SAN: A ROBUST END-TO-END ASR MODEL ARCHITECTURE

Zeping Min\(^*1\) Qian Ge\(^*1\) Guanhua Huang\(^*2\)

\(^{1}\) Peking University
\(^{2}\) University of Science and Technology of China

ABSTRACT

In this paper, we propose a novel Siamese Adversarial Network (SAN) architecture for automatic speech recognition, which aims at solving the difficulty of fuzzy audio recognition. Specifically, SAN constructs two sub-networks to differentiate the audio feature input and then introduces a loss to unify the output distribution of these sub-networks. Adversarial learning enables the network to capture more essential acoustic features and helps the models achieve better performance when encountering fuzzy audio input. We conduct numerical experiments with the SAN model on several datasets for the automatic speech recognition task. All experimental results show that the siamese adversarial nets significantly reduce the character error rate (CER). Specifically, we achieve a competitive 4.37 CER result on the AISHELL-1 dataset without any language models. To reveal the generality of the siamese adversarial net, we also conduct experiments on the phoneme recognition task, which also shows the superiority of the siamese adversarial network.

Index Terms— Automatic speech recognition, adversarial learning, fuzzy audio, siamese net

1. INTRODUCTION

Automatic speech recognition (ASR) is a traditional task with great application value. There has a long history on automatic speech recognition (ASR). In the early days, Kaldi employed the HMM-GMM probability models and achieved quite good performance.

Recently, the E2E models \([1, 2, 3, 4]\) have become a hot topic in the field of speech recognition. But these models often face challenges when faced with fuzzy audio, especially when high background noise is superimposed. To some extent, the fuzzy audio can be seen as an attack on the E2E model. For example, the word "wood" and "world" have very similar pronunciations. Most previous works add an additional language model to the decoder to deal with the fuzzy audio problem. Adding a language model \([5, 6]\) to the decoding process can solve some of the problems. The language model performs well when fuzzy audio is easy to distinguish semantically. For instance, in the audio with transcription of "There are billions of people in the world.", the raw model output may have a high probability on both "There are billions of people in the world." and "There are billions of people in the wood." predictions. And it is easy to eliminate the latter interference option by using the language model.

Unfortunately, this does not work all the time. When the interference option is also semantically meaningful, the language model can not help eliminate the wrong option and may even mislead the output, such as the "I like the wood" and "I like the world.". Both of them are semantically meaningful, but "I like the world" are more common. For input wave with the ground truth transcription "I like the wood.", the speech recognition model output may also have a high probability on the transcription of "I like the world.". Further, the language model may also vote for the wrong option "I like the world.". Finally, the model is prone to make mistakes on the audio input.

In this paper, inspired by \([7]\), we propose a novel siamese adversarial net (SAN) architecture for automatic speech recognition, which aims at solving the difficulty of recognizing fuzzy audio. In detail, SAN consists of two weight-shared sub-networks, which employ the dropout layers to make the acoustic features of the two sub-networks different. Then a Kullback–Leibler (KL) divergence is leveraged to minimize the output distributions of these two sub-networks, which boosts the model to extract and focus on the essential acoustic features to help the model deal with the fuzzy audio input. As experimental results, we achieve a competitive 4.37 CER result on AISHELL-1 dataset. In summary, our contributions are as follows:

- We propose a novel siamese adversarial net (SAN) architecture, solving the difficulty of recognizing fuzzy audio by adversarial learning with two subnets.
- We fulfill the gap that few works take care of the fuzzy audio recognition in the acoustic model itself.
- We achieve a competitive 4.37 CER without language model on the AISHELL-1 dataset. Besides, a large number of experiment results show that our SAN architecture is effective.
2. RELATED WORK

Deep learning has revolutionized the task of speech recognition. Common E2E speech recognition deep learning models include [1, 8, 9]. While these deep learning models have made big difference in speech recognition, they tend to make mistakes when faced with fuzzy audio input.

Combining the language model (LM) can temper part of the fuzzy audio problem. There is a lot of research focused on using language models such as [10, 11]. Among them, shallow fusion [10], adding weighted language model prediction probability in the decoding stage, is one of the most direct ways to fuse language model information. Cold fusion [11] is another way to efficiently use language models. Moreover, [6] introduce the language model for enhancing rare word recognition.

