FAKULTA APLIKOVANÝCH VĚD
ZÁPADOČESKÉ UNIVERZITY
V PLZNI

STUDENTSKÁ
VĚDECKÁ KONFERENCE 2022

Transfer Learning of Transformers for Spoken Language Understanding

Adam Frémund[1], Martin Bulín, Jan Švec[2]

1 Introduction

Pre-trained models used in the transfer-learning scenario are recently becoming very popular. Such models benefit from the availability of large sets of unlabeled text data. In this paper, we proposed two kinds of transformer models for dialog systems. Specifically the Wav2Vec 2.0 and T5 text-to-text transformer models are used as speech recognizer and the spoken language understanding modules. The aim of this work was to outperform the baseline model based on the DNN-HMM speech recognizer and CNN understanding.

2 Transfer Learning for Spoken Dialog Systems

The Transformer architecture is a building block of many modern approaches in speech and language processing. The training process consists of two steps: (1) pre-training a generic model and (2) fine-tuning the pre-trained model on in-domain data.

Wav2Vec 2.0 Transformer (Baevski (2020)) is one of the most studied self-supervised end-to-end automatic speech recognition (ASR) model architectures. It is a deep neural network pre-trained to reconstruct the corrupted signals. Because there is no Wav2Vec 2.0 model available for Czech language, which we are experimenting with, we decided to pre-train our own model.

The Text-to-Text Transfer Transformer (T5) model is a self-supervised trained variant of the generic textual Transformer architecture (Raffel (2020)). The T5 model is able to construct the internal representation of input on many linguistic layers: starting from phonetic and syntactic through semantic to the pragmatic layer. The T5 model is pre-trained in a self-supervised manner by generating a text restoration task from unlabelled training data.

3 Results

In the experiments, we use two Czech semantically annotated corpora: a Human-Human Train Timetable (HHTT) corpus which contains inquiries and answers about train connections; and an Intelligent Telephone Assistant (TIA) corpus containing utterances about meeting planning, corporate resources sharing and conference call management. These corpora contain unaligned semantic trees together with word-level transcriptions.

The comparison of the recognition word accuracy of the speech recognizers is presented in Tab. 1. First, we present the performance of two different DNN-HMM recognizers on the HHTT and TIA datasets. Then, we recognized the same data using the W2V recognizer. Al-

[1] student of the doctoral study program Applied Sciences and Informatics, field of study Cybernetics, specialization Spoken dialog systems, e-mail: afremund@kky.zcu.cz

[2] University of West Bohemia, Faculty of Applied Sciences, Dept. of Cybernetics, email:bulinm@kky.zcu.cz, honzas@kky.zcu.cz

Table 1: Speech recognition word-level accuracy.

	% Acc	
	devel	test
DNN-HMM TIA	71.31	77.88
DNN-HMM HHTT	70.40	74.05
W2V TIA	83.70	86.08
W2V HHTT	68.93	73.66
W2V TIA normalized	86.39	89.14
W2V HHTT normalized	73.80	79.48
DNN-HMM TIA+HHTT	70.92	76.23
W2V TIA+HHTT	77.30	80.72
W2V TIA+HHTT normalized	80.96	85.01

Table 2: Spoken language understanding performance.

	% cAcc		% sAcc	
	devel	test	devel	test
DNN-HMM ASR + CNN SLU (baseline)	76.04	80.24	69.70	74.51
DNN-HMM ASR + T5 SLU	76.09	81.50	70.98	74.84
W2V ASR + T5 SLU	80.81	84.29	73.09	79.04
W2V ASR normalized + T5 SLU	81.19	85.37	73.55	79.33
ground truth transcription + T5 SLU	87.54	87.69	81.27	83.41

though the Wav2Vec recognizer does not use domain knowledge, we report the recognition accuracy for TIA and HHTT datasets separately. The comparison of W2V with the DNN-HMM shows that the W2V provides a significant performance boost on the TIA dataset but no improvement on the HHTT dataset. The error analysis on the HHTT dataset showed that a large number of errors come from the orthographic transcription produced by the W2V recognizer scored against a normalized ground-truth reference. Therefore, we applied the rule-based normalization, for example: *na shledanou ← nashledanou, naschledanou, na schledanou (**Lit.*** good bye*)*. Using these rules, we normalized the recognizer output as well as the ground truth transcription.

In the next set of experiments, we compared the CNN SLU baseline with the T5 SLU model. We have to note, that the CNN SLU baseline is a special model designed for the SLU task and is able to process the input in the form of word lattice and also generate the probabilistic distribution over the set of semantic trees. By contrast, the T5 SLU model is a text-to-text transformer working only with the 1-best input and 1-best output. From this point of view, the results shown in Tab. 2 are very promising – the much simpler fine-tuning and prediction of the T5 model is fully compensated by the knowledge extracted during self-supervised pre-training. For comparison we used Acc/cAcc metrics, higher values of these metrics means better results.

The results presented in this paper are very promising and outline future research in the applications of transfer-learning Transformers in spoken dialog systems.

Acknowledgement

This work was supported by the University of West Bohemia grant, project No. SGS-2022-017

References

Baevski, Alexei and Zhou, Yuhao and Mohamed, Abdelrahman and Auli, Michael. (2020) wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*. pp. 12449–12460

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li and Peter J. Liu (2020) Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*. pp. 1-67.