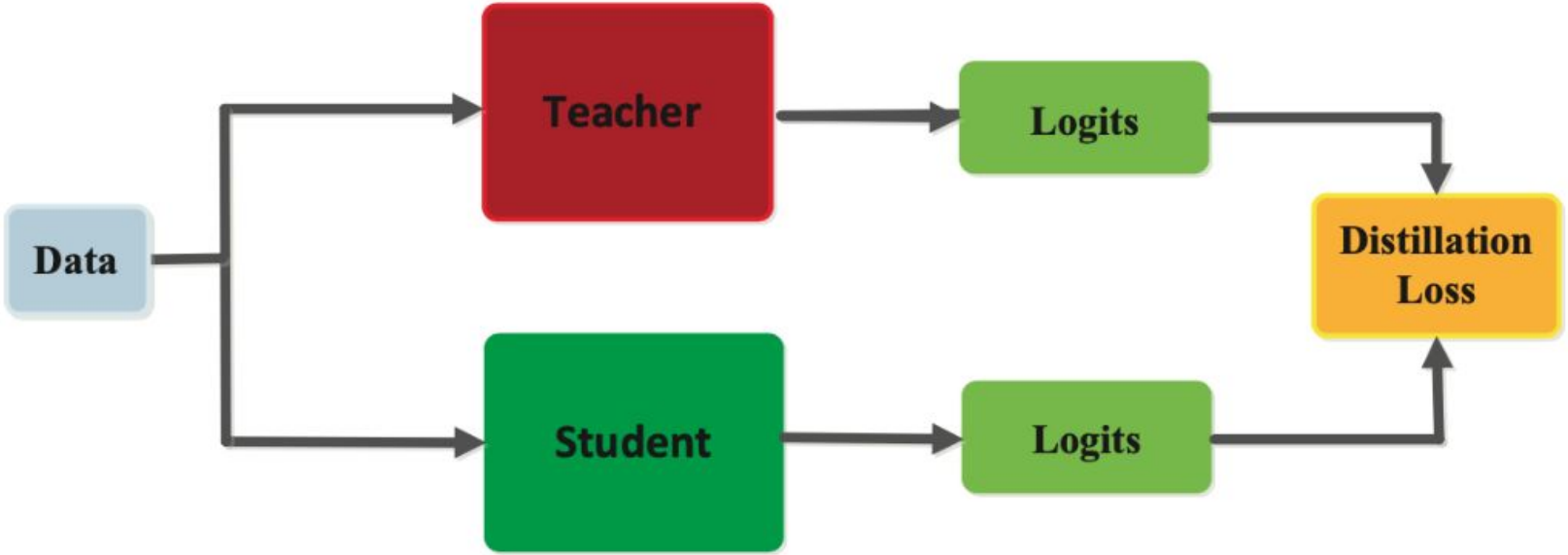
Laboratory of Applied Digital Technologies,
Novosibirsk State University
https://mca.nsu.ru/labadt/

AINL AutoML Workshop 2024

* End-to-end speech recognition systems optimization for “smart” devices

- End-to-end ASR systems as connectors between a human and the artificial intelligence located in the cloud🤝
- Achieving good speech recognition quality by using self-supervised learning algorithms and transfer learning📈
- Quantization (reducing the precision of the net's weights or activations) and **knowledge distillation (KD)**

