Training Deep Neural Network Models for Time Series Sensory Data v4

Lijiang Guo
Department of Intelligent Systems Engineering
Indiana University, Bloomington, IN 47405
lijguo@indiana.edu

Time-series sensory data is one of the most common types of data used by human. In this thesis research, we investigate Deep Neural Network (DNN) based machine learning algorithms to advance temporal prediction from human sensory data, for example speech and vision. In the first half we focus on supervised learning under multiple time-series sensory inputs. In the second half, we investigate methods for improving weakly-supervised single time-series model training from unaligned sequences and task loss.

Combining Multimodal Time-Series

Human senses the world with multiple modalities such as vision, sound, touch, etc. An AI agent, e.g. a robot, also relies on multiple sensors to collect data to perform tasks such as localization, path planning, control a robotic arm, etc. Baltrušaitis et al. [2018] refers a *sensory modality* as one of our primary channels of communication and sensation, such as vision. A problem is characterized as *multimodal* when it includes multiple sensory modalities. Different from classical statistics, in machine learning we often face very high dimensional input data, and each input modality has essentially different representation forms; for example, audio and vision. The challenging question in creating intelligent systems for processing multimodal sensory inputs is how to combine different input modalities when there are no straight forward physical model to describe them.

Multimodal sensor fusion algorithms can be divided into two big categories [Baltrušaitis et al., 2018]: *model-agnostic* and *model-based*. Model-agnostic sensor fusion does not depend on a specific machine learning method. It usually combines different sensors in a blind way and the relationship between sensors are not explicitly exploited. Model-based methods explicitly address fusion in their construction, e.g. using graphical models to impose a structural prior distribution on signals from different modalities. In this work, we focus on model-based approach for sensor fusion.

The most common fusion approach is to concatenate the inputs from different sensors at certain modeling stage to solve a unimodal learning problem. Ngiam et al. [2011] proposed a multimodal autoencoder which first use deep Boltzmann machine (DBM) to learn each audio and video features independently, then concatenate the learned features to an multimodal autoencoder to learn a shared representation. To reproduce original inputs, the shared representation is mapped back to each modality. Srivastava and Salakhutdinov [2012] also use DBM to learn a generative model of multimodal input data. Simonyan and Zisserman [2014] proposed a two-steam convolutional neural networks for action recognition in video. The two-stream model try to capture the complementary information from two modalities: still image and motion between frames. In their work, the last layer concatenates two streams into a final classification layer. Torfi et al. [2017] proposed a coupled 3D-CNN to map multiple modalities into a representation space to evaluate the correspondence of audio-visual streams. In Ephrat et al. [2018] the audio and visual streams are combined by concatenating the feature maps of each stream, which are subsequently fed into a BLSTM followed by three FC layers.

Dynamically combining multiple functions in supervised learning can be traced back to *mixture of experts* model [Jacobs et al., 1991]. A mixture of experts model is driven by the assumption that a set of training cases may be naturally divided into subsets that correspond to distinct tasks. However, the interference between different subsets of tasks would lead to slow learning and poor generalization. Such interference can be reduced by using a system composed of several different *expert functions* where each one is trained for a subset of tasks. A *gating function* is trained to decide which of the experts should be used for each training case. Instead of a hard decision such as decision tree, the gating function makes probablistics decision by assigning mixing weights to the experts. Neural networks can be used for both the expert functions and gating function, which is known as *mixture density network* (MDN) [Bishop, 1994]. One advantage of MDN is it can approximate a flexible family of distributions, including distribution of multiple modes.

Motivated by these works, we propose a spatial attention model for sensor fusion. The temporal attention mechanism [Bahdanau et al., 2014] can dynamically generate a probability distribution over a time segment. This probability distribution can be interpreted as a measure of the importance of a past time point to current prediction. From this interpretation, we propose a spatial attention which dynamically generates a probability distribution over input sensors at each time step. This spatial attention allows the system to dynamically combine information from different sensors in a sequential decision making task. To maximize utility of each type of input information each sensor is bonded with a modular neural network. Another gating modular neural network is dedicated to dynamically generate the spatial attention which is used as a set of mixing weights for outputs from sensor networks to balance utility of all sensors' information. We also experiment with a co-learning

mechanism to encourage co-adaption and independent learning of each sensor at the same time, and propose a regularization based co-learning method. We demonstrate the proposed method in an audio-visual speech activity detection task.

