Mask-Net: Learning Context Aware Invariant Features Using Adversarial Forgetting (Student Abstract)

Hemant Yadav , Rajiv Ratn Shah
IIIT Delhi, India
{hemantya, rajivratn}@iiitd.ac.in

Abstract

Training a robust system, *e.g.,* Speech to Text (STT), requires large datasets. Variability present in the dataset such as unwanted nuances and biases are the reason for the need of large datasets to learn general representations. In this work, we propose a novel approach to induce invariance using adversarial forgetting (AF). Our initial experiments on learning invariant features such as accent on the STT task achieve better generalizations in terms of word error rate (WER) compared to traditional models. We observe an absolute improvement of 2.2% and 1.3% on out-of-distribution and in-distribution test sets, respectively.

Introduction

Given a task with input X to predict a set of labels Y, deep learning (DL) systems often learn an undesired dependence on unwanted factors present in the input given the task (Domingos 2012) *e.g.,* STT. Therefore the need arises to learn invariant features to avoid the undesired dependencies. Previous work in this direction includes enforcing the invariance throughout all of the layers of the encoder network using the gradient reversal learning (GRL) technique (Ganin and Lempitsky 2015). (Jaiswal et al. 2020) propose to model the problem of learning invariant features by learning a mask $M \in (0, 1)$, given X, which on element wise multiplication with the encoder output Z produces an invariant feature $\tilde{Z}$. Our proposed method builds upon the two major shortcomings of this work.

Firstly, in their model setup, the authors learn low level features twice, once in the encoder and then again in the forget net module. It results in learning redundant features and increased training time since both of the modules have an identical architecture. Secondly, if the main task is based on time series data, *e.g.,* STT or machine translation, there is no notion of global context when learning the mask *M*. This hampers learning what to retain and what to forget *i.e.,* the M values. Furthermore, each value of mask M is learnt independently such that there is no correlation between the mask values over different time steps. This results in different mask values for different time steps, which is an undesirable behaviour.

Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

In this student abstract work, we propose a novel approach to address these issues and propose a new method combining two papers i.e., (Jaiswal et al. 2020) and (Han et al. 2020) with details provided in the next section.

Method

We propose a novel approach to induce invariance to unwanted factors, *e.g.,* accent present in the audio for the STT task. Let X be the input, Y be the output and, A be the unwanted bias. We propose an architecture with one encoder module *E* with output Z, one forget net module *F* with output $M \in (0, 1)$, one discriminator module *D* with output A and, one predictor module *P* with output Y. The mask M is then multiplied element wise with Z to get $\tilde{Z}$ as shown in Figure 1.

As shown in Figure 1 the input to forget net is Z instead of the raw input X. Therefore we bypass the redundant step to learning the low level features through the forget net module leading to reduction in training time. Secondly, in order to introduce the notion of global context we propose to create the forget net module similar to the squeeze and excitation module presented in ContextNet (Han et al. 2020). As shown in Figure 1, we first average pool over all the time steps to introduce the global context and then apply a linear layer such that the output dimension is reduced by eight (squeeze step). In the next step,a linear layer is applied again such that the output dimension is restored to the previous step (excitation step). In the last step we tile the output to match the dimension of Z. The last step ensures that all the time steps have the same mask value and thus the embeddings at each time step, map to a similar feature. We propose the above two architectural changes compared to (Jaiswal et al. 2020) and call the resulting network mask-Net (M-Net) such that it learns context aware invariant features using adversarial forgetting as shown in Figure 1.

The mask M is learned in an adversarial manner similar to the GRL technique with a distinction that the gradients from the discriminator only flow up to the forget net as shown by the dotted red lines in Figure 1. Therefore the encoder learning is purely driven by the main task. This is very useful in case we want to use Z for some other downstream task which requires the accent information. A practical application is to learn a common strong encoder using self supervised learning (SSL) and then use our method to learn n

