ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9414379

MIXTURE OF INFORMED EXPERTS FOR MULTILINGUAL SPEECH RECOGNITION

Neeraj Gaur[*]*, Brian Farris*[*]*, Parisa Haghani, Isabel Leal,*
Pedro J. Moreno, Manasa Prasad, Bhuvana Ramabhadran, Yun Zhu

Google Inc.

ABSTRACT

When trained on related or low-resource languages, multilingual speech recognition models often outperform their monolingual counterparts. However, these models can suffer from loss in performance for high resource or unrelated languages. We investigate the use of a mixture-of-experts approach to assign per-language parameters in the model to increase network capacity in a structured fashion. We introduce a novel variant of this approach, 'informed experts', which attempts to tackle inter-task conflicts by eliminating gradients from other tasks in these task-specific parameters. We conduct experiments on a real-world task with English, French and four dialects of Arabic to show the effectiveness of our approach. Our model matches or outperforms the monolingual models for almost all languages, with gains of as much as 31% relative. Our model also outperforms the baseline multilingual model for all languages by up to 9% relative.

Index Terms— end-to-end speech recognition, multilingual, RNN-T, language id, mixture of experts

1. INTRODUCTION

Multilingual automatic speech recognition (ASR) models can be trained to transcribe speech in different languages. They are attractive as they can improve the performance of low-resource languages by learning shared representations from data available from other languages [1, 2, 3]. With the advent and progress of end-to-end speech recognition models [4] in the recent years, multilingual modeling has been further simplified, replacing conventional ASR system such as the acoustic, language, and the pronunciation models, with a single unified model to provide multilingual recognition.

With the exception of a few recent works [5, 6, 7], most previous work on multilingual speech recognition focuses on the benefits of these models for lower-resource or related languages. Nevertheless, in order for these models to be utilized in real-world scenarios and replace their monolingual counterparts, they need to target a variety of languages, with large variations in amounts of training data. In this paper, we propose one multilingual model to transcribe languages with varied amounts of training data. We use the mixture-of-experts approach (MOE) [8] and adapt it to exploit the inherent structure of the data to simultaneously learn per-language experts. We make the following novel contributions in this paper:

- Introduction of an informed mixture-of-experts layer, used in the encoder of an RNN-T model, where each expert is assigned to one language, or set of related languages. The identity of the spoken language (LangID) is used to combine the different expert activations and back-propagate errors from only the relevant training data.

- An LSTM-based gating mechanism that replaces the simple projection layer in the mixture-of-experts model and implicitly learns the LID.

These techniques allow us to train a streaming multilingual model that does not require prior knowledge of LangID at inference time. Our experiments on 6 languages with $\sim$ 100K hours of available training data show that this outperforms a baseline multilingual model by as much as 9%. This model also outperforms most of the monolingual models improving the performance of the lowest resource language, Arabic Maghrebi by 31% relative.

2. BACKGROUND

A variety of approaches have been explored for changing the structure of the neural network model to make it more amenable to multilingual modeling. In the context of encoder-decoder models, [5] used adapter layers to account for different amounts of available data per language, [9] parameterized the attention heads of a Transformer-based encoder to be per-language, while [6] showed that a multi-decoder multilingual model, where each decoder is assigned to a cluster of languages, can achieve good performance.

Since the introduction of Mixture of Experts (MOE) in [10], these models have found popularity in machine translation [8], and speech recognition [11, 12]. In the context of end-to-end multilingual ASR, [13] investigated using MOE

[*] The first two authors have equal contribution. The rest of the list is sorted alphabetically.

Authorized licensed use limited to: IEEE Xplore. Downloaded on May 25,2021 at 06:38:24 UTC from IEEE Xplore. Restrictions apply.

for a bi-lingual code switching system using separate encoders initialized from pre-trained monolingual encoders. The MOE model is composed of multiple expert models. Each expert learns to do well on a subset of the data and the outputs from each expert can be combined together, via a gating function, to produce a model that does well on the complete data set. This technique was further expanded by [8], applying it at a more granular level with a neural network being composed of different sub-networks, each of which was a Mixture of Experts. In this work we use the model originally proposed by [8] and investigate the application of MOE at the level of an encoder layer. Unlike [13], our proposed model does not require any pre-training, and the whole model can be jointly trained from scratch leading to much simpler training recipes.

