Multi-Task Learning for Speech Recognition: An Overview

Gueorgui Pironkov, Stéphane Dupont, Thierry Dutoit

TCTS Lab, University of Mons, Belgium
{gueorgui.pironkov, stephane.dupont, thierry.dutoit}@umons.ac.be

Abstract. Generalization is a common issue for automatic speech recognition. A successful method used to improve recognition results consists of training a single system to solve multiple related tasks in parallel. This overview investigates which auxiliary tasks are helpful for speech recognition when multi-task learning is applied on a deep learning based acoustic model. The impact of multi-task learning on speech recognition related tasks, such as speaker adaptation, or robustness to noise, is also examined.

1 Introduction

Over the last few years, deep learning based acoustic models significantly outperformed Gaussian Mixture Models (GMM) for Automatic Speech Recognition (ASR) [1]. Specifically, Deep Neural Networks (DNN) discriminative learning offers a better fit for data modeling than the generative GMM [2]. This is due to the many levels of non-linearities of DNNs and their ability to assimilate higher levels of abstract concepts as the network deepness increases. Additionally, recent years advances in hardware and machine learning algorithms drastically improved DNN efficiency compared to GMM. In contrast with classic fully connected DNNs, more sophisticated methods take advantage of various specific hidden-connections architectures to further improve recognition accuracy. Among them can be cited Convolutional Neural Networks (CNN), where shared connection weights are applied to different localized patches [3]. Or Recurrent Neural Networks (RNN), which contain backward connections, thereby adding a temporal memory [4]. Despite this progress, overfitting tends to be a major issue for this deep learning algorithms. Indeed, with limited training data, the network learns good representation for the training set, which does not necessarily generalize well to test data, a problem commonly know as "overfitting". Several regularization methods try to improve generalization. For instance it is possible to stop training immediately when recognition drops on a validation set (early-stopping) [5]. Other methods investigate L1 and L2 regularization, adding a term to the cost function favoring sparse internal representation, which has shown to generalize better [6]. Lately, dropout has demonstrated promising results, by randomly dropping units during training, leading to a thinned neural network [7]. The main limitation with these methods is that the network ability to generalize is constrained by the recognition task. This leads to the intuition that overfitting can be reduced if the network is also asked to learn *meaningful* information, while estimating phoneme posterior probabilities for ASR. This scheme is referred to as Multi-Task Learning (MTL) [8]. The main idea is to train a single neural network to solve in parallel a main task, plus at least one auxiliary task.

In this paper, we will review the literature in MTL applied to ASR and see how MTL can increase the neural network generalization ability. This overview is organized as follows. In Section 2, MTL mechanisms will be described. The auxiliary tasks applied to ASR are presented in Section 3. Also, Section 3 gives a quick glance at MTL impact on other speech recognition related tasks. Finally we conclude and discuss the benefits of MTL for ASR in Section 4.

2 Multi-Task Learning

Multi-Task Learning was initially introduced in 1997 [8]. As suggested earlier, the idea is to train a neural network jointly for several different, but related tasks. Most often, the network learns one main task, and additionally to it one, two or more auxiliary tasks. The auxiliary tasks aim at helping the model to converge better, to the benefit of the main task. An illustration is presented in Figure 1, where the MTL has one main task and N auxiliary tasks. All MTL systems share two fundamental characteristics: a) all tasks are trained on the same input features, b) all tasks share the same internal representation. In order to update the network's parameters, the error will be backpropagated through the hidden layers of the network. In MTL, each task contributes to the cost function with a term:

$$\epsilon_{MTL} = \epsilon_{Main} + \sum_{n=1}^{N} \lambda_n * \epsilon_{Aux_n} \ ,$$

where ϵ_x is the cost function to be minimized, λ_n is a nonnegative weight and N the total number of auxiliary tasks. A λ_n closer to 1 means that the n^{th} auxiliary task will be as impacting as the main task, whereas a λ_n near 0 means that the auxiliary task has no influence on training. Usually, at test time, the auxiliary tasks are dropped, keeping only the neural network outputs for the main task. When the auxiliary tasks are well selected, MTL can help the model to improve its robustness to unseen data, thus, leading to better generalization. This method is especially efficient on limited datasets. Sharing the information among several tasks leads to higher performances compared to processing each task independently [8].