We next propose a probabilistic model which combines multiple sensor signals into separate modality-dependent and modality-invariant features. The core technique is a variational RNN (VRNN) [Chung et al., 2015] model where the transition between hidden units are stochastic. The VRNN model is a non-Markovian model in the sense the conditional independence assumption is broken in both transition and emission. We discuss this property by comparing VRNN with hidden Markov model and linear dynamic system. Based on VRNN, we propose a multimodal VRNN (MVRNN) model which imposes a structural prior on the generative model. As a result, the modality dependent and modality invariant factors are encouraged to separate into different latent variables. We emphasis that as oppose to PCA [Bishop et al., 2006] or VAE [Kingma and Welling, 2013, Rezende et al., 2014] where latent factors are not known a priori without looking at the posterior distribution, our model explicitly matches latent variables with concepts. This is a result of the explicit graphical model structure.

Summary

In this research we are motivated by the question of how to combine multiple modalities of sensory data in an explainable and controllable way. In doing so, sensory inputs can complement each other to gather information in a complicated environment. Such ability is valuable to the development of an intelligent autonomous system. We propose a spatial attention model that can dynamically combining sensor inputs in a sequential decision making task, and propose a multimodal variational RNN model to separate signal from noise. Our experiment results show the recurrent attention filter method can reduce the misclassification error by 51.3% and 57.8% when the audio is corrupted with non-human voice noise and human-voice noises respectively, while the video is corrupted, compared to a video or audio only system. We also find the model can detect when an input signal is corrupted and reduce its attention weight, while increase attention weights of more reliable inputs. This shows the improvement of decision accuracy attributes to the attention mechanism. To our knowledge, this algorithm is the first to use attention mechanism for streaming decision making with multiple inputs streams.

Learning from Unaligned Input-Output Sequences

We next focus on single time-series sensory data model. One major challenge of time-series sensory data is that the input and output sequences differ in lengths and both lengths are variable. For example

in speech recognition, the speech audio input and speech transcription output are of different lengths. Supervised learning of time-series model usually requires temporally aligned input and output. Taking speech recognition as an example, supervised learning of RNN based speech recognition algorithm requires speech transcriptions are given at each time step of speech audio [Graves et al., 2012]. Such temporally aligned data is often hard to prepare, and it is well recognized that DNN benefits from training with large dataset; in some situations, hand-prepared alignment data contain human errors, or there could simply is no single alignment between input and output sequences. Hence it is desirable to have algorithms that only require input and output sequences are *monotonically related* but *not temporally aligned*.

Many fields are interested in time-series models for unaligned data: DNA sequencing [Teng et al., 2018], machine translation [Yu et al., 2016], speech recognition and keyword detection [Lengerich and Hannun, 2016], and action recognition from video [Huang et al., 2016][Assael et al., 2016]. We are interested in one type of time-series model called *transducer* model which can learn to transcribe an input sequence to an output sequence from monotonically indexed unaligned examples. The most successful pre-deep learning, statistical speech recognition model — Hidden Markov Model (HMM) is a kind of transducer model which has thrived on the development of Baum-Welch algorithm [Rabiner, 1989]. Many recent efforts have been made to marry the strength of DNN with transducer models, including the Connectionist Temporal Classification (CTC) [Graves et al., 2006], RNN Transducer (RNN-T) [Graves, 2012], neural transducer [Jaitly et al., 2016], and Recurrent Neural Aligner (RNA) [Sak et al., 2017]. These transducer models generate *hard monotonic alignments* between input and output sequences, and use dynamic programming to efficiently summarizes all possible alignments. When evaluated from multiple criteria, RNN-T has been found to be favored over the rest in terms of superior prediction accuracy, ease of decoding, and simpler unified structure [Battenberg et al., 2017][Prabhavalkar et al., 2017][Chen et al., 2020].