Previous work on multilingual models also emphasizes the benefits of including the LangID as an input to the model. Language gating was proposed in [14]. LangID features have also been combined with either the input acoustic feature itself [15, 16, 5] or with the target label embedding [15]. We also make use of the LangID in the MOE method for guiding the gating mechanism. Furthermore, we introduce an LSTM-based gating mechanism that can implicitly learn LangID and allows the model to be used without prior knowledge of the spoken language of an utterance during inference.

3. MIXTURE OF INFORMED EXPERTS

The basic architecture of the *Mixture of Informed Experts (MIE)* is similar to MOE, the difference being that we exploit the structure in multilingual data by *specializing* the experts on data from a particular language. The experts are each pre-assigned to a particular language group and the parameters for the experts are only updated by data from their pre-assigned language groups. Optionally, one of the experts can be a *generalist*, i.e. all language groups get assigned to that expert. Figure 1 shows an encoder-decoder model with MIE encoder layers.

During forward propagation all the experts in a layer are fed the same input and the output is a convex combination of the outputs of each of the expert (see Figure 1). We refer to this as *collaboration*. This is similar to how MOE processes the data in forward propagation. *Specialization* is achieved during backpropagation by modifying the gradients such that the weights for an expert are only modified by the utterances from it's pre-assigned language group. This can be achieved by simply zeroing out the gradients corresponding to utterances which do not belong to the language group of interest. The relative importance of *collaboration*, *specialization*, and the *generalist* will be discussed in the ablation study described in Section 5.

The output of an MOE/MIE layer is the weighted summation of the outputs of the experts.

$$l_{out} = \sum_i \alpha_i * e_i$$

where, l_{out} is the output of the MOE/MIE layer, and e_i is the output of expert i. The weight vector $\alpha = [\alpha_i, \alpha_2, ..., \alpha_n]$ is given by

$$\alpha = \text{softmax}(G(g_{in}))$$

where G is the gating network and n is the number of experts. In our standard MIE setup, G is an affine projection of the oracle LangID. We refer to this as a "language gate". To remove the dependence on oracle LangID, we tried two variations which use the layer input for g_{in} instead of the oracle LangID; an affine projection and an LSTM gating network. When using an LSTM gate, the number of parameters could become very large if a different LSTM gate is used for each layer. Instead, we tie the weights at every layer. Results from these gate variations are discussed in Section 5.6.

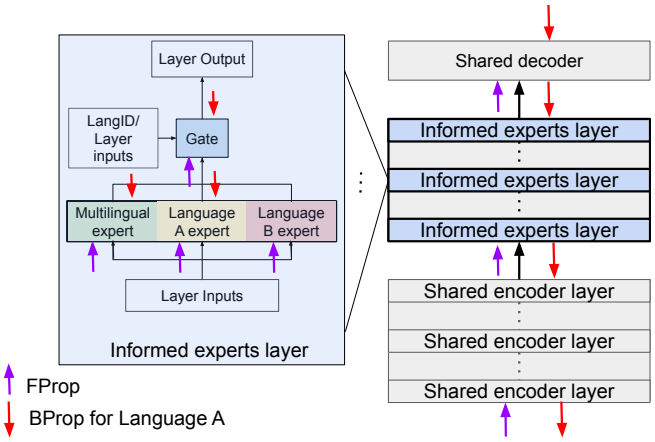

Fig. 1: *Mixture of Informed Experts(MIE) layers in the encoder of an encoder-decoder neural network. Here each MIE layer has 3 experts, 2 experts that are specific to specific language groups and one multilingual expert. During the forward pass, the output of the layer is a convex combination of the output of each expert with the weights given by the gate. The experts are specialized during the backward pass. For example, for language A, only the language A expert and the multilingual expert are updated.*

4. EXPERIMENTAL SETUP

4.1. Languages and Data

We conduct our experiments on six locales: four dialects of Arabic (Egyptian, Gulf, Levant, and Maghrebi), English and French. We chose this combination of languages as a phonetically and lexically diverse set with subsets that are generally mutually intelligible (Arabics). These languages also

Authorized licensed use limited to: IEEE Xplore. Downloaded on May 25,2021 at 06:38:24 UTC from IEEE Xplore. Restrictions apply.

