ACCOUNTING FOR THE RESIDUAL UNCERTAINTY OF MULTI-LAYER PERCEPTRON BASED FEATURES

Ramón Fernandez Astudillo[1], *Alberto Abad*[1,2] *and Isabel Trancoso*[1,2]

[1] L^2F - Spoken Language Systems Laboratory, INESC-ID Lisboa, Portugal
[2] IST - Instituto Superior Técnico, Universidade de Lisboa, Portugal
{ramon.astudillo, alberto.abad, isabel.trancoso}@l2f.inesc-id.pt

ABSTRACT

Multi-Layer Perceptrons (MLPs) are often interpreted as modeling a posterior distribution over classes given input features using the mean field approximation. This approximation is fast but neglects the residual uncertainty of inference at each layer, making inference less robust. In this paper we introduce a new approximation of MLP inference that takes under consideration this residual uncertainty. The proposed algorithm propagates not only the mean, but also the variance of inference through the network. At the current stage, the proposed method can not be used with soft-max layers. Therefore, we illustrate the benefits of this algorithm in a tandem scheme. We use the residual uncertainty of inference of MLP-based features to compensate a GMM-HMM backend with uncertainty decoding. Experiments on the Aurora4 corpus show consistent improvement of performance against conventional MLPs for all scenarios, in particular for clean speech and multi-style training.

Index Terms— Multi-Layer Perceptron, Mean Field Theory, Tandem, Uncertainty Decoding

1. INTRODUCTION

The use of Multi-layer perceptrons (MLP) for acoustic modeling [1] and feature extraction [2] is widely spread in automatic speech recognition (ASR). Recently, the interest in MLPs has spiked due to their central role in Deep Neural Networks (DNNs) [3, 4, 5]. In the context of ASR with DNNs, MLPs are often interpreted as a probabilistic model attained by stacking log-linear models [3, 5]. Under this interpretation, conventional MLP inference can be seen as approximate probabilistic inference using the mean field approximation. This approximation implies neglecting the uncertainty of inference at each hidden layer and passing only the average value to the next layer.

In this paper, we propose a closed form solution for inference in sigmoid layers named Gaussian Marginalization MLP (GM-MLP), in which this uncertainty is taken into consideration. The resulting inference algorithm is closer to the exact inference through marginalization. Furthermore, it provides a measure of inference uncertainty at the output of each layer that can be used for dynamic compensation.

At the current stage of development of the proposed method, no approximation has been found for the soft-max layers. Consequently, the method is here exemplified with a tandem scheme [2] in which GM-MLP-based extracted features are fed to a conventional Gaussian Mixture Model Hidden Markov Model (GMM-HMM) speech recognition system. The uncertainty of inference of the MLP is then used to dynamically compensate the GMM-HMM system using Uncertainty Decoding [6]. Results on the AURORA4 show that the presented approach consistently outperforms the conventional tandem approach. In particular, compensating the GMM-HMM for the uncertainty of inference of the GM-MLP produces notable improvements in the multi-style training scenario, which is usually not the case.

This paper is divided as follows. Section 2 reviews the mean field approximation for MLPs/DNNs. Section 3 introduces the Gaussian Marginalization MLP (GM-MLP) and discuses related works. Section 4 details the experimental setup and Section 5 presents the conclusions.

2. THE MEAN FIELD APPROXIMATION

This section is a review, see [7, 8, 5] for further details. In the context of ASR, a DNN/MLP of N layers models the posterior probability $p(q_l|\mathbf{x}_l)$ of each acoustic unit q_l e.g. monophones, senones, given a feature vector $\mathbf{x}_l$. This posterior can be seen as originating from the marginalization

$$p(q_l|\mathbf{x}_l) = \sum_{\mathbf{h}^N \in \mathbf{H}^N} \cdots \sum_{\mathbf{h}^1 \in \mathbf{H}^1} p(q_l, \mathbf{h}^N, \cdots, \mathbf{h}^1|\mathbf{x}_l) \qquad (1)$$

where $\mathbf{h}^N, \cdots, \mathbf{h}^1$ are binary vectors representing the hidden layer activations. The distribution in (1) factorizes as

$$p(q_l, \mathbf{h}^N, \cdots, \mathbf{h}^1|\mathbf{x}_l) = p(q_l|\mathbf{h}^N) \prod_{n=1}^{N} p(\mathbf{h}^n|\mathbf{h}^{n-1}). \qquad (2)$$