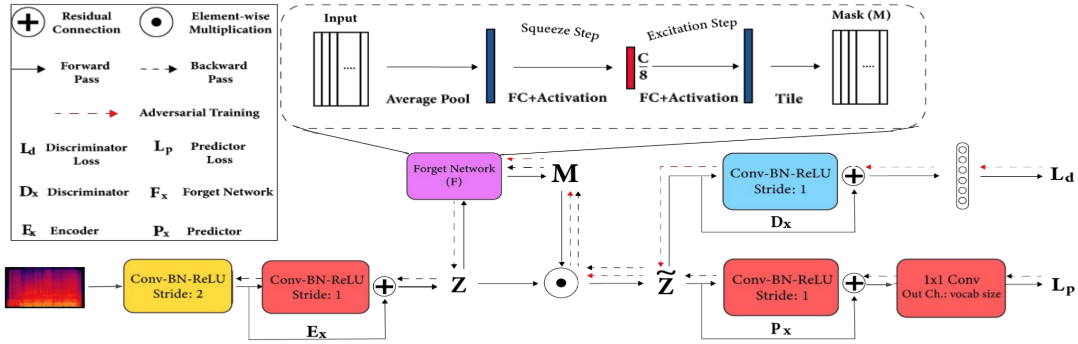

Figure 1: Mask-Net architecture for learning invariant features using adversarial forgetting. The rectangular blocks represent 1D CNN layers with Batch-Norm (BN) and ReLU and x denotes the number of repeated blocks as proposed by the authors of Jasper paper (Li et al. 2019). Each block can have n layers internally. Residual connections are only active if the number of layer within the block is greater than 1. Dotted arrows show the flow of gradients corresponding to each loss function. Red and black dotted line shows the contribution of the discriminator loss and the predictor loss in updating the weights of different modules. Note: figure better viewed on a device.

masks given n biases. This is not possible in the standard GRL technique. Lastly we expect our method to achieve results comparable to the standard GRL training technique if not better.

Preliminary Evaluation and Results

Similar to (Jain and Jyothi 2018), we evaluate our approach on a subset of the Common Voice V1 dataset, 34 hrs for training. Where accent classification is the discriminator task and STT is the predictor/main task. No data augmentation has been applied on the input spectrogram for training and testing purposes. Predictor and discriminator are trained with connectionist temporal classification (Graves et al. 2006) and cross entropy loss respectively. We report WER with a beam width of 512. All models are trained using mixed precision with Pytorch

Surprisingly our Japser (Li et al. 2019) based baseline achieves far better WERs when compared to recent work by (Jain and Jyothi 2018), in which the authors use utterance-level accent embeddings. This goes on to show that the baseline we provide is very strong. Therefore we only make comparison to our baseline and try to improve on that. We observe an overall improvement across all the test cases when compared with baseline as shown in Table 1.

Our proposed performs comparable to GRL technique as shown in Table 1. But our method is more desirable for achieving true invariant features (Jaiswal et al. 2020) as it achieves an absolute 5% lead over the other, when classifying the accents from the trained features ($\tilde{Z}$). We hope to scale our experiments on large datasets with varying biases to report a general trend.

Conclusion and Future Work

In this work we proposed a novel approach to learn invariant representations with the added benefits of global context aware mask, faster training time and, uniform mask values over all the time steps. Future work includes running experiments on bigger dataset with different biases. Furthermore to achieve true global context, we would try attention based forget net module to learn the mask M.

%	Test	TestIN	TestNZ
(Jain and Jyothi 2018)	19.74	51.2	22.7
Baseline (Li et al. 2019)	7.7	33.6	6.4
Multi-task	6.8	32.9	5.4
Gradient-reversal	6.2	30.3	4.9
Ours	**6.4**	**30.5**	**5.1**

Table 1: WER is reported on in-distribution test set *i.e.,* Test and out-of-distribution test sets *i.e.,* TestIN and TestNZ.

Acknowledgments

Rajiv Ratn Shah was partly supported by the Infosys Center for Artificial Intelligence and the Center of Design and New Media at IIIT Delhi, India.

References

Domingos, P. 2012. A few useful things to know about machine learning. *Communications of the ACM*, 55(10): 78–87.

Ganin, Y.; and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *ICML*, 1180–1189.

Graves, A.; Fernández, S.; Gomez, F.; and Schmidhuber, J. 2006. Connectionist temporal classification. In *ICML*.

Han, W.; Zhang, Z.; Zhang, Y.; Yu, J.; Chiu, C.-C.; Qin, J.; Gulati, A.; Pang, R.; and Wu, Y. 2020. Contextnet. *arXiv:2005.03191*.

Jain, A.; and Jyothi, P. 2018. Improved Accented Speech Recognition Using Accent Embeddings and Multi-task Learning. In *Interspeech*, 2454–2458.

Jaiswal, A.; Moyer, D.; Ver Steeg, G.; AbdAlmageed, W.; and Natarajan, P. 2020. Invariant Representations through Adversarial Forgetting. In *AAAI*, 4272–4279.

Li, J.; Lavrukhin, V.; Ginsburg, B.; Leary, R.; Kuchaiev, O.; Cohen, J. M.; Nguyen, H.; and Gadde, R. T. 2019. Jasper. arXiv:1904.03288.