MULTI-MODAL PRE-TRAINING FOR AUTOMATED SPEECH RECOGNITION

David M. Chan[⋆†1] *Shalini Ghosh*[†] *Debmalya Chakrabarty*[†] *Björn Hoffmeister*[†]

[⋆] University of California, Berkeley
[†] Amazon Alexa AI

ABSTRACT

Traditionally, research in automated speech recognition has focused on local-first encoding of audio representations to predict the spoken phonemes in an utterance. Unfortunately, approaches relying on such hyper-local information tend to be vulnerable to both local-level corruption (such as audio-frame drops, or loud noises) and global-level noise (such as environmental noise, or background noise) that has not been seen during training. In this work, we introduce a novel approach that leverages a self-supervised learning technique based on masked language modeling to compute a global, multi-modal encoding of the environment in which the utterance occurs. We then use a new deep-fusion framework to integrate this global context into a traditional ASR method, and demonstrate that the resulting method can outperform baseline methods by up to 7% on Librispeech; gains on internal datasets range from 6% (on larger models) to 45% (on smaller models).

***Index Terms*—** Automated Speech Recognition, Multi-Modal Learning, BERT, Conformer, Video

1. INTRODUCTION

Despite considerable research, automated speech recognition (ASR) remains an extremely challenging task, especially in noisy environments. Correctly understanding spoken phonemes requires an understanding of speech patterns, as well as an understanding of myriad varieties of background noise, much of which may never have been encountered by a model during the training process. Many traditional ASR methods such as the RNN-T and Conformer [1, 2, 3] focus on a local understanding of phonemes predicted from small 10-30ms segments on audio. Unfortunately, such local-first representations may leave ASR models vulnerable to extreme noise such as frame drops or sudden loud noises at the local level, as demonstrated by Chiu et al. [4]. Not only are such models vulnerable to local disruptions, they can be affected as well by global-level noise unseen during training.

In this paper, we target the problem of such global-level noise in utterances. Many ASR datasets such as Librispeech [5] are collected in lab-specific environments with a canonical set of situations, which leaves a long tailed distribution of noisy environments uncovered. While local level disruptions can in part be solved by introducing semantic-level language modeling [3, 6], the global out-of-distribution noise problem has no such simple solution. While self-supervised learning has been used in the student/teacher context in ASR [7, 8, 9], we speculate (and believe that it is important future work to confirm) that the additional performance gained by student teachers is a direct response to exposure to a large tail of environmental effects, regardless of the ASR content.

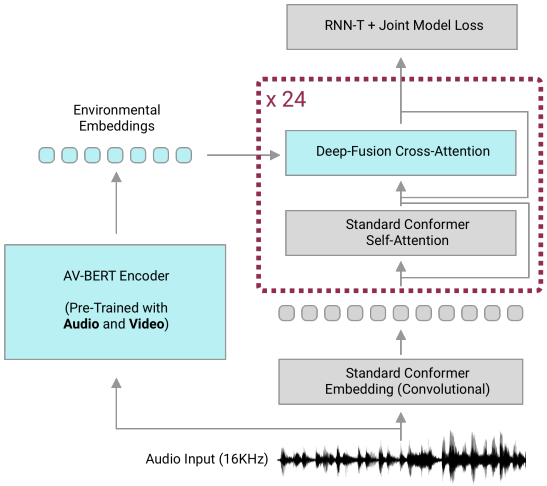

Fig. 1. An overview of our proposed approach to the ASR training process using deep-fusion with environmental embeddings. Our audio is fed to the pre-trained environmental representation model, trained on large-scale multimodal data. We then use a stack of 24 deep-fusion cross-attention layers in the base conformer architecture to deeply fuse the environmental representations with a standard conformer model. The RNN-T and joint model loss remain unchanged from [3].

Recently, both vision and NLP communities have introduced several methods [10, 11, 12] which make use of large unlabeled data to build self-supervised representations transferable to downstream tasks. Such representations can both provide exposure to the long-tailed data distribution and reduce the required labeled training data in transfer [12, 13]. Notably, Hsu et. al. [13] recently introduced HuBERT, which shows that general representations of audio are beneficial for ASR tasks.

Our proposed method goes beyond HuBERT, and expands

the context of environment-level global representation to both audio **and** visual data. We hypothesize that such visual data provides pseudo-labels, which help to weakly annotate auditory features, allowing the model to capture robust information in the long-tail. Recently [14] confirmed that even in the absence of the video representation at test time, audio-video self-supervised model representations can outperform audio-only models on *audio only* tasks. Concurrently with this work, [15] modified HuBERT for multi-modal downstream tasks but did not focus on the potential applications to ASR.