This work has been partially supported by the FCT through the grant number SFRH/BPD/68428/2010 and project PEst-OE/EEI/LA0021/2013, and by the DIRHA EU project FP7-ICT-2011-7-288121.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

where the activation of the j^{th} node of the n^{th} layer depends on all activations of the previous layer through

$$p(h_j^n|\mathbf{h}^{n-1}) = \frac{\exp\left(h_j^n z_j^n\right)}{\exp\left(0\right) + \exp\left(1 \cdot z_j^n\right)}, \tag{3}$$

with

$$z_j^n = \sum_{i=1}^{I^{(n-1)}} w_{ij}^n h_i^{n-1} + b_j^n. \tag{4}$$

Here $n \in \{1 \cdots N\}$ and the convention $\mathbf{h}^0 = \mathbf{x}_l$ has been used for simplicity.

Due to the factorization in (2), the marginalization can be carried out layer by layer in a step-wise fashion as e.g.

$$p(h_j^n|\mathbf{x}_l) = \sum_{\mathbf{h}^{n-1} \in \mathbf{H}^{n-1}} p(h_j^n|\mathbf{h}^{n-1}) p(\mathbf{h}^{n-1}|\mathbf{x}_l) \tag{5}$$

Unfortunately, this is computationally unfeasible for a reasonable layer size $I(n-1)$ since $|\mathbf{H}^{n-1}| = 2^{I(n-1)}$. For this reason, the so called mean field approximation[1] is used. This approximation assumes that the sum of hidden binary random variables in (4) collapses into the expected value of its sum. In other words, we have

$$p(z_j^n|\mathbf{x}_l) \approx \delta(z_j^n - E\{z_j^n|\mathbf{x}_l\}). \tag{6}$$

where $\delta()$ is the Dirac delta.

This approximation greatly simplifies the marginalization at the cost of neglecting the uncertainty in z_j^n. We need no longer to compute the whole posterior $p(\mathbf{h}^n|\mathbf{x}_l)$, it suffices to compute its expected value, which using (3) and (6) yields

$$\begin{aligned} E\{h_j^n|\mathbf{x}_l\} &= \sum_{\mathbf{h}^{n-1} \in \mathbf{H}^{n-1}} p(h_j^n = 1|\mathbf{h}^{n-1}) p(\mathbf{h}^{n-1}|\mathbf{x}_l) \\ &\approx \int_{-\infty}^{\infty} p(h_j^n = 1|z_j^n) p(z_j^n|\mathbf{x}_l) dz_j^n \\ &= \int_{-\infty}^{\infty} p(h_j^n = 1|z_j^n) \delta(z_j^n - E\{z_j^n|\mathbf{x}_l\}) dz_j^n \\ &= \frac{1}{1 + \exp\left(-E\{z_j^n|\mathbf{x}_l\}\right)}. \end{aligned} \tag{7}$$

Furthermore, due to the linearity of the expectation operator, the mean value of z_j^n can be directly computed from (4) as

$$E\{z_j^n|\mathbf{x}_l\} = \sum_{i=1}^{I^{(n-1)}} w_{ij}^n E\{h_i^{n-1}|\mathbf{x}_l\} + b_j^n. \tag{8}$$

This leads to the well known formulas for inference in MLPs/DNNs commonly referred as forward-pass. The same approximation can be applied to derive the inference at the last layer of the network $p(q_l|\mathbf{h}^N)$, which uses a soft-max non-linearity to yield a categorical posterior distribution over the acoustic units q_l.

