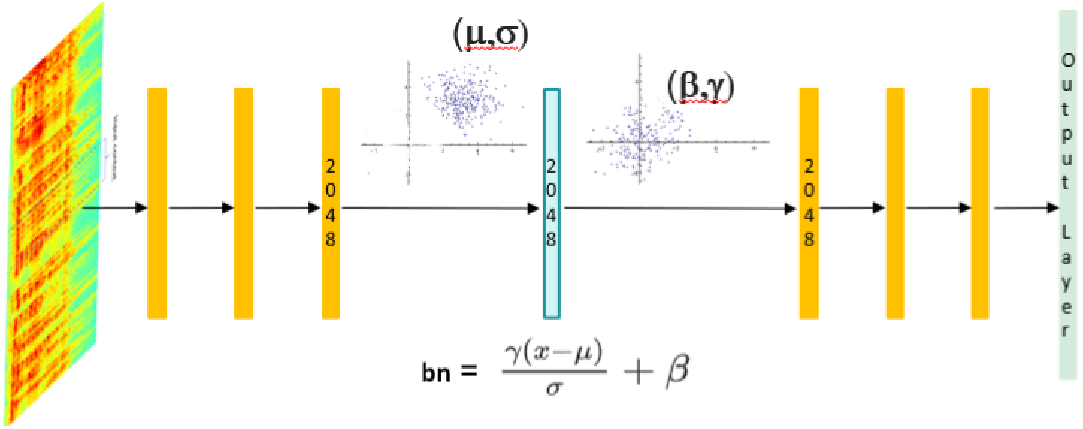
NUANCE | WHAT'S NEXT BLOG

Enterprise R&D, Innovation & Research

Making speech recognizers more robust in the wild

Innovation at Nuance | Employee Guest Blogger

January 13, 2020



Live adaptation capability of an ASR system is desirable but can be hard to achieve due to computational limitations, since it must operate while the service is up and running. The technique we propose fits both needs due to its simplicity and light requirement in terms of computational resources, while being applicable to any feed-forward architecture. In an Internet-of-Things scenario, we achieve word error rate reductions of 8.0% and 12.1%, respectively for two popular DNN architectures as Linear-Augmented DNN and Residual Net. The achieved result significantly improves over an i-vector baseline.

About the author: Franco Mana is a Principal Research Scientist in Nuance's Central Research department. Franco's interest in neural networks started in 1990, and since then, he has contributed to various developmental stages of neural network technologies as applied to speech recognition systems. These contributions include adopting graphic cards (GPU) for neural network acceleration and introducing parallel algorithms to make training large neural network models possible. He is a co-inventor of multiple patents ranging from neural network technology to signal processing. Prior to joining Nuance in 2011, Franco held a research position at a small technology company controlled by an Italian telephone company operating in speech recognition and synthesis systems. He holds a Bachelor of Science degree in computer science from Turin University in Italy.

Acknowledgements: This paper was co-authored by Franco Mana, Felix Weninger, Roberto Gemello, and Puming Zhan. It was presented at ASRU, the International Conference on Automatic Speech Recognition and Understanding, in January 2019.

Automatic Speech Recognition (ASR) facilitates humans interacting with machines in the most natural way: by speaking to them. Ideally, we would like to be able to talk with machines without limitation. This implies an ASR system should work well for any user and under any environment. Unfortunately, there are tremendous challenges towards achieving this goal. Today's state-of-the-art ASR systems are still fragile in the sense that their recognition accuracy may vary substantially under different operating environments. The root cause of this fragile performance is the mismatch between the data used to train the ASR models and the data they operate on in practice, despite attempts to train on data collected from a large number of users across various environments.

One of the most popular approaches to reduce the degradation of an ASR system in a new environment is to adapt the model with data collected in the targeted environment where the system operates. However, this usually requires collecting a certain amount of data in the targeted environment and adapting the model with the data beforehand. Obviously, such an approach is not convenient for users and is not amenable to applications in which environments can change dynamically during operation.

ASR performance has been improved significantly in recent years thanks to the application of Deep Neural Networks (DNN) technologies to acoustic and language modeling in ASR systems. However, these DNN-based systems still suffer from mismatched data in practice. Therefore, a great deal of research work has been done, including the aforementioned model adaptation, towards making DNN based ASR systems more robust in mismatched environment.

In the paper titled "*Online Batch Normalization Adaptation for Automatic Speech Recognition* (*ASRU 2019, Automatic Speech Recognition session III*), we envision a live adaptation (aka online adaptation) procedure which dynamically adjusts the ASR system. Live adaptation is appealing from the user point of view, because it operates without user-perceived interruption. It is also capable of capturing any dynamic change in the environment during the operation. However, it puts strong constraints on the underlying computational cost of the system in order to satisfy the latency requirement of the service. In this work, we present a simple and effective technique suitable for live adaptation to compensate train-test mismatch for acoustic models. The effectiveness of the method is measured in a domain mismatched scenario for two state-of-the-art DNN based acoustic models: Linear-Augmented DNN (LA-DNN) and Residual Networks (ResNet).

