Review of the Doctoral Thesis

Exploring Benefits of Transfer Learning in Neural Machine Translation

Candidate: Tom Komi
Faculty of Mathematics and Physics, Charles University in Prague

The submitted PhD thesis "Exploring Benefits of Transfer Learning in Neural Machine Translation" focuses on highly topical theme in the field of Natural Language Processing (NLP) – the Neural Machine Translation (NMT). The use of deep neural networks became the primary approach in the machine translation just few years ago and currently it is the dominating methodology in the academic research as well as in the commercial machine translation services. The methods based on neural networks are in general highly data demanding that can cause problems for low-resource languages where the predecessor – phrase-based machine translation – can still outperform NMT in many cases.

The main goal of this thesis was to propose and evaluate transfer learning techniques to improve NMT especially for the low-resource languages, but it has appeared that these techniques can significantly improve the performance also for high-resource languages and that the transfer learning used as an initialization method in NMT does not have negative effects known in other NLP applications. Huge number of experiments has been performed on broad variation of languages and it has been shown that the quantity of parallel corpora plays more important role in the transfer learning than relatedness among the languages used for initialization of the network and the trained language pair. Furthermore, the transfer learning improves the performance even when no language is shared between the models used for the initialization and for the actual translation.

The first three chapters of the thesis give us an overview of language resources and terminology used in the thesis, introduction to neural machine translation, and description of evaluation metrics in NMT. Chapter 2 includes the description and evaluation of a language identification tool, called LanideNN, developed by the candidate to help with cleaning and filtering multilingual data. According to the author, this was the first attempt to use neural networks for language identification and it has reached the state-of-the-art in multilingual language identification at the time of publication, in 2017.

The fourth chapter presents and evaluates the proposed methods for transfer learning in the context in NMT. Two approaches are compared and studied: a cold-start scenario, where the "parent" model used for initialization of the network is not modified in advance, and a warm-start scenario where the "child" model data are available at the time of the parent model training, which can be thus adjusted for the transfer learning to better initialize the "child" model for the final translation of a given language pair. For the cold-start scenario, two approaches are studied: direct transfer and transformed vocabulary, for the warm-start scenario the balanced vocabulary approach is studied and evaluated. Both cold-start and warm-start transfer learning improved performance for both low-resource and high-resource language pairs, the warm-start method reached in most cases significantly better results in BLEU score.

Chapter 5 provides a detailed analysis of various phenomena in transfer learning; the observations are again supported by a broad range of experiments. Noteworthy is for example the evaluation of traces of parent language pair in the final translations (child model). The experiment shows that during the training of the child model, the neural net forgets the parent target language and adapts to the child target language, so an accidental appearance of words or segments from parent target language in the final translation is very rare. Next promising observation concerns the fact that transfer learning helps to train translation models for extremely low-resource language pairs that would not be possible to train at their own. Not so surprising but valuable observation is that the transfer learning can harm the child model performance when the parent language pair has substantially less training data. As already mentioned, very promising observation is that with enough parallel data for the parent model, the transfer learning technique improves the performance even for no-shared language scenario.

Another set of experiments in Chapter 5 concerns comparing the importance of languages relatedness vs. the data size. For these experiments the author decides to create artificially related language pairs by several levels of harmful modification of the source language. Another trick is to train the parent model on different language pairs (for example with shared target language). The observation says that more data in the parent models produce better results of the child model even when the parent model is trained on a mix of languages. Interesting fact is that the English -> Mix model (trained as parent) produces outputs in various languages but never mixes different languages in one sentence. Very promising observation is the use of a revers direction model as the parent model for transfer learning. I.e. for translation model from Estonian to English use English to Estonian as the parent model for initialization. With this approach the author observed better results for all the evaluated language pairs in both directions of translation, in more than half of the cases it resulted in significantly better BLEU score even for low-resource languages like Odia.

Subsequent case study applies the transfer learning to the standard backtranslation approach, i.e. generates parallel data by translating monolingual data using model trained on the initial set of parallel sentences. Related work shows that the standard approach applied potentially in multiple rounds improves the performance for low-resource languages. The candidate proposes to combine the backtranslation methodology with the transfer learning and shows really large improvement to the performance on two low-resource language pairs: Gujarati-English and Kazakh-English. This observation may lead to the tendency of using huge amount of the monolingual data for the backtranslation approach and then supplement this huge amount of synthetic data by oversampling the authentic data. One must be careful with this approach as for the low-resource languages it has be shown that oversampling authentic data hurts NMT performance, but still the improvement over using solely the authentic data is huge.

The final chapter summarizes the results, the author nicely highlights the observation continuously through the whole thesis and provides a separate list of all the observations in the appendix, which is a great idea. The thesis is concluded by the analysis of an ecological trace of all the experiments run by the author during the last 20 months. The total energy consumption and derived CO_2 emissions may look scary but in the case of this thesis it was definitively worth to spend it on. The whole thesis is well structured, nicely written and I believe that the observations are great contributions to the research in neural machine translation.

Two clarification questions:

1. What it is meant by the statement that shared-target language scenario is harder task compared to the shared-source language scenario (Section 5.2.4, page 88)?
2. I know that the thesis primarily focuses on the low-resource languages, the languages using logograms, such as Chinese, Japanese, and Korean were not included in the experiments. What is the candidate opinion on the potential of transfer learning in NMT applied to these languages?

In summary, this thesis represents an important work in the field of neural machine translation by uncovering a huge potential of using transfer learning to improve performance for both low-resource and high-resource languages. This is supported by the fact that the candidate co-authored 17 publications; he is the first author of 11 of them. According to Google Scholar, he is cited by 109 different publications. The candidate demonstrates his ability to work in research community and to provides replicable approaches that are highly relevant for the research community. I recommend this thesis for presentation with the aim of receiving the Ph.D. degree. This thesis fulfills all conditions of creative work and I recommend it to the defense.

RNDr. Jan Cuřín, Ph.D.

Prague, November 21, 2019