MLPs/DNNs are usually trained with Backpropagation, which uses stochastic gradient and the cross-entropy (CE) criterion. This can also be seen as maximizing the log posterior probability under the mean field approximation

$$\mathcal{F}^{\text{CE}} = \sum_{r=1}^{R} \sum_{l=1}^{T_r} \log p(q_l|\mathbf{x}_l^r) \tag{9}$$

over all frames T_r of each train utterance r, see [7, 5].

3. BEYOND THE MEAN FIELD APPROXIMATION

3.1. The Gaussian Marginalization Approximation

The mean field approximation previously described only demands for the computation of the mean at each step and thus it is a very fast approximation. However, it completely neglects the uncertainty at each hidden node given the input $p(z_j^n|\mathbf{x}_l)$. This means that each layer is unaware of the error of the previous layer, making inference less robust.

Consequently, it would be desirable to consider this uncertainty of inference, while simultaneously keeping the low computational cost of the mean field approximation.

In this paper, we present a method that provides both characteristics. The basic idea underlying this principle is to approximate the large sum of hidden variables at each layer (4) by a Gaussian distribution

$$p(z_j^n|\mathbf{x}_l) \approx \mathcal{N}\left(\mu_j^n, \Sigma_j^n\right), \tag{10}$$

rather than the Dirac delta in (6). The parameters of the Gaussian are then

$$\mu_j^n = E\{z_j^n|\mathbf{x}_l\}, \qquad \Sigma_j^n = \text{Var}\{z_j^n|\mathbf{x}_l\}. \tag{11}$$

The Gaussian assumption is justified by the fact that node outputs of a MLP have a weak statistical dependence [9] and the central limit theorem, since z_j^n is a very large sum of random variables (4).

With this model, inference follows the same procedure as the mean field approximation in (7), only that second order information is also needed. The mean activations can be obtained by solving

$$\begin{aligned} E\{h_j^n|\mathbf{x}_l\} &\approx \int_{-\infty}^{\infty} p(h_j^n = 1|z_j^n) \mathcal{N}\left(\mu_j^n, \Sigma_j^n\right) dz_j^n \\ &= \int_{-\infty}^{\infty} \frac{1}{1 + \exp\left(-z_j^n\right)} \mathcal{N}\left(\mu_j^n, \Sigma_j^n\right) dz_j^n. \end{aligned} \tag{12}$$

This integral is here approximated using the PIecewise Exponential (PIE) Sigmoid approximation [10]. This solution approximates the sigmoid function by

$$\frac{1}{1 + e^{-z_j^n}} \approx 2^{z_j^n - 1} u(-z_j^n) + (1 - 2^{-z_j^n - 1}) u(z_j^n) \tag{13}$$

[1] The term mean field theory is also used to encompass more complex variational approaches to inference, see [8].

where $u(x)$ is the unit step function. For this approximation a closed form solution for (12) exists and is given by [10, Eq. 13]

$$\begin{aligned} E\{h_j^n|\mathbf{x}_l\} \approx\ & 2^{\left(\mu_j^n + \frac{1}{2}\log(2)\Sigma_j^n - 1\right)} \\ & \cdot \Phi\left(-\frac{\mu_j^n}{\sqrt{\Sigma_j^n}} - \log(2)\sqrt{\Sigma_j^n}\right) \\ & - 2^{\left(-\mu_j^n + \frac{1}{2}\log(2)\Sigma_j^n - 1\right)} \\ & \cdot \Phi\left(\frac{\mu_j^n}{\sqrt{\Sigma_j^n}} - \log(2)\sqrt{\Sigma_j^n}\right) \\ & + \Phi\left(\frac{\mu_j^n}{\sqrt{\Sigma_j^n}}\right) \end{aligned} \quad (14)$$

where $\Phi()$ is the Cumulative Density Function (CDF) of a normal variable.