3 Auxiliary Tasks

Recently, MTL associated to DNN modeling has been successfully applied to several areas of speech and language processing, such as speech synthesis [9], speaker verification [10], multilingual speech recognition[1] [11, 12, 13], spoken

[1] MTL for Multilingual ASR is a particular case of MTL, as there is no main or auxiliary tasks, but all tasks/languages have the same impact. Additionally, for this scenario, several databases are used, which is not the case for classic ASR MTL.

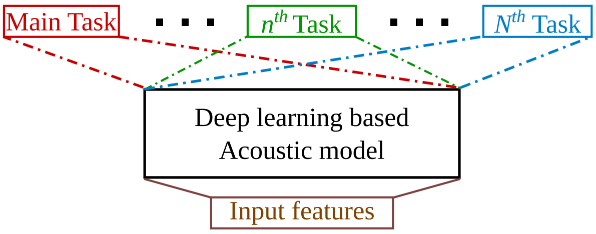

Fig. 1: A Multi-Task Learning network with one main task and N auxiliary tasks.

language understanding [14, 15], or natural language processing [16]. This study focuses on the use of MTL for the specific task of automatic speech recognition. In this section we will first investigate which auxiliary tasks are promising for ASR in general. We will also consider the helpful auxiliary tasks for other problems related to ASR, such as speaker adaptation, or ASR in noisy environment. A summary can be found in Table 1, where the relative improvement of MTL is compared to Single-Task Learning (STL), depending of the auxiliary tasks.

3.1 For Speech Recognition

3.1.1 Gender

Gender classification is one of the first auxiliary task that has been tested for ASR. Two (male/female) [17] or three (male/female/silence) [18] additional output nodes are added as auxiliary task after a RNN model. Even though this auxiliary task by itself does not seem very interesting, its association with another auxiliary task gives encouraging outcomes [17]. Another study also applies this auxiliary task after a RNN, and obtains its best Word Error Rate (WER) in comparison to the auxiliary task presented in the next subsection [18].

3.1.2 Phonetic units

Phonetic units are also considered as auxiliary task. A simple approach consists of using phoneme classification as an auxiliary task of a MTL DNN system [19], the purpose being to give the system indications about the similarity among acoustic states. Using even broader phonetic classes (such as plosive, fricative, nasal, . . .) is not efficient for MTL speech recognition [18].

3.1.3 Symbolic units

Instead of using phonetic-related representations, some studies focused on other representations of speech, such as graphemes. A grapheme is as symbolic representation of a phoneme, namely a character or group of characters from the alphabet that represent a sound. Estimating only graphemes as an auxiliary task degrades recognition accuracy [18].

3.1.4 Context

Another way to augment ASR generalization ability is to force the model to estimate local context. Giving a temporal context, by using as auxiliary task the next and previous frame's acoustic state, is an effective approach. Furthermore, estimating the left and right phoneme context is even more efficient [19]. Lastly, using graphemes can be helpful, if the context is added. For instance computing trigrapheme posterior probabilities as an auxiliary task improves recognition [20]. Besides using trigrapheme, Chen et al. also investigate two variants of the classic MTL DNN scheme. A first model (MTL_1) does not drop the auxiliary task at test time, but combines both tasks outputs using a weighted voting scheme, ROVER [21]. While a second model (MTL_2), uses a MTL with trigrapheme posterior probability estimation as a subtask that is fed to a second DNN. MTL_2 is depicted in Figure 2. Both MTL_1 and MTL_2 outperform the classic MTL approach, with MTL_1 showing slightly better accuracy results.