In designing our proposed method, we exploit the ability of visual information to organize audio representations, as demonstrated by Wang et. al. [14]. However, unlike the model proposed by Wang et al., our method leverages masked language modeling (MLM) as a joint training objective, as opposed to contrastive representation learning. In contrastive representation learning, samples from the audio and video domains are pushed into a joint latent space at a global level. This mode of training, however, *inherently suggests that the modalities should lie in the same latent space*, which we believe to be sub-optimal for automated speech recognition, where we want to focus on globally aware local-first representations (one phoneme should retain its' independence of other phonemes, but the representation should still be contextually aware). On the other hand, MLM, popularized in BERT [11] and further extended to multiple modalities in VideoBERT [12] and UniT [16], focuses on contextualizing local representations with the ability to reconstruct the full sequence, leading to local-first global aware representations.

In an effort to address the global representation problem, we (1) introduce a multi-modal pre-training scheme (AV-BERT) based on masked language modeling (Section 2.1), (2) develop a novel deep-fusion scheme for training on joint local/global representations in the ASR domain (Section 2.2) and (3) demonstrate the benefits of generating robust environmental representations with visual information, even when no visual information is present at test time (Section 3).

2. APPROACH

Our method consists of a two-stage approach (an overview is given in Figure 1) inspired by ideas from [12], [13] and [6]. In the first stage we build a video-augmented audio representation using a pre-training task based on masked language modeling. In the second stage, we use these video-augmented audio representations to provide additional context to a conformer-based ASR model.

2.1. Multimodal Pre-Training

While many methods for learning multimodal representations focus on self-supervised learning with a contrastive objective, our proposed method AV-BERT differs in that it uses a masked language modeling objective. Our pre-training encoder model, shown in Figure 2, takes inspiration from the UniT [16] model for unified transformer architectures, however instead of using multiple encoders, we use a single unified encoder layer. We take further inspiration from ViViT [17] and use a video-patch based encoding to the transformer, while taking inspiration from HuBERT's [9] iterated quantization training method for masked-language modeling from raw signals. In this section, we dive deeper into each of the components, and discuss our modeling choices from the perspective of an ASR-first multi-modal pre-training model.

To build a multimodal representation learning method based on masked language modeling principles, we first consider a token-based representation of modalities. We draw the representation from a discrete quantization of both the video and audio domains.

For the video modality, we extract non-overlapping voxels of shape $3 \times 16 \times 16$ of the input data, and 3-frame patches of audio data. These patches are then quantized using a K-means model with 8192 video centers, and 4096 audio centers trained offline. While we could use the quantized tokens directly as input to the masked language model as was done in VideoBERT [12], similar to HuBERT [13] we use the raw data as input while classifying based on the quantization. This allows the model to see the full input audio/video, and also allows the model to respond to subtle changes in the input which cannot be captured by our audio-visual language (which only consists of a total of 12288 audio/video tokens).

Thus, as input to our model, we use a set of modality-specific convolutions (matching the patch dimensions) to embed both the video and audio in the model dimension. We then apply a learned modality embedding, as well as learned position embedding [11]. For the audio, we use the frame index as the position. For the video, we apply a spatio-temporal position embedding identical to that from Timesformer [18]. We then flatten the spatial dimensions, and concatenate the video and audio sequences along the temporal axis, to form the input to the multimodal BERT-style encoder.

To perform masked language modeling, we use an architecture similar to BERT [11], which allows for full cross-attention between all points (both in the audio and video modalities, as well as spatiotemporally). This can lead to very long sequence lengths, so we compensate by reducing the per-node batch size and distributing the model across several GPUs. Because of the distribution across many GPUs, we do not use batch normalization — we instead use instance normalization to enhance training efficiency, since distributed batch normalization requires cross-GPU communication and is under-defined for small per-node batch sizes.

The training of the AV-BERT model is heavily dependent on the choice of masking technique. If we mask tokens uniformly with some rate, it is unlikely that the model will learn cross-model representations, as both audio and video are highly local representations (information in one location tends to share a large amount of mutual information with other neigh-