Once the mean activations are computed, the second order information can be computed by taking into account that

$$E\{(h_j^n)^2|\mathbf{x}_l\} = 1^2 \cdot p(h_j^1 = 1|\mathbf{x}_l) + 0 = E\{h_j^n|\mathbf{x}_l\} \quad (15)$$

thus, the variance can be obtained as

$$\begin{aligned} \text{Var}\{h_j^n|\mathbf{x}_l\} &= E\{(h_j^n)^2|\mathbf{x}_l\} - E\{h_j^n|\mathbf{x}_l\}^2 \\ &= (1 - E\{h_j^n|\mathbf{x}_l\})E\{h_j^n|\mathbf{x}_l\}. \end{aligned} \quad (16)$$

Due to the properties of the variance of a random variable we have that

$$\Sigma_j^n = \text{Var}\{z_j^n|\mathbf{x}_l\} = \sum_{i=1}^{I^{(n-1)}} \left(w_{ij}^n\right)^2 \text{Var}\{h_i^{n-1}|\mathbf{x}_l\}, \quad (17)$$

with which we can completely characterize the distribution (10) and solve the marginalization one layer at a time. The new inference method for MLPs proposed is hereinafter referred to as Gaussian Marginalization (GM)-MLP.

3.2. Computational Costs and Current Limitations

The computational cost of the approach presented here is around twice that of a conventional MLP. For the usual network sizes used in ASR, the cost of inference is dominated by the linear step (4) and thus replacing the sigmoid computation by (14) and (16) has a small impact. Since the linear transformation has to be now applied to both mean (8) and variance (17), the cost approximately doubles.

The main limitation of the proposed approach is that no solution for inference with a soft-max layer has yet been found. Consequently, the GM-MLP can not be currently used as acoustic model. This limits its use to feature extraction on a GMM-HMM ASR system as e.g [2]. Interestingly, since the GM-MLP also produces a variance of estimation of the features, this variance can be used to compensate the GMM-HMM system by using observation uncertainty techniques [11] and thus attain additional robustness.

Finally, another current limitation of the GM-MLP approach is the lack of a training method under the exact same approximation. As discussed in Section 2, Backpropagation can be seen as maximizing the log posterior probability under the mean field approximation. Since the GM-MLP is a different approximation for the same posterior, it can be expected that the conventional training process remains valid to some extent when the new GM-MLP approximation is used in test. In practice, this creates a mismatch which probably hinders the performance of the method, as commented in the experimental section.

3.3. Comparison with Related Works

The solution here proposed can be related to works which approximate (12) for other purposes e.g. non-linear ICA [12] and uncertainty propagation through the MLP non-linearity [10]. These works use similar approximations, but solve a different problem since they ignore the probabilistic view of the MLP. In particular [10] uses the PIE approximation to propagate uncertainty coming from the front-end, while considering the MLP as a deterministic function. This requires solving different problems i.e. computing second order moments with respect to the sigmoid non-linearity. Finally, with regard to inference approximations, variational approximations for belief networks under the mean field theory [8] can also be related to the work presented here. However, their iterative nature makes them ill suited for ASR.

4. EXPERIMENTS AND RESULTS

4.1. Experimental Setup

The presented method is here tested in a tandem scheme [2] on the AURORA4 corpus [13], an artificially corrupted version of the Wall Street Journal corpus of 5k words. Two major steps can be differentiated in the training process.

In a first step, similar to that of [3], we train a hybrid MLP-HMM system on the alignments attained from a conventional GMM-HMM system. For this purpose, we use MMSE-Mel-Frequency Cepstral Coefficients (MFCC) features [14] with cepstral mean subtraction per utterance[2] and HTK [15]. Training followed Vertannen's HTK recipe for a word internal triphone-based ASR system[3], we denote this system as MMSE-MFCC. As hybrid MLP-HMM ASR system, we use our in-house system AUDIMUS. This is a classic three layer MLP using 321 acoustic units, including state-dependent monophones and phoneme transition units [16].