The basic building block of RNN-T for modeling temporal relation is RNN. While RNN has been the long time standard DNN model for temporal data, it is not efficient in memorizing long-time-span relations [Bengio et al., 1994] and not friendly to parallel computation due to temporal dependency in hidden states. The recently proposed *Attention-based Encoder-Decoder* (AED) models have been shown to successfully overcome these issues for sequence to sequence prediction [Chorowski et al., 2015] [Bahdanau et al., 2014]. AED models use attention mechanism to compute *soft non-monotonic alignment* between each output step and every input step. Attention mechanism allows direct interaction between any two time steps in a sequence despite their distance in time. Early AED models were built on top of RNN models. The more recent Transformer model with self-attention [Vaswani et al., 2017] completely eliminated RNN and only use feed-forward layers, which allows for parallel computation during training. Self-attention allows neural network to learn

flexible features with long-contextual-span properties, which are critical to the success of structural discriminative methods. Transformer model features multiple encoder-decoder attention layers and multihead self-attention at each layer. Recent works have found Transformer outperform RNN in many sequence-to-sequence tasks such as machine translation [Vaswani et al., 2017] and speech recognition [Karita et al., 2019].

A sequence prediction model is capable of *streaming prediction* if the model can generate outputs while receiving input signals. A model is *non-streaming* if it can generate outputs only after receiving the entire input signal. Because the attention mechanism maps every input step to each output step, AED models naturally are not capable of streaming prediction. The best configuration of AED model for streaming prediction is still unknown and has been a hot research area. Some works have explored *monotonic attention* variants for time-series data. Hard Monotonic Attention (HMA) [Raffel et al., 2017] uses hard monotonic sampling over the input sequence to generate a hard monotonic alignment between input and output sequences. But each output time step can only attend one input time step due to hard alignment. To relax the strict hard alignment, Monotonic Chunkwise Attention (MoChA) [Chiu and Raffel, 2017] uses soft attention[1] over a small fixed-size chunk of input time steps for each output time step, where the location of each chunk is selected by hard monotonic attention. Monotonic Infinite Look-back Attention (MILk) [Arivazhagan et al., 2019] further extends the fixed-size attention window to a soft attention over all previous input time steps. These monotonic attention approaches also provide a closed form expression for the expected alignment between source and target tokens, and avoid unstable reinforcement learning, which trades computation for stability. However, chunkwise methods ignore the relationship between different chunks.

In contrast to AED model, transducer models are streamable and have low runtime computation cost. Moritz et al. [2019] tried to combine transducer model with attention in the Triggered Attention (TA) model for streaming prediction. TA model utilizes the alignment capability of transducer to generate monotonically growing attention regions. The decoder then oversees each attention region to generate an output. The trigger network is essentially a CTC decoder, which provides a one-best alignment between input and output sequences. There are a few weaknesses with TA model. First the encoder is a compromise for two tasks, i.e. alignment and attention. Second, because of the hybrid loss function in TA model, the hybrid decoding algorithm is complicated and requires hyper-parameter tuning. It is more desirable to have a unified loss such as RNN-T. In addition, due to using CTC, TA model does not model the structure of output sequence.

The integration of Convolutional Neural Network (CNN) and Transformer is another promising direction. While Transformers are good at modeling long-range global context, they are less capable of extracting fine-grained local feature patterns. CNN can capture local context progressively via a

[1] Xu et al. [2015] has an excellent discussion on the difference between soft and hard attention.

local receptive field layer by layer. Recent works have shown that CNN is also capable of capturing temporal relations for sequence prediction [Gehring et al., 2017]. Collobert et al. [2016] used convolutional layer to extract feature from raw audio signals. Mohamed et al. [2019] showed convolutional layers not only helps Transformer to extract local temporal features, but also provides a new way for relative positional encoding in addition to the distance-based relative position encoding in Transformer-XL [Dai et al., 2019].

