EXTRACTING DEEP BOTTLENECK FEATURES USING STACKED AUTO-ENCODERS

Jonas Gehring¹  Yajie Miao²  Florian Metze²  Alex Waibel¹,²

1. Introduction

We propose a novel training scheme for generating bottleneck features from deep neural networks. A stack of denoising auto-encoders is first trained on standard audio features in a layer-wise, unsupervised manner. Afterwards, the bottleneck layer and an additional hidden layer are added and the whole network is fine-tuned to predict target phoneme states. Our architecture yields significant improvements on challenging tasks and benefits from adding more hidden layers as well as pre-training on additional unlabeled data.

Further training details:
- Input corruption with masking noise, setting a random 20% of the elements to zero
- Tied weights: encoding with $W$ and decoding with $W^T$
- Auto-encoder activation function $\sigma$ is standard sigmoid
- Mean squared error loss function $L$ for first layer, cross-entropy loss in subsequent layers
- Supervised fine-tuning with constant learning rate
- Network selection with early stopping using held-out validation data

2. A Deep Architecture for Bottleneck Feature Extraction

- Deep learning techniques are effective on many signal processing tasks, e.g. acoustic modeling
- How to leverage their modeling power for bottleneck feature extraction?

Proposed architecture on the left:
- Pre-train a stack of auto-encoders on standard input features
- Add randomly initialized bottleneck, additional hidden layer and classification layer on top
- Fine-tune the whole network to predict monophone target states
- Usage of bottleneck activations like in standard approaches

- Layers prior to the bottleneck are pre-trained as denoising auto-encoders that learn to reconstruct input vectors from randomly corrupted versions
- Formally, an input vector $x$ is first corrupted by a stochastic process $q_D$. By using a weight matrix $W$ and bias vectors $c$ and $b$, the hidden representation $y$ and reconstruction $z$ are computed from $x$ as follows:

\[
x \sim q_D(x) \quad y = \sigma(Wx + b) \quad z = \sigma(W^Ty + c)
\]

3. Evaluation

Evaluation was performed on recently released Cantonese and Tagalog corpora with varying sizes containing conversational telephone speech as used in the IARPA Babel Program.

- Deeper networks extract more useful features
- Pre-training of auto-encoder layers is required for good performance
- Unsupervised pre-training on more data helps slightly
- Supervision required to fully leverage the additional data
- Persistent improvements with BMMI/SAT training
- Similar results obtained on Switchboard (39.0% to 35.6%, network trained on 120 of 300 hours)

Supported in part by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense U.S. Army Research Laboratory (DoD/ARL) contract number W911NF-12-C-0015.