[2] Note that, unlike in [14], the MMSE-MFCC variances are not used. The input to the network is deterministic.

[3] http://www.inference.phy.cam.ac.uk/kv227/htk/

Two variants of MLPs were used: clean training and multi-style training. The latter included various noisy utterances and recordings with a different microphone in the train-set.

In a second step, we construct a feature extraction by using the trained MLP output without the soft-max layer, applying Principal Component Analysis (PCA) to attain dimensionality reduction from 321 to 39 features and applying an additional mean subtraction per utterance. These features are then used to train a new GMM-HMM system on HTK, which is labeled MLP.

MMSE-MFCC and MLP systems are here considered as baselines. These are compared with a GM-MLP variant attained by replacing the MLP by the GM-MLP on the second stage of training. It is important to underline that the GM-MLP approximation was not used during the training of the MLP model in the first step, due to the restrictions on the use of soft-max discussed in Section 3.2. Therefore, the GM-MLP system is not optimally trained. Two variants are tested. The GM-MLP variant uses only the feature means at the last layer. A second variant, GM-MLP+UD, uses also the variance of inference of the GM-MLP features with Uncertainty Decoding (UD) [6]. This technique only implies adding the variance obtained from the GM-MLP features to the variance of each GMM mixture. Note that, since both PCA and mean normalization are linear operations, obtaining the variances of the GM-MLP features from the variances at the last layer (17) of the GM-MLP, is trivial.

The test set is based on the November 1992 ARPA WSJ evaluation set, but includes six additional versions with different types of noise. Although the algorithm here presented is not specifically designed for robustness against environment distortions, the results on noisy speech are also provided for completeness. Results for all noises are averaged into one single coefficient but they are considered independently for the total average.

4.2. Analysis of the Results

Tables 1 and 2 contain the results in terms of Word Error Rate (WER) for clean and multi-style training respectively. Using tandem based features (MLP, GM-MLP and GM-MLP+UD) consistently provides performance improvements on clean test data with respect to conventional cepstral features, as it is known [2]. On the contrary, performance for the noisy test set decreases greatly on clean training conditions (Table 1). This is likely due to over-fitting of the MLP to the clean training data. This hypothesis is also reinforced by the results obtained in multi-style training conditions Table 2 for which the MLP outperforms the MMSE-MFCC baseline.

The GM-MLP method introduced in this work outperforms the MLP baseline in all scenarios, but suffers the same over-fitting problem with the clean trained models as the MLP. Interestingly, using UD to compensate the GMM-HMM system for the uncertainty of inference brings notable improvements in all scenarios, and leads to the best results overall. It is important to underline that UD is normally used to compensate for uncertainty arising from environmental distortions e.g. [14]. In these scenarios little or no improvements are attained for clean tests and in particular multi-style trained models. Compensating for the uncertainty of inference of the GM-MLP brings however notable improvements in these scenarios.

Table 1. Aurora 4 WER scores for clean trained GMM-HMM and MLP. Baselines (top), GM-MLP (bottom). Best results displayed in bold.

Features	Clean	Noisy	Avg
MMSE-MFCC	9.5	30.4	27.4
MLP	9.2	*38.6*	*34.4*
GM-MLP	8.5	36.4	32.4
GM-MLP+UD	**7.5**	**29.2**	**26.2**

Table 2. Aurora 4 WER scores for multi-style trained GMM-HMM and MLP. Baselines (top), GM-MLP (bottom). Best results displayed in bold.

Features	Clean	Noisy	Avg
MMSE-MFCC	13.3	19.7	18.8
MLP	9.8	18.9	17.6
GM-MLP	9.4	18.5	17.2
GM-MLP+UD	**8.4**	**16.0**	**14.9**

Finally, it is also worth noting that the mean normalization after PCA played a fundamental role on the performance of the GM-MLP with no UD. This speaks for the mismatch between training and test commented in Section 3.2.