Classic knowledge distillation approach



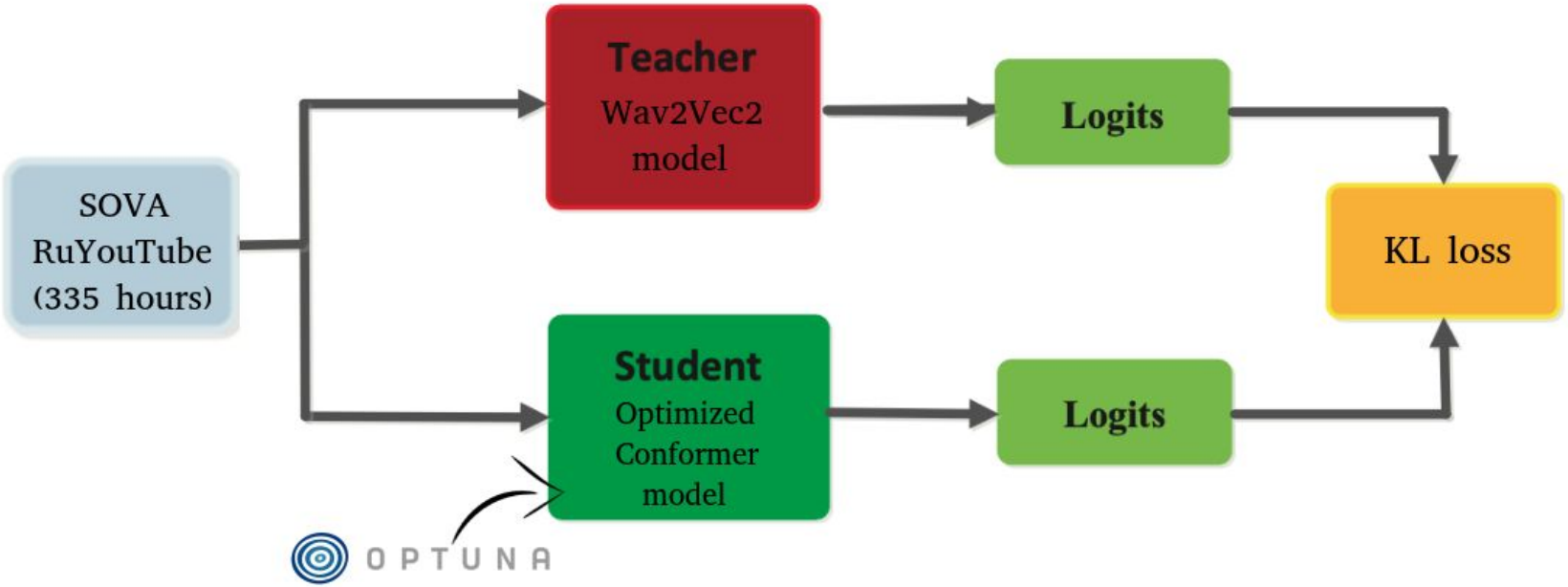
Gou, J., Yu, B., Maybank, S.J. *et al.* Knowledge Distillation: A Survey. *Int J Comput Vis* **129**, 1789–1819 (2021). https://doi.org/10.1007/s11263-021-01453-z

* Highlights of previous works dedicated to ASR models KD

- Kullback–Leibler divergence as distillation loss;
- N-stage knowledge distillation to reduce the number of teacher model parameters by five and more times (Conformer Transducer)
- Two-stage KD: get a smaller model in the first stage and to fine-tune it for downstream task on the second stage (DistilHuBERT, FitHuBERT, DistillW2V2 etc.)
- …

What student model we want to get? The best one!

* Solution: AutoML for Student parameters optimization



* Teacher model

Wav2Vec2 teacher model based on a pretrained multilingual model XLSR-53 and fine-tuned on Russian speech corpora in 2 stages:

1. Fine-tuning on Golos dataset by mini-batch sampling technique (which takes the complexity of each speech sample into account, based on its duration) to improve the generalization of the teacher model;

2. Fine-tuning on Golos, Russian LibriSpeech, and RuDevices to increase the model's ability to generalize in different domains.

Teacher model: https://huggingface.co/bond005/wav2vec2-large-ru-golos

* Conformer parameters optimization: step by step

1. Speaker clusterization of Golos dataset

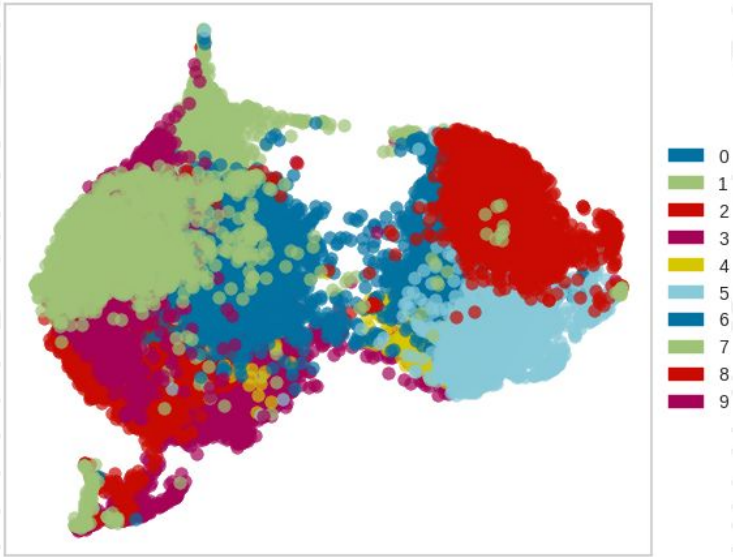
- Generation of audio embeddings by applying the **wav2vec2-base-superb-sv** model to audio data;
- Embeddings clustering by using different clustering algorithms to get speaker labels for every audio file;
- The original dataset splitting to training and testing sets according to obtained labels.

