A Minimal Example of Multi-Task Learning Using The Librispeech and Tunisian MSA Corpora

John Morgan Michelle Vanni Stephen LaRocca

April 17, 2018

1 ABSTRACT

Multi-Task Learning was applied to a Large corpus of English and a small corpus of Modern Standard Arabic read speech. An improvement in Word Error Rate over the best Single-Task Learning method was observed.

2 INTRODUCTION

For the experiments described here, we used the Neural Network (nn) as a framework for acoustic modeling in Automatic Speech Recognition (asr). Recently, large asr systems have been trained on tens of thousands of hours of speech data [3]. nn models have performed well when these amounts of data are available. We assumed we are under conditions of severe training data sparcity in the Target Language (tl). We also assumed that we have access to a large corpus of speech in another language. We refer to this other language as the Background Language (bl). Multi-Task Learning (mtl) [2] is a framework that enables the advantages of Deep Learning (dl) to be applied in the situation where we have access to large amounts of data in the bl and scarce resources in the tl.

an nn was built with two kinds of layers.

1. Shared Layers.
2. Language Specific Layers.

The shared layers were trained on all the training data from all the languages.

The language specific layers were trained only on data from the tl.

Our research question was formulated as follows.

Can nn acoustic models trained with mtl on data from two different languages improve performance of an asr system for a Low Resource (lr) tl?

3 DATA

We ran an experiment to test the mtl method on two specific publically available corpora: the large Librispeech English corpus and a small corpus of Tunisian Accented Modern Standard Arabic (msa).

3.1 Librispeech

LibriSpeech is a corpus of read speech, based on LibriVox's public domain audio books. The corpus is available at:

`http://www.openslr.org/resources/12`.

We used the cleaned training fold of 960 hours of speech.

3.2 Tunisian msa:

This is a corpus of ten hours of msa as spoken by 120 male and female Tunisians in 2003. The informants provided recitations and answers to questions. It can be downloaded at:

`http://www.openslr.org/resources/46`.

4 EXPERIMENT

We used the kaldi toolkit [4] to build our ASR systems. We derived our setup from the kaldi babel multilang recipe. We tried to implement our setup to contain only methods that are required to run the mtl method. Thus, We did not use i-vectors which are standard in many kaldi recipes. We did include however a bottleneck layer.

We built baseline Speaker Adaptive Training (sat) Gaussian Mixture Model (gmm) Hidden Markov Model (hmm) Acoustic Model (am)s for both the Librispeech and Tunisian msa corpora. For the Tunisian msa baseline system we derived our pronouncing dictionary from the 2 million entries from the Qatar Computing Research Institute (qcri) vowelized dictionary [1] available at: `http://alt.qcri.org/resources/speech/dictionary/ar-ar_lexicon_2014-03-17.txt.bz2` We added the Out Of Vocabulary (oov) words from the Tunisian msa training set and the test set. We trained our 3-gram language model with the Stanfor Research Institute Language

Model (srilm) toolkit [5] on the transcripts from the training and test data. Our best Word Error Rate (wer) results for Tunisian msa were obtained with online chain models.

We followed the kaldi standard recipe for the librispeech cleaned 960 hours of speech task with one exception. We extracted and trained with Perceptual Linear Prediction $(\text{plp})_{\text{pitch}}$ features instead of MMel Frequency Cepstral Coefficients (mfcc) features. We followed this path since we derived our scripts from the kaldi babel multilang recipe.

4.1 Neural Network Configuration

The nn for both languages had 8 layers.

1. One input layer,
2. 6 hidden layers,
3. One Bottleneck layer,
4. One affine layer, and
5. One soft max layer.

The dimension of the hidden layers was 1024. The dimension of the Bottleneck layer was 512.

The soft max layer outputs a probability density function over the clustered triphones.

The frame Context was set to: 16 frames to the left, 12 frames to the right.

4.2 Training

A bilingual raw deep nn was trained on the combined set of training examples from the English Librispeech and Tunisian msa corpora. The data from the Tunisian msa corpus was used to readjust the parameters in the last two layers of the bilingual Deep Neural Network (dnn) model to produce a new monolingual Tunisian msa acoustic model. Similarly, a new monolingual English model was produced. These two models shared the parameters in their first six layers, only their final 2 layers were different.

4.3 Decoding

The monolingual system with the Tunisian msa am was used to decode a test set of speech from four speakers, 3 Libyan males and one Tunisian female. The same Finite State Transducer (fst) decoding graph that was built for the Tunisian msa sat gmm hmm system was used for decoding with the mtl am set.

5 RESULTS

TThe Single-Task unisian msa Baseline system with chain model trained ams yielded a wer of 11.03.

After MTL, the Tunisian msa system gave a wer of 7.12.

References

[1] A. Ali, Yifan Zhang, P. Cardinal, N. Dahak, S. Vogel, and J. Glass. A complete kaldi recipe for building arabic speech recognition systems. In *Spoken Language Technology Workshop (SLT), 2014 IEEE*, pages 525–529, Dec 2014.

[2] Richard Caruana. Multitask learning: A knowledge-based source of inductive bias. In *Proceedings of the Tenth International Conference on Machine Learning*, pages 41–48. Morgan Kaufmann, 1993.

[3] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean. Multilingual acoustic models using distributed deep neural networks," icassp, 2013.

[4] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Nagendra Goel, Mirko Hannemann, Yanmin Qian, Petr Schwarz, and Georg Stemmer. The kaldi speech recognition toolkit. In *In IEEE 2011 workshop*, 2011.

[5] Andreas Stolcke. Srilm - an extensible language modeling toolkit. pages 901–904, 2002.

Acronyms

am Acoustic Model. 3, 4

asr Automatic Speech Recognition. 1, 2

bl Background Language. 2

dl Deep Learning. 2

dnn Deep Neural Network. 4

gmm Gaussian Mixture Model. 3, 4

hmm Hidden Markov Model. 3, 4

lr Low Resource. 2

mfcc MMel Frequency Cepstral Coefficients. 3

msa Modern Standard Arabic. 1–4

mtl Multi-Task Learning. 2, 3

nn Neural Network. 1–4

oov Out Of Vocabulary. 3

plp Perceptual Linear Prediction. 3

qcri Qatar Computing Research Institute. 3

sat Speaker Adaptive Training. 3, 4

srilm Stanfor Research Institute Language Model. 3

tl Target Language. 2

wer Word Error Rate. 3, 4