Microsoft Research Our research All Microsoft

Return to Blog Home

Microsoft Research Blog

Layer Trajectory BLSTM: New evolution enhances speech recognition technology

Published September 16, 2019

Research Area

- Artificial intelligence
- Human language technologies

Speech is a signal that can enable natural interaction between human and machine. In order to facilitate this exchange, machines have to be able to recognize what a human has spoken, both the words and the context in which those words appear. This is the task of speech recognition—a seemingly simple one from a human perspective but an incredibly difficult one for machines. For decades, researchers have been trying to develop algorithms to make speech recognition accuracy by machines close to what human beings can achieve. A big breakthrough happened only when deep feedforward neural networks (DNNs) were introduced to speech recognition, a pioneer work in which Microsoft was a key contributor.

DNNs, however, have limitations. They can’t identify the temporal relationship between speech frames from one moment in time to the next. As speech recognition technology has advanced, better models for identifying this temporal relationship have been created alongside the development of more accurate detection and classification of small units of speech, called senones.

Subscribe today ›

In the upcoming Interspeech 2019 paper, "Layer Trajectory BLSTM," Microsoft AI researchers Eric Sun, Jinyu Li, and Yifan Gong successfully advanced the speech recognition modeling technology by re-designing the modeling units for speech recognition. Specifically, this research improves on the current model of speech recognition technology, bi-directional LSTM (BLSTM), by adding layer trajectory to take over senone (target) classification so that the BLSTM can focus on temporal modeling.

Adding on new layers of speech recognition technology: Getting from RNN to ltBLSTM

As speech is a time-sequence signal, it is important to model the property of a speech signal over time. Therefore, recurrent neural network (RNN) models with long short-term memory (LSTM) units have been adopted and result in significant accuracy improvement. RNNs are able to identify the temporal relationship across speech frames, while LSTM units provide additional functions that better emulate human thinking when it comes to speech (for example, LSTM adds a "forget" gate that allows machines to better understand how to sift relevant and irrelevant speech frames it has encountered, along with input and output gates). LSTM is, however, limited by the fact that it can only read information in one direction.

Given the understanding that future information is beneficial to predicting a current speech frame's senone label, bi-directional LSTM (BLSTM) has been proposed to replace uni-directional LSTM, advancing speech recognition accuracy further. In other words, BLSTM can read inputs backwards and forwards, which enables it to use future context to recognize speech more accurately. Figure 1 shows the flowchart of BLSTM modeling, which uses T(ime)-BLSTM to indicate it works on temporal modeling across the time axis. Thanks to all of these technologies, speech recognition systems now perform similarly to humans.

Figure 1: BLSTM in traditional systems uses BLSTM units to complete both temporal modeling and target classification.

Moving beyond the current model utilizing a trajectory layer

Traditionally, the speech recognition model builds in a layer-by-layer, frame-by-frame fashion (a frame of speech typically spans several tenths of milliseconds with 10 milliseconds between frames). Note that BLSTM in Figure 1 not only does the temporal modeling job across the (horizontal) time axis but also the target classification job across the (vertical) depth axis. It is accomplishing two tasks at once, limiting its ability to focus solely on either of those tasks.

The researchers believe it is not optimal to work on both temporal modeling and target classification using the same modeling units. Although a speech recognition system using BLSTM units achieves very good recognition accuracy, it could be improved by decoupling the tasks of temporal modeling and target classification. To that end, in the paper the researchers have proposed a novel structure called layer trajectory BLSTM (ltBLSTM), which uses T(ime)-BLSTM units to focus on temporal modeling and D(epth)-LSTM units to take on the target classification job.

Figure 2: The layer trajectory BLSTM (ltBLSTM) model uses both T-BLSTM and D-LSTM units to allow both types of unit to specialize in individual tasks: T-BLSTM units focus on temporal modeling, while D-LSTM models focus on target classification.

As shown in Figure 2, now the speech recognition model has T-BLSTM working on temporal modeling across the time axis and D-LSTM working on target classification across the depth axis. The D-LSTMs at different time steps do not have any time dependency: they scan the hidden states of each T-BLSTM layer and use the summarized layer trajectory information for final target classification. Furthermore, the D-LSTMs create auxiliary connections for gradient flow, thereby making it easier to train deeper models. The decoupling makes the design of ltBLSTM very flexible—it is not necessary to use LSTM for depth processing. Any other model units suitable for classification can be used.

As a result, this new ltBLSTM model moves the speech recognition system one stage further. Training with a large amount of data, this new model can relatively improve the traditional BLSTM model by up to 14%. By using a low-latency design, the proposed ltBLSTM is now powering Microsoft's conversational meeting transcription and is expected to be widely used in speech recognition scenarios.

The speech recognition capability demonstrated by ltBLSTM works on the senone units, smaller units of speech when compared with sub-words or words. The researchers hope that this technology will lead to future developments that allow for sub-word and word units.

The researchers will present their research on ltBLSTM at Interspeech 2019. The presentation begins at 5:40pm on September 17, during the “ASR Neural Network Architectures 1” session.

Related to this article

Events

Microsoft at Interspeech 2019

Publications

Layer Trajectory LSTM

Layer Trajectory BLSTM

Up next

See all blog posts

Announcing the ORBIT dataset: Advancing real-world few-shot learning using teachable object recognition

Domain-specific language model pretraining for biomedical natural language processing

Making machines recognize and transcribe conversations in meetings using audio and video

FastSpeech: New text-to-speech model improves on speed, accuracy, and controllability

Follow us:

Share this page:

What's new

- Surface Pro 8
- Surface Laptop Studio
- Surface Pro X
- Surface Go 3
- Surface Duo 2
- Surface Pro 7+
- Windows 11 apps
- HoloLens 2

Microsoft Store

- Account profile
- Download Center
- Microsoft Store support
- Returns
- Order tracking
- Virtual workshops and training
- Microsoft Store Promise
- Flexible Payments

Education

- Microsoft in education
- Office for students
- Office 365 for schools
- Deals for students & parents
- Microsoft Azure in education

Enterprise

- Azure
- AppSource
- Automotive
- Government
- Healthcare
- Manufacturing
- Financial services
- Retail

Developer

- Microsoft Visual Studio
- Windows Dev Center
- Developer Center
- Microsoft developer program
- Channel 9
- Microsoft 365 Dev Center
- Microsoft 365 Developer Program
- Microsoft Garage

Company

- Careers
- About Microsoft
- Company news
- Privacy at Microsoft
- Investors
- Diversity and inclusion
- Accessibility
- Security

Sitemap Contact Microsoft Privacy Terms of use Trademarks Safety & eco About our ads