In this article, we propose a novel siamese adversarial network (SAN) architecture, which uses dropout to strengthen fuzzy audio attacks on the network and leverage KL divergence to help the model to extract and focus on essential features. Similar to us, there exist some works in NLP that employs dropout to help the model extracts and focus on essential features. Sim-ilar to us, there exist some works in NLP that employs dropout to help the model extracts and focus on essential features. Similar to us, there exist some works in NLP that employs dropout to help the model extracts and focus on essential features.

In this article, we propose a novel siamese adversarial network (SAN) architecture, which uses dropout to strengthen fuzzy audio attacks on the network and leverage KL divergence to help the model to extract and focus on essential features. Similar to us, there exist some works in NLP that employs dropout to help the model extracts and focus on essential features. Similar to us, there exist some works in NLP that employs dropout to help the model extracts and focus on essential features.

In this article, we propose a novel siamese adversarial network (SAN) architecture, which uses dropout to strengthen fuzzy audio attacks on the network and leverage KL divergence to help the model to extract and focus on essential features. Similar to us, there exist some works in NLP that employs dropout to help the model extracts and focus on essential features. Similar to us, there exist some works in NLP that employs dropout to help the model extracts and focus on essential features.

3. METHOD

As shown in Figure 1, the subnet in SAN use the joint CTC/Attention framework architecture as Wenet [3]. To deal with the fuzzy audio problem, SAN employs the KL loss to help extract the essential feature for predicting the fuzzy part.

3.1. Encoder of the subnet

Given a sequence of preprocessed audio feature

$$S = \{s_1, s_2, ..., s_{N_s}\}$$ (1)

where \(s_i \in \mathbb{R}^d\), \(d\) is the embedding dimension and \(N_s\) is the length of the preprocessed audio feature sequence. We firstly encode the preprocessed audio feature by conformer [2]:

$$H = \text{Conformer}([s_1, s_2, ..., s_{N_s}])$$ (2)

and obtain the speech’s hidden representation

$$H = [h_1, h_2, ..., h_{N_h}],$$

3.2. Decoder of the subnet

3.2.1. CTC

After getting the hidden representation \(H\) of speech, the CTC loss \(\mathcal{L}_{\text{CTC}}\) is calculated in each SAN subnet.

$$\mathcal{L}_{\text{CTC}} = \text{CTC}([h_1, h_2, ..., h_{N_h}])$$ (3)

For attention part of SAN subnet, we employ a transformer decoder with the cross attention to hidden representation \(H\). And we calculate the loss \(\mathcal{L}_{\text{Attn}}\) in each SAN subnet.

During inference, for the attention rescore decode mode, the transformer decoder rescores the text sequence results from CTC decoding.

3.2.2. Attention

For attention part of SAN subnet, we employ a transformer decoder with the cross attention to hidden representation \(H\). And we calculate the loss \(\mathcal{L}_{\text{Attn}}\) in each SAN subnet.

During inference, for the attention rescore decode mode, the transformer decoder rescores the text sequence results from CTC decoding.

3.3. Siamese Computing

To alleviate the fuzzy audio problem, we extract the essential feature by achieving a siamese adversarial training.

We build a siamese network [14], which consists of two sub-encoder-decoder networks sharing the same parameters.

In the adversarial learning of two sub-networks, the randomness of dropout layers differentiate the acoustic features. For fuzzy audios, this randomness of dropout is very likely to cause a significant change in the prediction distribution. The KL loss aims to forcing the two subnets having the same output distribution, which help the model extracts and focus on the essential feature in the prediction of the fuzzy part.

![Fig. 1. The model architecture of SAN. SAN consists of two weight-shared sub-networks. The dropout layers differentiate the acoustic features in two sub-networks, while the KL loss is employed to unify these acoustic features. In the sub-network of SAN, a conformer encoder is used to encode the raw audio sequence, and a CTC decoder and a transformer decoder are leveraged to combine the streaming decoding and the attention decoding.](image-url)
In detail, the same model input is first fed to the two encoder-decoder modules independently. Then two different output distributions are obtained after the computation of two sub-networks since the dropout layers in these two sub-networks randomly mask the model features, which differentiates the acoustic features.

\[
P_1 = SubNetwork - 1([s_1, s_2, ..., s_N])
\]

\[
P_2 = SubNetwork - 2([s_1, s_2, ..., s_N])
\]

Where \( P_1, P_2 \) are attention decoder output distributions of two sub-networks. Finally, a KL-divergence loss \( L_{KL} \) is employed to unify the output of two sub-network.