In DNN training, it is common practice to normalize input features to zero mean and unit variance to ease convergence of the training process. But after non-linear transformation layers, the output again becomes un-normalized, which makes training deep DNNs harder. The batch normalization (BN) technique enforces a normalization of the inputs to *each layer* in a DNN. This is achieved by computing the means and variances of each hidden layer outputs on a subset of the training data (called a mini-batch), then normalizing the outputs to zero mean and unit variance (again).

Our idea is to adapt the means and variances used for BN at test time in order to counteract a change in environment, speakers, etc. that could cause a domain mismatch and hence performance degradation. By doing this, we combine the benefits of BN (being able to train deep networks for better performance) and live adaptation (being able to counter domain mismatch).

All of this is more formally depicted in Figure 1. The BN layer shifts and scales each value of the inputs from the original distribution to a new one. It has some parameters, i.e. (m, s, b, g), which control the processing and are learned during the training stage. The input distribution of the BN layer, i.e. (m, s), is appealing for live adaptation, because it can be used to quantify the magnitude of the changes of the environment domain where the ASR system is operating. We can compensate these variations by re-estimating the (m, s) distribution dynamically during the operation of the ASR system. We call this enhancement "Online Batch Normalization" (OBN) technique. After the re-estimation of the input means and variances, the BN layer will realize the compensation automatically by shifting and scaling these updated values toward the pre-trained and fixed output (b, g) distribution.\

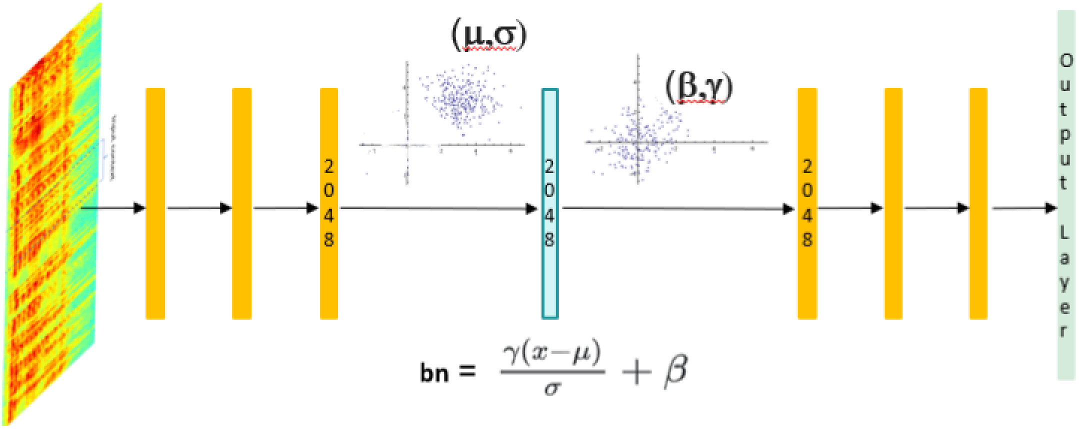


Figure 1 – The "batch normalization" layer can shift and scale the incoming activations from an input distribution toward an output one. Any new domain, that is reflected in the input distribution by a mis-matching condition with respect to the training, can be compensated by re-estimating the new input distribution during test. Then, the usual shift and scale processing can keep the output distribution stable.

The capability of the OBN layer to compensate for domain mismatched conditions is applicable to any feed-forward DNN model architecture. We demonstrated the generality of the method by applying OBN layer in LA-DNN and ResNet acoustic model. The models are trained by using anonymized field data that are representative for human-machine interactions and collected from devices with close-talk microphones in our various cloud services. The domain chosen for evaluation consists of far-field speech data collected in anonymized way from an IoT (Internet of Things) application, which presents a strong mismatch with the training data. In this scenario, we can reach a word error rate reduction of 8.0% and 12.1% for LA-DNN and ResNet, respectively. For comparison, we conducted the same experiments with i-vector based online adaptation and observed a word error rate reduction of 2.9% and 8.8% for LA-DNN and ResNet. This demonstrates the relative effectiveness of our technique versus a strong baseline for the domain mismatched scenario.

In conclusion, a live adaptation capability of an ASR system is desirable but can be hard to achieve due to computational limitations, since it must operate while the service is up and running. The proposed technique fits both needs due to its simplicity and light requirement in terms of computational resources, while being applicable to any feed-forward architecture. In an IoT scenario, for LA-DNN and ResNet DNNs we achieve word error rate reductions of 8.0% and 12.1%, respectively, significantly improving over an i-vector baseline.

Tags: Automatic speech recognition