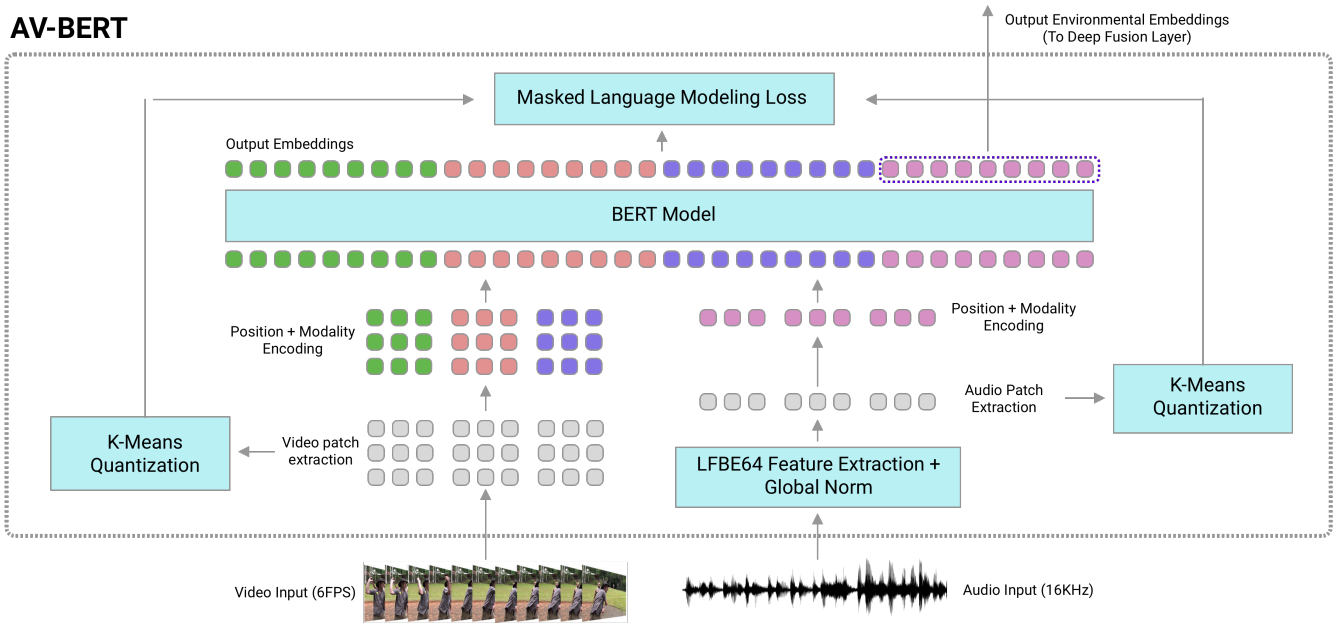

Fig. 2. An overview of our pre-training model. First, a set of patches are extracted from the multimodal inputs. Next, these patches are quantized using k-means and embedded directly using convolutional layers, modality encodings, and positional encodings. The embedded patches form the input sequence, which is passed to a standard BERT masked-language model. The quantized token labels are used along with the output of the masked BERT model to perform masked-language prediction.

bors). A naive approach would be to mask entire modalities at a time, however often the modalities are not heavily correlated enough to reconstruct a quantized representation of the other. Instead, we apply a progressive masking technique along the sequence dimension, where we begin the training by masking local information to encourage local-level representations, and progressively increase the size of the masks during training. This encourages the model to first learn local representations, and then eventually learn more global representations as the training process continues. Explicitly, we initialize the masking with a random probability of 0.15 and a mask width of 1. As training progresses, we increase the mask width and probability, ending with a final mask width of 11 and a mask probability at any center of 0.45. The probability is increased on a exponential decay schedule over 10,000 steps — every time we reach 10,000 optimization steps, the mask width is increased and the masking probability is reset.

We perform our pre-training using the publicly available splits of the Kinetics-600 dataset [19]. The Kinetics-600 dataset consists of 366K 10 second videos, each with a corresponding audio track and an associated action label (from 600 classes). For video, we reduce the frame-rate to 6FPS and resize the short side of the video to 256 pixels, and take a 256x256 center crop of the resulting resized video. For the audio, we resample the raw input audio to 16KHz, and stack 3 adjacent Log-Filterbank Energy features of dimension 64 (LFBE-64) from the resulting audio frames. The features are whitened using the mean and standard deviation computed globally from the training data, then clipped to lie between $(-1.2, 1.2)$ for numerical stability.

In our pre-training experiments, we use a model dimension of 128 with six BERT encoder blocks. The model is implemented in Tensorflow and is trained using 32 Nvidia-V100 GPUs, each with a batch size of 8, for 100 epochs (or until the validation perplexity converges, whichever comes first). To perform the optimization, we use an Adam [20] optimizer with a fixed learning rate of $3e^{-4}$, and $\beta_1 = 0.9, \beta_2 = 0.99$.