3.4. Training

For the loss of training SAN model, we plus the loss from two sub-networks together for CTC loss and Attention loss. That is,

\[
L_{SubNetwork-1} = \lambda_{CTC} L_{CTC}^{SubNetwork-1} + \lambda_{Attn} L_{Attn}^{SubNetwork-1}
\]

\[
L_{SubNetwork-2} = \lambda_{CTC} L_{CTC}^{SubNetwork-2} + \lambda_{Attn} L_{Attn}^{SubNetwork-2}
\]

Then we make a weighted sum over three losses \( L_{SubNetwork-1}, L_{SubNetwork-2} \) and \( L_{KL} \).

\[
L_{All} = \lambda_1 L_{SubNetwork-1} + \lambda_2 L_{SubNetwork-2} + \lambda_3 L_{KL}
\]

More training details are listed in Section 4.2.

4. EXPERIMENTS

4.1. Datasets

In order to evaluate our proposed SAN model, we firstly conduct experiments on AISHELL-1 [15] and AIDATATANG-200zh\(^1\) datasets. Both the AISHELL-1 [15] dataset and AIDATATANG-200zh dataset are open-source Mandarin speech dataset.

4.2. Experimental Settings

For each subnet in SAN, we adopt 12 conformer layers as the subnet encoder. For the subnet attention decoder, 6 transformer decoder layers are used and each layer is built with an embedding dimension of 256 and 4 heads. As the description above, the two subnets in SAN are weight-shared.

We train the SAN models for 240 epochs on AISHELL-1 and 100 epochs on AIDATATANG using Adam optimizer with a learning rate of 0.002 during the training process, and a learning rate schedule with 25000 warmup steps on AISHELL-1 and 15000 warmup steps on AIDATATANG.

In experiments on both AISHELL-1 and AIDATATANG datasets, the weights of losses are \( \lambda_1 = 0.5, \lambda_2 = 0.5, \lambda_3 = 2 \). Besides, we conducted experiments on AISHELL-1 and AIDATATANG with four 16Gb memory P100 GPUs and two 24Gb memory RTX3090 GPUs.

4.3. Main Results

| models                  | AISHELL-1 | AIDATATANG |
|-------------------------|-----------|------------|
| ESPnet Transformer [16] | 6.70      | /          |
| CAT (with language model) [17] | 6.34 | /          |
| WeNet [3]               | 4.73      | 4.72       |
| SAN(Ours)               | **4.37**  | **4.46**   |

Table 1. The results of CER (character error rate) on the different test sets under different models. SAN achieves the best performance on all datasets.

Table 1 shows the character error rate (CER) on the test set of two different datasets using End2End architecture. We compare our model with ESPnet Transformer [16], CAT [17] and WeNet\(^2\) [3]. From the results, we can see that our model achieves competitive results of 4.37 CER (character error rate) on AISHELL-1, which outperforms the previous WeNet model and achieves around 7.6% relative character error rate reduction. On AIDATATANG, our model also returns competitive results with 4.46 CER and outperforms the WeNet, and achieves around 5.5% relative CER reduction.

4.4. Case Study

In this section, we show that our SAN model can mitigate fuzzy speech recognition errors. In Fig. 2, we show an intuitive example. The first line is the output text of our model towards audio and is exactly the same as the ground truth. The second line is the inference output from WeNet facing the same input audio.

Fig. 2. A case of a fuzzy audio problem: The words in the circle, pronounced as “q`u” and “q¯u” in Pinyin, are fuzzy audio since their pronunciations are similar. We can see our SAN model successfully recognized.

We can find that our proposed SAN is more discriminative for the fuzzy audio above. And Table 2 is an overall error statistic. We can see that our SAN significantly alleviates

\(^1\)https://openslr.org/62

\(^2\)We refer to some code and results from https://github.com/wenet-e2e/wenet
the substitution errors, which contain fuzzy audio scenarios. We speculate it is because the siamese adversarial structure in SAN helps the model to have the capability of capturing and focusing on the essential features. Besides, deletion errors and insertion errors are also reduced.

| Error Type   | WeNet   | SAN     |
|--------------|---------|---------|
| Substitution | 4737    | 4378(-359) |
| Deletion     | 131     | 117(-14) |
| Insertion    | 91      | 80(-11)  |
| N            | 104765  | 104765  |

Table 2. The character number of different error types on the AISHELL-1 dataset. N is the total character number of AISHELL-1 and the differences are listed in the (brackets). Compared with WeNet, our SAN reduces the errors for all types, especially for substitution error which contains fuzzy scenarios.