Table 1: *Summary of the number of hidden LSTM cells in the shared encoder layers N_{hid}^{shared}, number of hidden LSTM cells in the MIE layers N_{hid}^{mie}, number of shared layers N_{layer}^{shared}, number of MIE layers N_{layer}^{mie}, and number of experts(generalists) $N_{expert}(N_{gen})$ for each experiment. mie/moe refers to the mie, nospec, nocollab, proj, and LSTM experiments.*

Experiment	mono	cotrain	mie/moe	nogen
N_{hid}^{shared}	2048	4096	2048	2048
N_{hid}^{MIE}	NA	NA	2048	2730
N_{layer}^{shared}	8	8	4	4
N_{layer}^{MIE}	0	0	4	4
$N_{expert}(N_{gen})$	NA	NA	3(1)	3(0)

have a large range in the amount of training data for each language, namely, 3600 hours in Maghrebi to 37K hours in English or French, thus presenting challenges for multilingual ASR (similar to [6]). This set of languages also contains transcriptions with distinct grapheme sets (Arabic and Latin). For all of these languages, the training and test data are anonymized and transcribed by humans. Similar to [9], we use two forms of data augmentation to mitigate overfitting and improve generalization, namely noise and reverberation based augmentation detailed in [17] and SpecAugment [18].

4.2. Models

The input acoustic features to the model are 80-dimensional log-Mel features stacked over three frames as described in [16]. Our model follows the encoder-decoder RNN-T streaming architecture detailed in [19] with changes to the encoder layers. The model architectures used throughout this paper are detailed in Table 1. In order to have a fair comparison, the number of parameters in the multilingual baseline are increased to match that of the MIE architecture. For each of our MOE/MIE experiments, we use the same architecture for the first 4 layers of the encoder, while the latter 4 layers are replaced by MOE/MIE layers, as described in Section 3. Each expert within a given MIE layer is an LSTM layer with 2048 hidden units, except in the no-generalist experiments (denoted "nogen"), where the number of hidden units is increased by a factor of $4/3$ to 2730 in order to match the capacity of the models with a generalist. When an LSTM layer is used for the gating network, it has 2048 hidden units. The target vocabulary is the union of graphemes from all languages. For all models with specialization, the outputs of the experts are uniformly averaged and we do not specialize, i.e., we backpropagate through every expert, for the first N_{warm} steps. $N_{warm} = 70k$ for all experiments except LSTM gate, for which $N_{warm} = 35k$. All models are trained in using 8x8 Tensor Processing Units with an effective batch size of 16384, except for our monolingual baselines, which are trained on 4x4 TPUs with an effective batch size of 4096.

5. RESULTS

5.1. Baseline monolingual and multilingual models

We train monolingual models as baselines for our experiments. The architecture for each monolingual model is described in Table 1 and each is trained with data from that respective language only. We train a "cotraining " model with the same model architecture as monolingual models, but with data combined from all languages. Model capacity is increased by doubling the number of hidden units in each encoder layer, and the effective batch size is increased by 4x. Results are shown in the "mono" and "cotrain" columns of Table 2. Compared to the monolingual baselines we find WER improvements as high as 25% relative for the relatively low-resource arabic dialects. However, we see WER regressions of 1% and 3% relative for en-us and fr-fr respectively. Similar performance issues on high-resource languages have been noted in previous multilingual ASR work as well [6].

5.2. MIE

In each MIE layer we include 3 experts, one with weights which are only updated by Arabic data, one which is only updated by English data, and one which is only updated by French data. We include a "generalist" with weights which are updated for all data. The output from each expert/generalist is combined by a language gate, as described in Section 3. We append the LangID as a one-hot vector to the input features [15]. We find that MIE can match or exceed the performance of both the cotraining baseline and monolingual baselines for every language except French, which shows a marginal regression under 1% relative. Results are shown in the "mie" column of Table 2.

5.3. Generalist

The motivation for the inclusion of the generalist in the MIE implementation was to allow each expert to focus on generating language-specific features, while allowing more universal features to be represented once by the generalist. In Figure 2 we show the relative weights given to each expert/generalist in each layer. When the generalist is included, we find that it receives close to half the weight, supporting it's inclusion. However, in order to measure it's importance, we trained a model with no generalist, but increased the number of hidden units in each expert LSTM layer by a factor of 4/3 in order to keep the number of model parameters consistent. Surprisingly, the results are largely unchanged (Table 2). From the "no generalist" gate outputs in Figure 2, we see that the weight from the generalist is almost entirely transferred to that of the correct language expert.

Authorized licensed use limited to: IEEE Xplore. Downloaded on May 25,2021 at 06:38:24 UTC from IEEE Xplore. Restrictions apply.