Summary

In this work, we study the unification of CNN, Transformer and RNN-T into a single streaming prediction model. We combine the strength of CNN for local feature extraction with Transformer for capturing long-context-span relation, and RNN-T for streaming prediction. We explore effective methods to modify self-attention for streaming prediction in a RNN-T framework such as time-restricted self-attention [Moritz et al., 2020] and use convolutional layers for relative positional encoding for Transformer. Our experiments have shown encouraging results. Comparing to RNN-T, the CNN-Transformer based model reduces word-error-rate (WER) from 11.50% to 8.36%. This suggests CNN-Transformer is a viable replacement to RNN in RNN-T. Interestingly, we find the performance of CNN-Transformer based model is more sensitive to the model size than RNN based model. When we reduce model parameter from 53 million to 15 million, WER increases from 8.36% to 11.79%, about 41% relative increase. Meanwhile, RNN based model increases from 11.50% to 12.11%, about 5.3% relative increase. This suggests self-attention mechanism needs more layers and attention heads to maximize its potential. Integration of Transformer and RNN-T has drawn many interests from industry. We notice related works [Zhang et al., 2020] [Tripathi et al., 2020] [Huang et al., 2020] were concurrently developed with this work. We plan to continue investigate the properties of CNN-Transformer-Transducer model such as prediction latency.

Minimizing Task Loss

While transducer and AED models have lifted the requirement for alignment between input and output sequences, they continue to use likelihood based objectives as surrogates to task performance criteria such as word-error-rate (WER) for speech recognition [Kaiser et al., 2002] or BLEU score for machine translation [Papineni et al., 2002]. This is due to these evaluation criteria are often non-decomposable over time and are non-differentiable functions, thus are difficult to optimize by gradient based methods in batches. In addition, because of the difficulty in learning long-context-span structure in time series, it has been a standard practice in RNN model to adopt *teacher-forcing* [Bengio et al., 1994] to feed previous prediction targets rather than predictions to the model and update the model parameters in direction towards the next target. However, while teacher-forcing can

find sequence-level optimal for data in the training set, the model never learn what to do once it has made an error in an intermediate time step. These compromises impose inherent limit on a model's prediction accuracy.

Task loss minimization has been of great interest to speech recognition. Previous works have relied on a second stage optimization over an objective that is more closely correlated with task loss to fine tune a pre-trained model [Gibson, 2008]. The second objective is an expected task loss that measures the difference between the target output sequence and the predicted output sequence weighted by its probability. This expected task loss does not have a closed form solution and is expensive to compute by enumerating the sampling space, thus is approximated in different ways such as beam search. It also has been found that learning a classification function by direct minimization of the classification loss, e.g. *empirical risk minimization* (ERM) gives lower prediction error than learning a posterior probability based Bayes classifier [Gibson, 2008]. Early works in this area have adopted Extended Baum-Welch (EBW) algorithm to compute gradient of expected loss and fine tune a pre-trained HMM [Kaiser et al., 2002] or shallow neural network models [Veselỳ et al., 2013] for task loss. These practices were restricted to small dataset of short sequences according to today's standard for DNN models. Graves and Jaitly [2014] has shown that the CTC transducer model can be directly optimized towards expected task loss, using an efficient sampling and variance reduction strategy. However, this strategy is tied to CTC model structure and cannot be extended to other transducer models such as RNN-T. Recent works have shifted towards more general Monte Carlo sampling-based approaches which can be applied to any kinds of model structure and task loss [Shannon, 2017] [Prabhavalkar et al., 2018].

Time-series prediction as sequence transduction can be considered as a kind of Markov Decision Process (MDP). A close examination reveals methods in [Prabhavalkar et al., 2018] has the same formulation as a kind of Reinforcement Learning (RL) model i.e. Policy Gradient (PG) method. In RL, an agent uses a policy to interact with an environment to sequentially generate a path that maximizes expected reward. Hence each policy induces a probability distribution over the path space. The policy is gradually improved by a *Monte Carlo gradient estimate* from sampling from this induced probability distribution. This motivates us to approach sequence task loss minimization from a reinforcement learning perspective.