5. ABLATION STUDY

In order to better understand the SAN architecture, we conduct some ablation experiments.

5.1. Effective of adversarial architecture

To show the effectiveness of adversarial architecture in SAN, we compare the experiment results of SAN and SAN without adversarial architecture. In order to be more convincing, we tried the cases where the encoder and subnet in the SAN are transformer and LSTM respectively. The result of SAN with transformer encoder (SAN with Transformer)) and SAN with LSTM subnet (SAN with LSTM)) are shown in Table 3 and Table 4. We can clearly see that the model with the siamese adversarial structure achieves a lower CER than the model without it.

| models                  | CER on dev set | CER on test set |
|-------------------------|----------------|-----------------|
| SAN with Transformer    | 16.82          | 17.45           |
| SAN with LSTM without adversarial | 15.86          | 17.04           |

Table 4. Performance comparison of SAN (LSTM subnet) with/without siamese adversarial learning architecture. After replacing the subnet with LSTM, SAN with siamese adversarial architecture still gains lower CER than SAN that without siamese adversarial architecture.

5.2. Phone recognition Task.

To demonstrate the effectiveness of our SAN model in other relative scenarios. We conducted experiments on phoneme recognition, which is an easier task. We use the SAN with GRU subnet and train on Timit dataset [18].

| models | PER on dev set | PER on test set |
|--------|----------------|-----------------|
| GRU only | 22.99          | 24.48           |
| SAN with GRU | 22.90          | 24.14           |

Table 5. The results of PER (phone error rate) on Timit under different models. We can see the simplified SAN with GRU significantly reduces the PER on the Timit dataset.

The results are shown in Table 5. We can see that the SAN with GRU model archives a distinct decrease in PER for both the test set and the dev set.

6. CONCLUSION

We propose a novel Siamese Adversarial Network (SAN) architecture for automatic speech recognition. To the best of our knowledge, this is the first architecture that specifically focuses on the difficult problem of recognizing fuzzy audio in an acoustic model. The SAN architecture can capture key acoustic features and helps the model achieve better performance when faced with fuzzy audio inputs. Experiments on multiple datasets show that our SAN architecture works well and achieves competitive result on the AISHELL-1 dataset, 4.37 CER without language model, leading an around 7.6% relative CER reduction compared to the wenet baseline. We hope that our SAN architecture can make a difference in ASR tasks, especially in fuzzy speech recognition scenes.

7. REFERENCES

[1] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, et al., “Espnet: End-to-end speech processing toolkit,” arXiv preprint arXiv:1804.00015, 2018.
[2] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al., “Conformer: Convolution-augmented transformer for speech recognition,” arXiv preprint arXiv:2005.08100, 2020.

[3] Zhuoyuan Yao, Di Wu, Xiong Wang, Binbin Zhang, Fan Yu, Chao Yang, Zhendong Peng, Xiaoyu Chen, Lei Xie, and Xin Lei, “Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit,” arXiv preprint arXiv:2102.01547, 2021.

[4] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” Advances in Neural Information Processing Systems, vol. 33, pp. 12449–12460, 2020.

[5] Anjuli Kannan, Yonghui Wu, Patrick Nguyen, Tara N Sainath, Zhijeng Chen, and Rohit Prabhavalkar, “An analysis of incorporating an external language model into a sequence-to-sequence model,” in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 1–5.

[6] Wang Weiran, Tongzhou Chen, Tara Sainath, Ehsan Varjiani, Rohit Prabhavalkar, W. Ronny Huang, Bhuvana Ramabhadran, Neeraj Gaur, Sepand Mavandadi, Cal Peysar, Trevor Strohman, Yanzhang He, and David Rybach, “Improving Rare Word Recognition with LM-aware MWER Training,” in Proc. Interspeech 2022, 2022, pp. 1031–1035.

[7] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky, “Domain-adversarial training of neural networks,” The journal of machine learning research, vol. 17, no. 1, pp. 2096–2030, 2016.

[8] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in Proceedings of the 23rd international conference on Machine learning, 2006, pp. 369–376.

[9] William Chan, Navdeep Jaitly, Quoc V Le, and Oriol Vinyals, “Listen, attend and spell,” arXiv preprint arXiv:1508.01211, 2015.

[10] Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, “On using monolingual corpora in neural machine translation,” arXiv preprint arXiv:1503.03535, 2015.