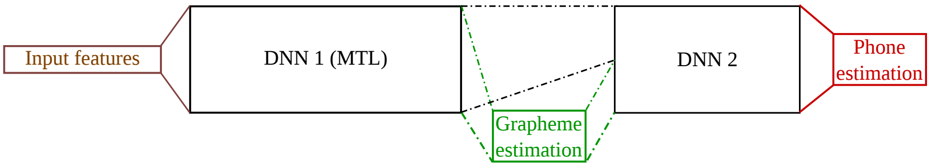

Fig. 2: MTL variant example. DNN 1 uses MTL to estimate trigrapheme posterior probabilities as an auxiliary task. The obtained estimates are then fed as complementary input to DNN 2 for classic ASR classification.

3.2 For ASR Speaker Adaptation

Speaker adaptation for ASR systems can be tricky, especially when the amount of adaptation data is limited. Huan et al. propose a MTL based adaptation scheme [22]: 1) A speaker-independent (SI) training is performed on a STL DNN for senone (tied-state triphones) classification using 7240 sentences. 2) The last weight matrix is expanded with connections corresponding to the auxiliary task: phoneme or senone-cluster estimation. 3) This weight matrix is fine-tuned using the 7240 sentences SI training data, while the rest of the DNN hidden parameters are fixed. 4) A Linear Hidden Network (LHN) [23] is inserted between the last hidden layer and the last weight matrix. The LHN is an identity matrix with zero bias. 5) Finally the LHN is updated by supervised adaptation using 1 to 40 sentences per speaker, while all the other parameters of the DNN are fixed.

Adapting only the LHN helps avoiding overfitting when the amount of adaptation data is limited, as the LHN number of parameters is way smaller than the DNN parameters. Results show that for a small number of adaptation sentences (1 or 2 sentences) using senone-cluster estimation as an auxiliary task gives better WER, whereas for more adaptation data (5 to 40 sentences) phoneme classification as auxiliary task performs better [22].

3.3 For Noise-Robust Speech Recognition

The degradations caused by noise and reverberation are a common problem for speech recognition. Learning complementary information about the acoustic environment can be fruitful for the speech recognition task. Thus, generating denoised speech, also referred to as speech enhancement (SE), is an effective auxiliary task for noise-robust ASR [17, 24]. Interestingly, using SE as an auxiliary task significantly improves the WER from lower Signal-to-Noise Ratio (SNR=-5dB) to relatively high SNR (SNR=25dB) [17].

Auxiliary task		**Database**	**Relative improvement**(%)	**Metric**
Gender [18]		WSJ0*	8.36	WER
Phonetic units	Broad phonetic classes (plosive, fricative, . . .) [18]	WSJ0*	-12.67	WER
	Phoneme [19]	TIMIT	0.46	PER
Symbolic units	Grapheme [18]	WSJ0*	-1.53	WER
Context	Frame [19]	TIMIT	3.00	PER
	Phoneme [19]	TIMIT	6.38	PER
	Grapheme [20]	Lwazi Speech Corpus*	4.80	ER
	MTL_1 Grapheme (see section 3.1.4) [20]	Lwazi Speech Corpus*	9.36	ER
	MTL_2 Grapheme (see section 3.1.4) [20]	Lwazi Speech Corpus*	8.59	ER
Speaker adaptation	Phoneme (1 adaptation sentence) [22]	WSJ0*	5.43	WER
	Senone-cluster (40 adaptation sentences) [22]	WSJ0*	10.75	WER
Speech enhancement [24]		CHiME 2	2.38	WER

Table 1: Relative MTL improvement comparing to STL for different auxiliary tasks. $<$ *Database name* $>$* implies that only a partial part of this database is used. *PER* stands for Phone Error Rate, and *ER* for Error Rate[2].

4 Conclusion and Discussion

In this paper, we gave an overview of multi-task learning applications for speech recognition, and more specifically the auxiliary tasks improving generalization. MTL is interesting technique as it increases generalization without requiring external data. Having one shared structure to update is also a positive aspect, leading to no major increase of the computational time, while improving the recognition accuracy. Nevertheless, using MTL also implies a preparation of the auxiliary task labels. Another problematic that occurs with MTL, consists of dealing with a possible temporality difference between the main task and the auxiliary task(s). For instance, gender recognition could require longer temporal information than phoneme recognition. Modeling features with memory-based deep neural networks such as the recurrent networks is then a promising solution.