Table 2: *Ablation study comparing the WER rates from mie model to versions with either specialization disabled (nospec), collaboration disabled (nocollab), or the generalist removed (nogen).*

	Baselines			Ablations			MIE Gate Variants	
Experiment	**mono**	**cotrain**	**mie**	**nogen**	**nospec**	**nocollab**	**proj**	**LSTM**
ar-eg	14.0	13.9	13.3	13.1	13.6	14.0	14.9	13.2
ar-x-gulf	12.8	11.4	10.9	10.9	11.1	11.6	11.7	10.9
ar-x-levant	15.0	14.1	13.7	13.4	13.8	14.0	14.8	13.6
ar-x-maghrebi	17.5	13.2	12.0	12.1	12.0	12.5	12.9	12.1
en-us	6.1	6.2	6.1	6.0	6.2	6.4	6.6	6.1
fr-fr	13.1	13.5	13.2	13.4	13.2	13.2	13.5	13.4

5.4. Specialization

In order to assess the role of specialization in the improvements we get from MIE, we trained a model with specialization turned off, meaning that the weights in each expert can be updated by data from any language. We find that this degrades the performance slightly, with relative WER regressions $\leq 2\%$, indicating that, when using language gate, specialization leads to marginal benefits. Results are shown in the "nospec" column of Table 2.

5.5. Collaboration

In order to determine the importance of collaboration in the performance of the MIE model, we also trained a model (denoted "nocollab") in which we pass the output of each expert in a given MIE layer directly as input to the corresponding expert in the next layer, without using the gate. The output of each expert in the last MIE layer is combined with a single language gate before being passed to the decoder. We find that removing collaboration leads to a significant degradation in performance with relative WER regressions as high as 6%. In future work we intend to investigate whether this holds true for other collections of languages. Results are shown in the "nocollab" column of Table 2.

5.6. MIE Gate variants

While MIE performs well, the use of the language gate demands that LangID be available at inference time. In order to circumvent this limitation, we trained models in which the language gate is replaced with either an LSTM gate or a affine projection gate, as described in Section 3. Note that while LangID is used during training in order to mask gradients in the MIE layers and achieve specialization, it is not used during inference. The input to the LSTM gate is the shared encoder output, and the output is a score for each expert, which is the same for every MIE layer. While the results for the affine projection gate are somewhat degraded in comparison with the language-gate model, the LSTM gate is able to essentially match the performance of the MIE model. This indicates that the LSTM gate, coupled with specialization, can learn a reasonably accurate implicit LangID model without modifying the loss function in any way. Results are shown in Table 2.

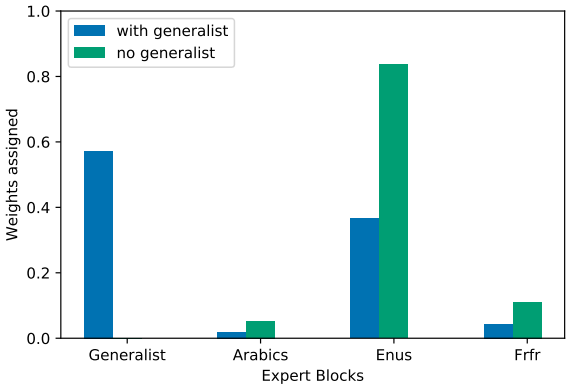

Fig. 2: *Weight assigned to each expert in the final MIE layer, from experiments using language gate, with the LangID corresponding to en-us. Outputs with a generalist included (not included) are shown in blue (green).*

6. CONCLUSION

This paper presents a multilingual ASR system that achieves significant improvements in WER of up to 9% relative on lower-resource languages while improving or maintaining performance on high resource languages. We demonstrated that we can remove inference time dependence on an external LangID for these languages by exploiting structure in multilingual data along with an LSTM based gating network. The proposed MIE model with its gating mechanism is trained jointly leading to a simple training procedure.

Authorized licensed use limited to: IEEE Xplore. Downloaded on May 25,2021 at 06:38:24 UTC from IEEE Xplore. Restrictions apply.

7. REFERENCES

[1] Alex Waibel, Hagen Soltau, Tanja Schultz, Thomas Schaaf, and Florian Metze, "Multilingual speech recognition," in *Verbmobil: Foundations of Speech-to-Speech Translation*, pp. 33–45. Springer, 2000.

[2] 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 *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7304–7308.

