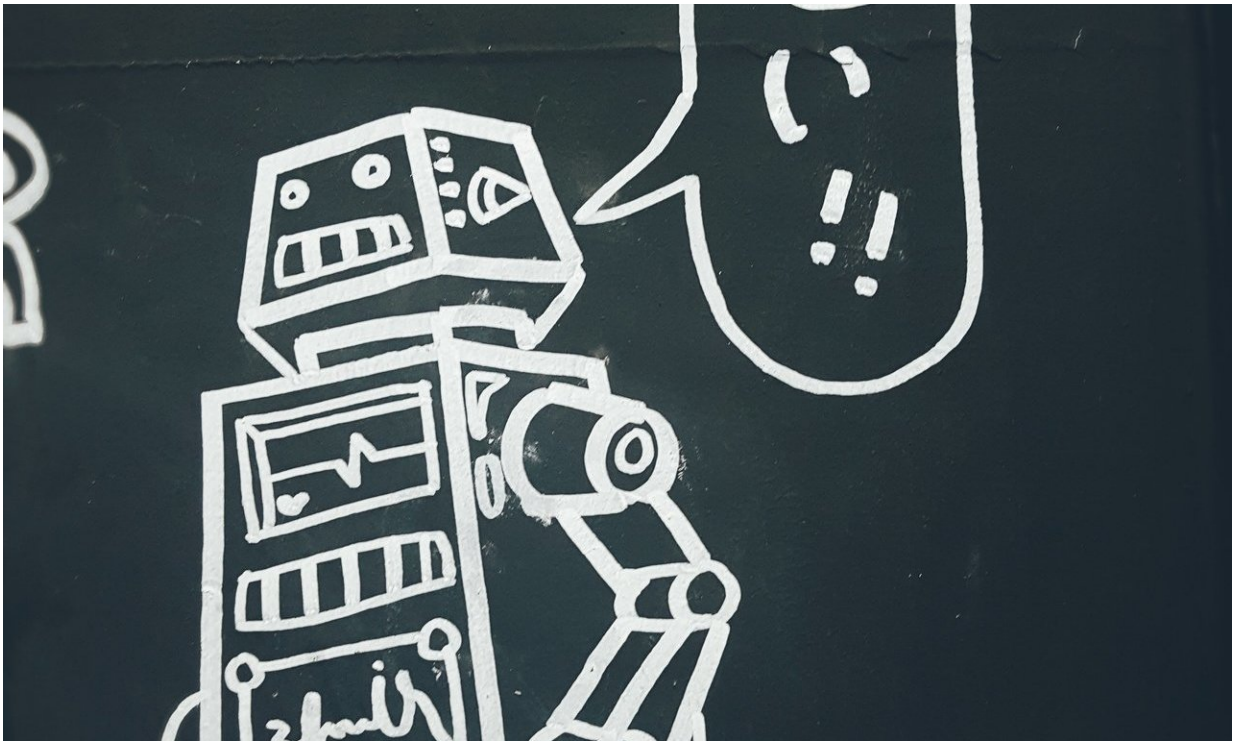
Using multi-task learning for low-latency speech translation

August 13 2018, by Ingrid Fadelli



Credit: Suan Moo, Unsplash.com

Researchers from the Karlsruhe Institute of Technology (KIT), in Germany, have recently applied multi-task machine learning to low-latency neural speech translation. Their study, which was pre-published on *ArXiv*, addresses some of the limitations of existing neural machine translation (NMT) techniques.

Advances in the field of deep learning have led to significant improvements in human speech and text translation. NMT, a widely used approach to machine translation, trains a large neural network to read a sentence and provide an accurate translation, generally by modeling entire sentences into an integrated model.

When compared to traditional approaches, such as rule-based or statistical machine translation, NMT typically achieves more fluent translations, both for speech and written text. While it can effectively capture more complex dependencies between source and target languages, to consistently perform well, this approach requires substantial amounts of training data.

"When applying partial sentence translation to neural machine translation systems, we encounter the problem that the MT system has only been trained on complete sentences, and thus the decoder is biased to generate complete target sentences," the researchers wrote in their paper. "When receiving inputs which are partial sentences, the translation outputs are not guaranteed to exactly match with the input content. We observe that the translation is often 'fantasized' by the model to be a full sentence, as would have occurred in the training data."

In other instances, the decoder can fall in an over-generation state, repeating the last word that was fed to it several times in its translation. To address these issues, the KIT researchers focused on speech translation in cases in which an NMT needs to provide an initial translation in real time, before a speaker has finished his/her sentence.

"In this work, we aim to remedy the problem of partial sentence translation in NMT," the researchers wrote. "Ideally, we want a model that is able to generate appropriate translations for incomplete sentences, without any compromise during other translation use cases."

As datasets with partial sentences are not readily available, the researchers created artificial data that could be used in the training process. They trained the network using multi-task learning, a deep learning strategy that has been often used in natural language processing (NLP) to train a single model for different tasks, reducing expenses and enhancing its performance.

Their study achieved promising results, suggesting that NMT systems could be adapted to perform well even in cases where task-specific data is not available, without losing performance on the original task they were trained for. "We first showed that simple techniques to generate artificial data are effective to get more fluent output with less correction," the researchers concluded in their paper. "We also illustrated that multi-task learning can help adapt the model to the new inference condition, without losing the original capability to translate full sentences."

Their adaptation of NMT achieved high-quality translations at low latency, minimizing the number of corrected words by 45 percent. In the future, their study could have meaningful practical implications, helping to develop better tools for real-time speech translation.

More information: Low-Latency Neural Speech Translation, arXiv: 1808.00491v1 [cs.CL]. arxiv.org/abs/1808.00491

Abstract
Through the development of neural machine translation, the quality of machine translation systems has been improved significantly. By exploiting advancements in deep learning, systems are now able to better approximate the complex mapping from source sentences to target sentences. But with this ability, new challenges also arise. An example is the translation of partial sentences in low-latency speech translation. Since the model has only seen complete sentences in training, it will

always try to generate a complete sentence, though the input may only be a partial sentence. We show that NMT systems can be adapted to scenarios where no task-specific training data is available. Furthermore, this is possible without losing performance on the original training data. We achieve this by creating artificial data and by using multi-task learning. After adaptation, we are able to reduce the number of corrections displayed during incremental output construction by 45%, without a decrease in translation quality.