Embedding generator model: https://huggingface.co/anton-l/wav2vec2-base-superb-sv

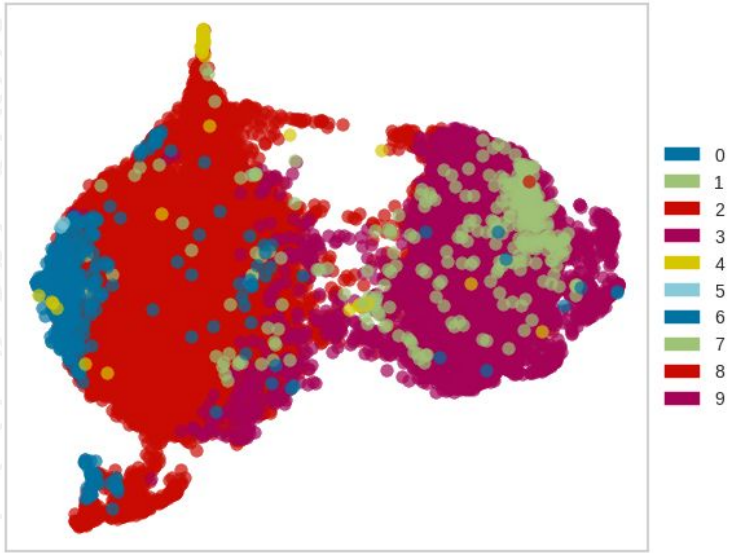
* Conformer parameters optimization: step by step

1. Speaker clusterization of Golos dataset

a)



b)



UMAP projections of clusterized speaker audio embeddings made by (a) K-Means algorithm, (b) Birch algorithm. The mean Silhouette Coefficient score of K-Means result is 0.066, the Birch result is 0.138.

* Conformer parameters optimization: step by step

2. Hyperparameter search with Optuna

- The hyperparameters were divided into 2 groups: **1)** Conformer student model architecture parameters, **2)** knowledge distillation learning hyperparameters.
- Optimization iteration:
  1.4 knowledge distillation training stages with 4 different random seed values on obtained training set of clusters;
  2.4 knowledge distillation testing stages with 4 different clusters of obtained testing set of clusters (Character Error Rate metrics);
  3.The estimation of geometric mean value of the CER measurements during the testing stages.

Goal: to find a well-balanced setup between the most suitable configuration of Conformer model for transferring knowledge and the training parameters to get the the lowest CER in the validation step.

* Conformer parameters optimization: step by step

2. Hyperparameter search with Optuna

Hyperparameters	Values
hidden_layer_size	512
num_layers	4
num_heads	8
dropout	0.497
ffn_dim	256
depthwise_conv_kernel_size	11
kernel_size	5
minibatch_size	4
learning_rate	0.000408

The values of optimized hyperparameters from the first optimization experiment. The Conformer model architecture parameters: **hidden layer size** – Conformer block's input dimension value, **num layers** – the number of Conformer blocks, **num heads** – the number of attention heads in each Conformer block, **kernel size** – the kernel size of 1-D convolutional layer, **depthwise conv kernel size** – the kernel size value of each Conformer block's depthwise convolution layer. The knowledge distillation learning hyperparameters: **minibatch size** – the number of samples per mini-batch, **learning rate** – learning rate value.

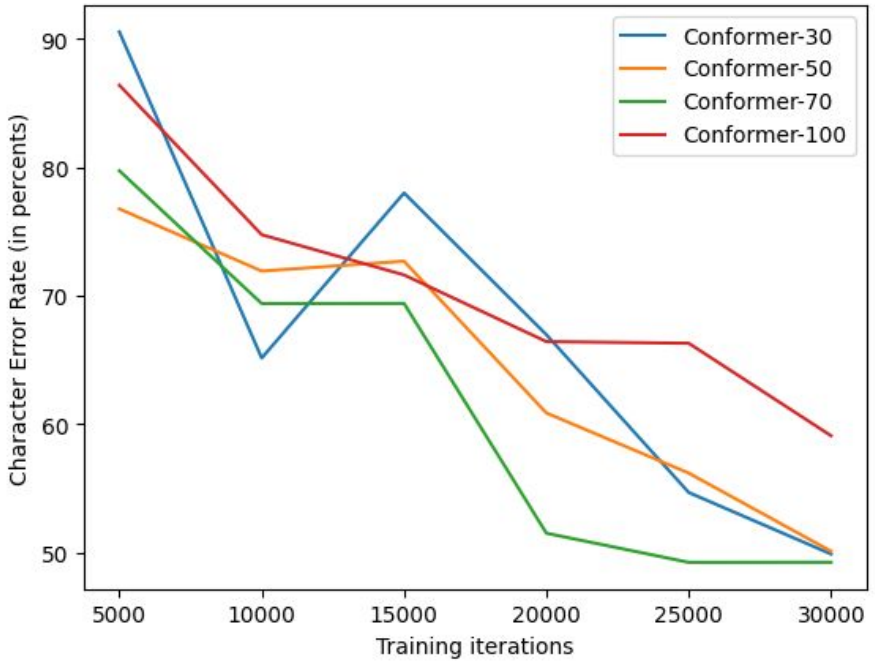
*Conformer parameters optimization: step by step