[3] Arnab Ghoshal, Pawel Swietojanski, and Steve Renals, "Multilingual training of deep neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7319–7323.

[4] Alex Graves, "Sequence transduction with recurrent neural networks," *arXiv preprint arXiv:1211.3711*, 2012.

[5] Anjuli Kannan, Arindrima Datta, Tara N Sainath, Eugene Weinstein, Bhuvana Ramabhadran, Yonghui Wu, Ankur Bapna, Zhifeng Chen, and Seungji Lee, "Large-scale multilingual speech recognition with a streaming end-to-end model," *arXiv preprint arXiv:1909.05330*, 2019.

[6] Vineel Pratap, Anuroop Sriram, Paden Tomasello, Awni Hannun, Vitaliy Liptchinsky, Gabriel Synnaeve, and Ronan Collobert, "Massively multilingual asr: 50 languages, 1 model, 1 billion parameters," in *Proc. INTERSPEECH*, 2020.

[7] Wenxin Hou, Yue Dong, Bairong Zhuang, Longfei Yang, Jiatong Shi, and Takahiro Shinozaki, "Large-scale end-to-end multilingual speech recognition and language identification with multi-task learning," in *Proc. INTERSPEECH*, 2020.

[8] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean, "Outrageously large neural networks: The Sparsely-Gated Mixture-of-Experts layer," *arXiv preprint arXiv:1701.06538*, 2017.

[9] Yun Zhu, Parisa Haghani, Anshuman Tripathi, Bhuvana Ramabhadran, Brian Farris, Xu Hainan, Han Lu, Hasim Sak, Isabel Leal, Neeraj Gaur, Pedro Moreno, and Qian Zhang, "Multilingual speech recognition with self-attention structured parameterization," in *Proc. INTERSPEECH*, 2020.

[10] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton, "Adaptive mixtures of local experts," *Neural Comput.*, vol. 3, no. 1, pp. 79–87, 1991.

[11] Sara Papi, Edmondo Trentin, Roberto Gretter, Marco Matassoni, and Daniele Falavigna, "Mixtures of Deep Neural Experts for Automated Speech Scoring," in *Proc. Interspeech 2020*, 2020, pp. 3845–3849.

[12] Abhinav Jain, Vishwanath P. Singh, and Shakti P. Rath, "A Multi-Accent Acoustic Model Using Mixture of Experts for Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 779–783.

[13] Yizhou Lu, Mingkun Huang, Hao Li, Jiaqi Guo, and Yanmin Qian, "Bi-Encoder Transformer Network for Mandarin-English Code-Switching Speech Recognition Using Mixture of Experts," in *Proc. Interspeech 2020*, 2020, pp. 4766–4770.

[14] Suyoun Kim and Michael L Seltzer, "Towards language-universal end-to-end speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4914–4918.

[15] Bo Li, Tara N Sainath, Khe Chai Sim, Michiel Bacchiani, Eugene Weinstein, Patrick Nguyen, Zhifeng Chen, Yanghui Wu, and Kanishka Rao, "Multi-dialect speech recognition with a single sequence-to-sequence model," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4749–4753.

[16] Austin Waters, Neeraj Gaur, Parisa Haghani, Pedro Moreno, and Zhongdi Qu, "Leveraging language id in multilingual end-to-end speech recognition," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 928–935.

[17] Chanwoo Kim, Ananya Misra, Kean Chin, Thad Hughes, Arun Narayanan, Tara Sainath, and Michiel Bacchiani, "Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in Google Home," in *Proc. INTERSPEECH*, 2017.

[18] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin Dogus Cubuk, and Quoc V. Le, "Specaugment: A simple augmentation method for automatic speech recognition," in *INTERSPEECH*, 2019.

[19] Yanzhang He, Tara Sainath, Rohit Prabhavalkar, Ian McGraw, Raziel Alvarez, Ding Zhao, David Rybach, Anjuli Kannan, Yonghui Wu, Ruoming Pang, Qiao Liang, Deepti Bhatia, Yuan Shangguan, Bo Li, Golan Pundak, Khe Chai Sim, Tom Bagby, Shuo yiin Chang, Kanishka Rao, and Alex Gruenstein, "Streaming end-to-end speech recognition for mobile devices," 2019.

Authorized licensed use limited to: IEEE Xplore. Downloaded on May 25,2021 at 06:38:24 UTC from IEEE Xplore. Restrictions apply.