2.2. Automated Speech Recognition Downstream Task

For the downstream automated speech recognition task, we have two goals: (a) maintain the performance of current state of the art machine learning techniques, and (b) augment the current models with additional global-level environment context to improve phoneme recognition. In order to accomplish these goals, we modify the conformer architecture (as shown in Figure 1) to include additional cross-attention layers, which attend across the vector-level representations generated by our pre-trained AV-BERT model. This method allows the model to selectively pay attention to global context information learned by our model, while preserving the local-first approach favored by ASR techniques. This helps resolve one of the major challenges faced by HuBERT [13] in the ASR domain: when you focus on learning global representations, you can fail to encode the information necessary for local-first tasks such as

Reference:	should i buy from the princess starfrost set royale high
Base (M):	should i buy from the princess stare froset in we're all rawhide
Ours (M):	should i buy from the princess star frost set royale high
Reference:	read all of lisa left eye lopes songs including the thirteen more
Base (M):	read all of lisa **** *** loeb songs including the thirteen horn
Ours (M):	read all of lisa left eye lopez songs including the thirteen more
Reference:	... signalman he lead tenor for telephone wires so soldiers ...
Base (M):	... signal map he'd late tenoff telephone wise ** soldiers ...
Ours (M):	... signalman he lead teno for telephone wires so soldiers...

Fig. 3. Examples showing improvements on long utterances in our model, vs the baseline model. Blue indicates deletions, pink indicates substitutions and yellow indicates insertions.

phoneme detection. During the training of the downstream model, we freeze the representations learned by AV-BERT, to both reduce the computational complexity and to maintain a more global representation level even after significant training. Note that AV-BERT does not introduce additional trainable parameters, as AV-BERT is trained offline and only the audio representations are leveraged during the ASR training.

We evaluate the proposed model on the LibriSpeech [5] dataset, which consists of 970 hours of labeled speech. Because our audio embedding method is frozen during the training process, to ensure that there is no domain shift we follow the same audio pre-processing technique as in Section 2.1 with additional SpecAugment [21]. In addition to Librispeech, we present results on several internal datasets: "Base" representing general speech, "Query", representing standard speech queries, "Rare" representing the long-tailed distribution of rare words, and "Messages" representing longer message-based utterances. All customer-specific data has been de-identified from these internal datasets. For these results, we use the same model architecture; however once pre-training of AV-BERT is complete, we fine-tune our final ASR model on a corpus consisting of 120K hours of labeled speech and 180K hours of unsupervised speech, using in-house teacher distillation.

For the ASR Conformer, we use a hidden dimension of 1024 and 24 self-attention/cross-attention blocks, with convolutional downsampling on the input waveform with a kernel size of 3 and a stride of 2. We use an identical joint model to Gulati et al. [3] with a graph-based ASR decoding framework. The optimization hyper-parameters are shared with AV-BERT and described in Section 2.1, with the exception that we use a per-node batch size of 28.

3. RESULTS

Our main results are presented in Table 1. We report results on two models, a model using AV-BERT trained with both the video and audio components of the Kinetics dataset, as well as a model trained with only the audio from Kinetics. In Table 1, the Baseline model is identical to the proposed model except all multi-modal cross-attention layers are replaced with self-attention layers to preserve the parameter count.

Table 1. Results summary of word error rate for the Librispeech dataset with no additional language model. The baseline model replaces the cross-attentions with self-attention (to closely preserve parameters) using the same training profile (See Section 2). "A" is the audio-only model, and "A/V" is the full Audio/Video BERT.

Method	Params (M)	test-clean	test-other
LAS			
Transformer [22]	370	2.89	6.98
Transformer [23]	-	2.2	5.6
LSTM [3]	360	2.6	6.0
Transducer			
Transformer [2]	139	2.4	5.6
ContextNet (M) [24]	31.4	2.4	5.4
ContextNet (L) [24]	112.7	2.1	4.6
Conformer			
Conformer (M) [3]	30.7	2.3	5.0
Conformer (L) [3]	118.8	2.1	4.3
Ours			
Conf. (M, base)	79	2.21	4.85
Conf. (L, base)	122	2.11	4.29
A + Conf. (M)	79	2.15 **(+2.7%)**	4.82 **(+0.6%)**
A/V + Conf. (M)	79	2.10 **(+4.8%)**	4.72 **(+2.7%)**
A/V + Conf. (L)	122	1.98 **(+7.0%)**	4.10 **(+4.4%)**

Table 2. Relative improvement over baseline WER for Alexa-AI datasets methods without a language model.