5. CONCLUSIONS

In this paper we have introduced a new approximation which goes beyond the conventional mean field approximation for MLPs/DNNs. This approximation propagates not only the mean activations but also their variance through the network. In this way, the resulting uncertainty of inference at each layer is taken into consideration. The resulting algorithm is also closer to the view of the MLP as inference through marginalization of the hidden activations. At the current stage, no approximation for the soft-max layers has been found, which limits the use and performance of the method due to train/test mismatch. However, using the uncertainty of inference for the dynamic compensation of GMM-HMM acoustic models based on tandem features consistently outperforms the conventional MLP-based tandem approach. This resulted particularly convenient for the case of clean speech and multi-style trained systems, where dynamic compensation brings usually little improvement. The extension of this approach to cope with network layers that use the soft-max activation function opens interesting future research directions.

6. REFERENCES

[1] N. Morgan and H. Bourlad, "An introduction to hybrid HMM/connectionist continuous speech recognition," *IEEE Signal Processing Magazine*, vol. 12, no. 3, pp. 25–42, 1995.

[2] Hynek Hermansky, Daniel P. W. Ellis, and Sangita Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. ICASSP*, 2000, pp. 1635–1638.

[3] Frank Seide, Gang Li, and Dong Yu, "Conversational speech transcription using context-dependent deep neural networks," in *Proc. Interspeech*, 2011, pp. 437–440.

[4] George E. Dahl, Dong Yu, , Li Deng, and Alex Acero, "Context-dependent pre-trained deep neural networks for large vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[5] Hang Su, Gang Li, Dong Yu, , and Frank Seide, "Error Back Propagation For Sequence Training Of Context-Dependent Deep Networks For Conversational Speech Transcription," in *Proc. ICASSP*, 2013, pp. 6664–6668.

[6] J. Droppo, A. Acero, and Li Deng, "Uncertainty decoding with SPLICE for noise robust speech recognition," in *Proc. ICASSP*, 2002, vol. 1, pp. I–57–I–60 vol.1.

[7] Steve Renals, Nelson Morgan, Herv Bourlard, Michael Cohen, and Horacio Franco, "Connectionist probability estimators in hmm speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 2, no. 1, pp. 161–174, 1994.

[8] Lawrence Saul, Tommi Jaakkola, and Michael I. Jordan, "Mean field theory for sigmoid belief networks," *Journal of Artificial Intelligence Research*, vol. 4, pp. 61–76, 1996.

[9] Y. Lee and S.-H. Oh, "Input Noise Inmunity of Multilayer Perceptrons," *ETRI*, vol. 16, pp. 35–43, Apr 1994.

[10] R. F. Astudillo and J. P. Neto, "Propagation of Uncertainty through Multilayer Perceptrons for Robust Automatic Speech Recognition," in *Proc. Interspeech*, 2011, pp. 461–464.

[11] D. Kolossa and R. Haeb-Umbach, Eds., *Robust Speech Recognition of Uncertain or Missing Data - Theory and Applications*, Springer, 2011.

[12] A. Honkela, "Approximating Nonlinear Transformations of Probability Distributions for Nonlinear Independent Component Analysis," in *Proc. of the IEEE International Joint Conference on Neural Networks (IJCNN)*, 2004, pp. 2169–2174.

[13] Guenter Hirsch, *Experimental Framework for the Performance Evaluation of Speech Recognition Front-ends on a Large Vocabulary Task*, Niederrhein University of Applied Sciences, November 2002.

[14] R. F. Astudillo and R. Orglmeister, "Computing MMSE Estimates and Residual Uncertainty directly in the Feature Domain of ASR using STFT Domain Speech Distortion Models," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 5, pp. 1023 – 1034, May 2013.

[15] S. Young, *The HTK Book (for HTK Version 3.4)*, Cambridge University Engineering Department., 2006.

[16] Alberto Abad and João Paulo Neto, "Incorporating acoustical modelling of phone transitions in an hybrid ANN/HMM speech recognizer," in *Proc. Interspeech*, 2008, pp. 2394–2397.