Toward speech recognition for uncommon spoken languages

November 4 2021, by Adam Zewe

PARP is a new technique that reduces computational complexity of an advanced machine learning model so it can be applied to perform automated speech recognition for rare or uncommon languages, like Wolof, which is spoken by 5 million people in West Africa. Credit: Jose-Luis Olivares, MIT

Automated speech-recognition technology has become more common

with the popularity of virtual assistants like Siri, but many of these systems only perform well with the most widely spoken of the world's roughly 7,000 languages.

Because these systems largely don't exist for less common languages, the millions of people who speak them are cut off from many technologies that rely on speech, from smart home devices to assistive technologies and translation services.

Recent advances have enabled machine learning models that can learn the world's uncommon languages, which lack the large amount of transcribed speech needed to train algorithms. However, these solutions are often too complex and expensive to be applied widely.

Researchers at MIT and elsewhere have now tackled this problem by developing a simple technique that reduces the complexity of an advanced speech-learning model, enabling it to run more efficiently and achieve higher performance.

Their technique involves removing unnecessary parts of a common, but complex, speech recognition model and then making minor adjustments so it can recognize a specific language. Because only small tweaks are needed once the larger model is cut down to size, it is much less expensive and time-consuming to teach this model an uncommon language.

This work could help level the playing field and bring automatic speech-recognition systems to many areas of the world where they have yet to be deployed. The systems are important in some academic environments, where they can assist students who are blind or have low vision, and are also being used to improve efficiency in health care settings through medical transcription and in the legal field through court reporting. Automatic speech-recognition can also help users learn new languages

and improve their pronunciation skills. This technology could even be used to transcribe and document rare languages that are in danger of vanishing.

"This is an important problem to solve because we have amazing technology in natural language processing and speech recognition, but taking the research in this direction will help us scale the technology to many more underexplored languages in the world," says Cheng-I Jeff Lai, a Ph.D. student in MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL) and first author of the paper.

Lai wrote the paper with fellow MIT Ph.D. students Alexander H. Liu, Yi-Lun Liao, Sameer Khurana, and Yung-Sung Chuang; his advisor and senior author James Glass, senior research scientist and head of the Spoken Language Systems Group in CSAIL; MIT-IBM Watson AI Lab research scientists Yang Zhang, Shiyu Chang, and Kaizhi Qian; and David Cox, the IBM director of the MIT-IBM Watson AI Lab. The research will be presented at the Conference on Neural Information Processing Systems in December.

Learning speech from audio

The researchers studied a powerful neural network that has been pretrained to learn basic speech from raw audio, called Wave2vec 2.0.

A neural network is a series of algorithms that can learn to recognize patterns in data; modeled loosely off the human brain, neural networks are arranged into layers of interconnected nodes that process data inputs.

Wave2vec 2.0 is a self-supervised learning model, so it learns to recognize a spoken language after it is fed a large amount of unlabeled speech. The training process only requires a few minutes of transcribed speech. This opens the door for speech recognition of uncommon

languages that lack large amounts of transcribed speech, like Wolof, which is spoken by 5 million people in West Africa.

However, the neural network has about 300 million individual connections, so it requires a massive amount of computing power to train on a specific language.

The researchers set out to improve the efficiency of this network by pruning it. Just like a gardener cuts off superfluous branches, neural network pruning involves removing connections that aren't necessary for a specific task, in this case, learning a language. Lai and his collaborators wanted to see how the pruning process would affect this model's speech recognition performance.

After pruning the full neural network to create a smaller subnetwork, they trained the subnetwork with a small amount of labeled Spanish speech and then again with French speech, a process called finetuning.

"We would expect these two models to be very different because they are finetuned for different languages. But the surprising part is that if we prune these models, they will end up with highly similar pruning patterns. For French and Spanish, they have 97 percent overlap," Lai says.

They ran experiments using 10 languages, from Romance languages like Italian and Spanish to languages that have completely different alphabets, like Russian and Mandarin. The results were the same—the finetuned models all had a very large overlap.

A simple solution

Drawing on that unique finding, they developed a simple technique to improve the efficiency and boost the performance of the neural network,

called PARP (Prune, Adjust, and Re-Prune).

In the first step, a pretrained speech recognition neural network like Wave2vec 2.0 is pruned by removing unnecessary connections. Then in the second step, the resulting subnetwork is adjusted for a specific language, and then pruned again. During this second step, connections that had been removed are allowed to grow back if they are important for that particular language.

Because connections are allowed to grow back during the second step, the model only needs to be finetuned once, rather than over multiple iterations, which vastly reduces the amount of computing power required.

Testing the technique

The researchers put PARP to the test against other common pruning techniques and found that it outperformed them all for speech recognition. It was especially effective when there was only a very small amount of transcribed speech to train on.

They also showed that PARP can create one smaller subnetwork that can be finetuned for 10 languages at once, eliminating the need to prune separate subnetworks for each language, which could also reduce the expense and time required to train these models.

Moving forward, the researchers would like to apply PARP to text-to-speech models and also see how their technique could improve the efficiency of other deep learning networks.

"There are increasing needs to put large deep-learning models on edge devices. Having more efficient models allows these models to be squeezed onto more primitive systems, like cell phones. Speech

technology is very important for cell phones, for instance, but having a smaller model does not necessarily mean it is computing faster. We need additional technology to bring about faster computation, so there is still a long way to go," Zhang says.

Self-supervised learning (SSL) is changing the field of speech processing, so making SSL models smaller without degrading performance is a crucial research direction, says Hung-yi Lee, associate professor in the Department of Electrical Engineering and the Department of Computer Science and Information Engineering at National Taiwan University, who was not involved in this research.

"PARP trims the SSL models, and at the same time, surprisingly improves the recognition accuracy. Moreover, the paper shows there is a subnet in the SSL model, which is suitable for ASR tasks of many languages. This discovery will stimulate research on language/task agnostic network pruning. In other words, SSL models can be compressed while maintaining their performance on various tasks and languages," he says.

More information: Cheng-I Jeff Lai et al, PARP: Prune, Adjust and Re-Prune for Self-Supervised Speech Recognition. arXiv:2106.05933v2 [cs.CL], arxiv.org/abs/2106.05933

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology

https://techxplore.com/news/2021-11-speech-recognition-uncommon-spoken-languages.html