2. Hyperparameter search with Optuna

Hyperparameters	Used training dataset parts			
	100%	70%	50%	30%
hidden_layer_size	512	512	512	512
num_layers	4	6	6	6
num_heads	8	8	8	8
dropout	0.497	0.213	0.490	0.411
ffn_dim	256	512	64	64
depthwise_conv_kernel_size	11	11	5	9
kernel_size	5	3	5	4
minibatch_size	4	6	6	5
learning_rate	0.000408	0.000282	0.000220	0.000420

The values of optimized hyperparameters from all of the optimization experiments. The Conformer model architecture parameters: **hidden layer size** – Conformer block's input dimension value, **num layers** – the number of Conformer blocks, **num heads** – the number of attention heads in each Conformer block, **kernel size** – the kernel size of 1-D convolutional layer, **depthwise conv kernel size** – the kernel size value of each Conformer block's depthwise convolution layer. The knowledge distillation learning hyperparameters: **minibatch size** – the number of samples per mini-batch, **learning rate** – learning rate value.

Evaluation experiments



The development of knowledge distillation evaluation Character Error Rate on voxforge dataset

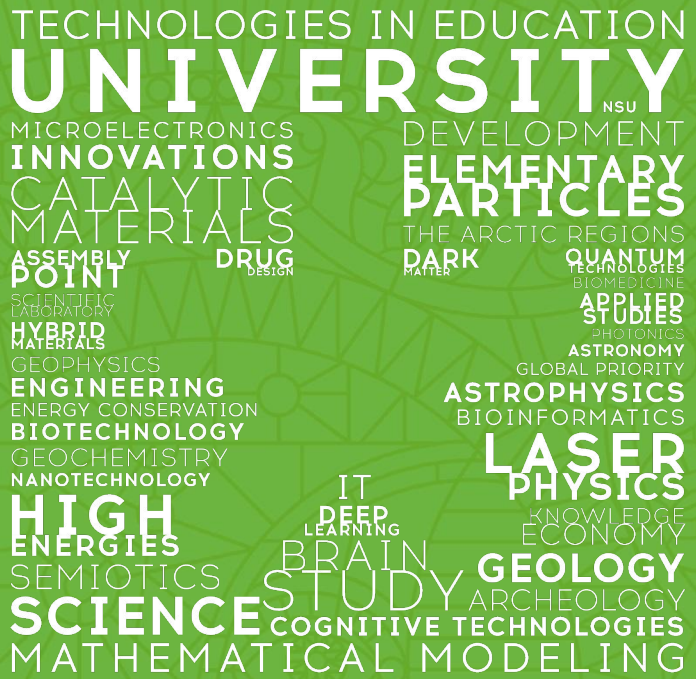
* Discussion

- The optimal hyperparameter search for the speech recognition model distillation can be carried out even on a small speech corpus ❗;
- The error rates are leveled after 25-30 thousand iterations🤔;
- Loss function affects the size of the speech corpus for preliminary hyperparameter search of the speech recognition model?

Conclusion

- We demonstrated that the optimal training configuration for conducting an efficient knowledge distillation of end-to-end teacher ASR model can be obtained with the different subsets of speech data.
- The tuning algorithm may be further improved, according to the desired model characteristics and its speech recognition quality metrics.
- The further work will investigate the efficiency of using the various parts of speech dataset and different loss functions for searching for the best parameters of SSL based ASR model on pretraining and fine-tuning stages.

Experiment code: https://github.com/dangrebenkin/w2v2_distillation_hyperparameter_auto_ml



Novosibirsk State University

*THE REAL SCIENCE

Thank you for your attention!

Contacts:

Daniil Grebenkin

d.grebenkin@g.nsu.ru

i.bondarenko@g.nsu.ru

Ivan Bondarenko