5 Acknowledgements

This work has been partly funded by the European Regional Development Fund (ERDF) through the DigiSTORM project.

[2] ER = 100% - percentage of correctly recognized words, while WER depends of the insertions, deletions and substitutions. Thus, WER can be greater than 100%, contrary to ER.

References

[1] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE*, 29(6):82–97, 2012.

[2] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[3] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, and Gerald Penn. Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 4277–4280. IEEE, 2012.

[4] Oriol Vinyals, Suman V Ravuri, and Daniel Povey. Revisiting recurrent neural networks for robust ASR. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 4085–4088. IEEE, 2012.

[5] Lutz Prechelt. Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer, 1998.

[6] Steven J Nowlan and Geoffrey E Hinton. Simplifying neural networks by soft weight-sharing. *Neural computation*, 4(4):473–493, 1992.

[7] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[8] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[9] Zhizheng Wu, Cassia Valentini-Botinhao, Oliver Watts, and Simon King. Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis.

[10] Nanxin Chen, Yanmin Qian, and Kai Yu. Multi-task learning for text-dependent speaker verification. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[11] Stéphane Dupont, Christophe Ris, Olivier Deroo, and Sébastien Poitoux. Feature extraction and acoustic modeling: an approach for improved generalization across languages and accents. In *Automatic Speech Recognition and Understanding, 2005 IEEE Workshop on*, pages 29–34. IEEE, 2005.

[12] Georg Heigold, Vincent Vanhoucke, Alan Senior, Patrick Nguyen, Marc'Aurelio Ranzato, Matthieu Devin, and Jeffrey Dean. Multilingual acoustic models using distributed deep neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8619–8623. IEEE, 2013.

[13] Jui-Ting Huang, Jinyu Li, Dong Yu, Li Deng, and Yifan Gong. Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 7304–7308. IEEE, 2013.

[14] Gokhan Tur. Multitask learning for spoken language understanding. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 1, pages I–I. IEEE, 2006.

[15] Xiao Li, Ye-Yi Wang, and Gökhan Tür. Multi-task learning for spoken language understanding with shared slots. In *INTERSPEECH*, volume 20, page 1, 2011.

[16] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.

[17] Youyi Lu, Fei Lu, Siddharth Sehgal, Swati Gupta, Jingsheng Du, Chee Hong Tham, Phil Green, and Vincent Wan. Multitask learning in connectionist speech recognition. In *Proceedings of the Tenth Australian International Conference on Speech Science & Technology: 8-10 December 2004; Sydney*, pages 312–315, 2004.

[18] Jan Stadermann, Wolfram Koska, and Gerhard Rigoll. Multi-task learning strategies for a recurrent neural net in a hybrid tied-posteriors acoustic model. In *INTERSPEECH*, pages 2993–2996, 2005.

[19] Michael L Seltzer and Jasha Droppo. Multi-task learning in deep neural networks for improved phoneme recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 6965–6969. IEEE, 2013.

[20] Dongpeng Chen, Brian Mak, Cheung-Chi Leung, and Sunil Sivadas. Joint acoustic modeling of triphones and trigraphemes by multi-task learning deep neural networks for low-resource speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 5592–5596. IEEE, 2014.

[21] Jonathan G Fiscus. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*, pages 347–354. IEEE, 1997.

[22] Zhen Huang, Jinyu Li, Sabato Marco Siniscalchi, I-Fan Chen, Ji Wu, and Chin-Hui Lee. Rapid adaptation for deep neural networks through multi-task learning. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[23] Ryan Price, Ken-ichi Iso, and Koichi Shinoda. Speaker adaptation of deep neural networks using a hierarchy of output layers. In *Spoken Language Technology Workshop (SLT), 2014 IEEE*, pages 153–158. IEEE, 2014.

[24] Zhuo Chen, Shinji Watanabe, Hakan Erdogan, and John R Hershey. Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.