In this part of work, we investigate Monte Carlo gradient estimation method to directly optimize a time-series task performance criterion in the context of speech recognition. There are two major types of Monte Carlo stochastic gradient estimators: score function method and pathwise method [Fu, 2006][Mohamed et al., 2020]. Recent works have investigated RL methods for sequence prediction in many areas [Gu et al., 2016] [Li et al., 2016] [Bahdanau et al., 2017] [Zhou et al., 2018][Bello et al., 2016], but most are limited to the PG algorithm which uses the score function method to

estimate the gradient of the expected reward with respect to model parameters. Because score function method has to marginalize an unknown stochastic path by sampling, its gradient estimate is extremely variable. Hence it is critical to employ variance reduction techniques in score function method. We investigate several variance reduction methods for score function gradient estimator, including action-independent baseline, self-critic baseline, and action-dependent baseline. The pathwise gradient estimator is closely related to the Variational Autoencoder (VAE) model [Kingma and Welling, 2013, Rezende et al., 2014]. When a latent variable is continuous, pathwise method can utilize the reparameterization trick to sample from a constant distribution instead of the sampling distribution generated by the policy. This has been found to significantly reduce the variance of gradient estimate and accelerates convergence. When output space is discrete, reparameterization trick cannot be used [Kingma and Welling, 2013]. The Gumbel distribution [Gumbel, 1948] provides a smooth controllable approximation to a categorical distribution. We can hence approximate the discrete random variable with a continuous random variable of Gumbel-softmax distribution [Maddison et al., 2016, Jang et al., 2016]. This allows us to choose action in forward pass with argmax and train the model using fully differentiable backpropogation with reparameterization trick in backward pass. The price we pay is the bias introduced by continuous relaxation of discrete random variable in the gradient estimate.

Summary

To summarize, we investigated sampling-based methods for direct optimization of arbitrary task loss, e.g. WER minimization for speech recognition, and proposed a new pathwise gradient estimator for discrete variable. With aforementioned theoretical analysis, we find it is challenging to have expected performance of these models in experiments. We find both score function and pathwise gradient estimation methods are unstable and their convergences require huge amount of computation beyond we could afford. Training reinforcement learning model is known to be difficult [Henderson et al., 2018]. Our observations echo the questions shared by many recent researches. We discuss our limitations and lessons learned.

Remark

The task loss training can be performed parallel to or after transducer model training. Besides, there is a common theme among the methods we studied in this thesis: we use latent variables to describe the temporal structure within time-series data. We assume the observed data are represented by some underlying latent variables which usually lives on a much lower dimension space. Hence the variational RNN and reinforcement learning methods both use Monte Carlo sampling to approximate

the integration over latent variables, while the transducer model can use dynamic programming to compute the exact integration over the latent variables.

References

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011.

Nitish Srivastava and Ruslan R Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, pages 2222–2230, 2012.

Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

Amirsina Torfi, Seyed Mehdi Iranmanesh, Nasser Nasrabadi, and Jeremy Dawson. 3d convolutional neural networks for cross audio-visual matching recognition. *IEEE Access*, 5:22081–22091, 2017.

Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*, 2018.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

Christopher M Bishop. Mixture density networks. Technical report, Citeseer, 1994.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. A recurrent latent variable model for sequential data. In *Advances in neural information processing systems*, pages 2980–2988, 2015.

Christopher M Bishop et al. *Pattern recognition and machine learning*, volume 1. springer New York, 2006.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

Alex Graves et al. *Supervised sequence labelling with recurrent neural networks*, volume 385. Springer, 2012.

Haotian Teng, Minh Duc Cao, Michael B Hall, Tania Duarte, Sheng Wang, and Lachlan JM Coin. Chiron: translating nanopore raw signal directly into nucleotide sequence using deep learning. *GigaScience*, 7(5):giy037, 2018.

Lei Yu, Jan Buys, and Phil Blunsom. Online segment to segment neural transduction. *arXiv preprint arXiv:1609.08194*, 2016.