Method	Base	Rare	Query	Messages
Conformer (M)	0	0	0	0
+ Audio (M)	+30.1%	+17.9%	+26.7%	+20.1%
+ Audio/Video (M)	+45.6%	+31.2%	+38.7%	+17.2%
Conformer (L)	0	0	0	0
+ Audio/Video (L)	+5.1%	+5.4%	+4.2%	+5.9%

The results show that both models — one trained on audio only and another with audio + video embeddings — outperform the baseline model. We further validate the method with experiments on internal Alexa AI datasets in Table 2.Our fine-tuned model is warm-started from the base Alexa model (300K hours of audio), and augmented with the additional training on the 377K samples. The benefit is more pronounced in smaller models, as the contextual representations are more useful with fewer training data, and parameters.

4. CONCLUSION

In this paper, we have introduced an initial approach for exploring global-level contextual embeddings for the automated speech recognition pipeline. We build a novel self-supervised vision + audio encoder, and demonstrate the performance of this method by using deep-fusion to directly connect the contextual embeddings with the local-first conformer model. Our model demonstrates strong performance on Librispeech, and presents a new direction for exploration into multimodal ASR.

5. REFERENCES

[1] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*. Ieee, 2013, pp. 6645–6649.

[2] Qian Zhang, Han Lu, Hasim Sak, Anshuman Tripathi, Erik McDermott, Stephen Koo, and Shankar Kumar, "Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7829–7833.

[3] 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.

[4] Chung-Cheng Chiu, Arun Narayanan, Wei Han, Rohit Prabhavalkar, Yu Zhang, Navdeep Jaitly, Ruoming Pang, Tara N Sainath, Patrick Nguyen, Liangliang Cao, et al., "Rnn-t models fail to generalize to out-of-domain audio: Causes and solutions," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 873–880.

[5] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[6] Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelovic, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman, "Self-supervised multimodal versatile networks.," *NeurIPS*, vol. 2, no. 6, pp. 7, 2020.

[7] Shinji Watanabe, Takaaki Hori, Jonathan Le Roux, and John R Hershey, "Student-teacher network learning with enhanced features," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5275–5279.

[8] Zi-qiang Zhang, Yan Song, Jian-shu Zhang, Ian Vince McLoughlin, and Li-rong Dai, "Semi-supervised end-to-end asr via teacher-student learning with conditional posterior distribution.," in *INTERSPEECH*, 2020, pp. 3580–3584.

[9] Vimal Manohar, Pegah Ghahremani, Daniel Povey, and Sanjeev Khudanpur, "A teacher-student learning approach for unsupervised domain adaptation of sequence-trained asr models," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 250–257.

[10] Hassan Akbari, Linagzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong, "Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text," *arXiv preprint arXiv:2104.11178*, 2021.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[12] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid, "Videobert: A joint model for video and language representation learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7464–7473.

[13] Wei-Ning Hsu, Yao-Hung Hubert Tsai, Benjamin Bolte, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "Hubert: How much can a bad teacher benefit asr pre-training?," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6533–6537.

[14] Luyu Wang, Pauline Luc, Adria Recasens, Jean-Baptiste Alayrac, and Aaron van den Oord, "Multimodal self-supervised learning of general audio representations," *arXiv preprint arXiv:2104.12807*, 2021.

[15] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," *arXiv preprint arXiv:2201.02184*, 2022.

[16] Ronghang Hu and Amanpreet Singh, "Unit: Multimodal multitask learning with a unified transformer," *arXiv preprint arXiv:2102.10772*, 2021.

[17] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid, "Vivit: A video vision transformer," *arXiv preprint arXiv:2103.15691*, 2021.

[18] Gedas Bertasius, Heng Wang, and Lorenzo Torresani, "Is space-time attention all you need for video understanding?," *arXiv preprint arXiv:2102.05095*, 2021.

[19] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman, "A short note about kinetics-600," *arXiv preprint arXiv:1808.01340*, 2018.

[20] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[21] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[22] Gabriel Synnaeve, Qiantong Xu, Jacob Kahn, Tatiana Likhomanenko, Edouard Grave, Vineel Pratap, Anuroop Sriram, Vitaliy Liptchinsky, and Ronan Collobert, "End-to-end asr: from supervised to semi-supervised learning with modern architectures," *arXiv preprint arXiv:1911.08460*, 2019.

[23] Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, Ryuichi Yamamoto, Xiaofei Wang, et al., "A comparative study on transformer vs rnn in speech applications," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 449–456.

[24] Wei Han, Zhengdong Zhang, Yu Zhang, Jiahui Yu, Chung-Cheng Chiu, James Qin, Anmol Gulati, Ruoming Pang, and Yonghui Wu, "Contextnet: Improving convolutional neural networks for automatic speech recognition with global context," *arXiv preprint arXiv:2005.03191*, 2020.