Chris Lengerich and Awni Hannun. An end-to-end architecture for keyword spotting and voice activity detection. *arXiv preprint arXiv:1611.09405*, 2016.

De-An Huang, Li Fei-Fei, and Juan Carlos Niebles. Connectionist temporal modeling for weakly supervised action labeling. In *European Conference on Computer Vision*, pages 137–153. Springer, 2016.

Yannis M Assael, Brendan Shillingford, Shimon Whiteson, and Nando de Freitas. Lipnet: Sentence-level lipreading. *arXiv preprint arXiv:1611.01599*, 2016.

Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

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*, pages 369–376. ACM, 2006.

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

Navdeep Jaitly, Quoc V Le, Oriol Vinyals, Ilya Sutskever, David Sussillo, and Samy Bengio. An online sequence-to-sequence model using partial conditioning. In *Advances in Neural Information Processing Systems*, pages 5067–5075, 2016.

Hasim Sak, Matt Shannon, Kanishka Rao, and Françoise Beaufays. Recurrent neural aligner: An encoder-decoder neural network model for sequence to sequence mapping. In *Interspeech*, pages 1298–1302, 2017.

Eric Battenberg, Jitong Chen, Rewon Child, Adam Coates, Yashesh Gaur Yi Li, Hairong Liu, Sanjeev Satheesh, Anuroop Sriram, and Zhenyao Zhu. Exploring neural transducers for end-to-end speech

recognition. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 206–213. IEEE, 2017.

Rohit Prabhavalkar, Kanishka Rao, Tara N Sainath, Bo Li, Leif Johnson, and Navdeep Jaitly. A comparison of sequence-to-sequence models for speech recognition. In *Interspeech*, pages 939–943, 2017.

Xie Chen, Yu Wu, Zhenghao Wang, Shujie Liu, and Jinyu Li. Developing real-time streaming transformer transducer for speech recognition on large-scale dataset. *arXiv preprint arXiv:2010.11395*, 2020.

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.

Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. In *Advances in neural information processing systems*, pages 577–585, 2015.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

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. *arXiv preprint arXiv:1909.06317*, 2019.

Colin Raffel, Minh-Thang Luong, Peter J Liu, Ron J Weiss, and Douglas Eck. Online and linear-time attention by enforcing monotonic alignments. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2837–2846. JMLR. org, 2017.

Chung-Cheng Chiu and Colin Raffel. Monotonic chunkwise attention. *arXiv preprint arXiv:1712.05382*, 2017.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015.

Naveen Arivazhagan, Colin Cherry, Wolfgang Macherey, Chung-Cheng Chiu, Semih Yavuz, Ruoming Pang, Wei Li, and Colin Raffel. Monotonic infinite lookback attention for simultaneous machine translation. *arXiv preprint arXiv:1906.05218*, 2019.

Niko Moritz, Takaaki Hori, and Jonathan Le Roux. Triggered attention for end-to-end speech recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5666–5670. IEEE, 2019.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. *arXiv preprint arXiv:1705.03122*, 2017.

Ronan Collobert, Christian Puhrsch, and Gabriel Synnaeve. Wav2letter: an end-to-end convnet-based speech recognition system. *arXiv preprint arXiv:1609.03193*, 2016.

Abdelrahman Mohamed, Dmytro Okhonko, and Luke Zettlemoyer. Transformers with convolutional context for asr. *arXiv preprint arXiv:1904.11660*, 2019.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.

Niko Moritz, Takaaki Hori, and Jonathan Le Roux. Streaming automatic speech recognition with the transformer model. *arXiv preprint arXiv:2001.02674*, 2020.

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)*, pages 7829–7833. IEEE, 2020.

Anshuman Tripathi, Jaeyoung Kim, Qian Zhang, Han Lu, and Hasim Sak. Transformer transducer: One model unifying streaming and non-streaming speech recognition. *arXiv preprint arXiv:2010.03192*, 2020.

Wenyong Huang, Wenchao Hu, Yu Ting Yeung, and Xiao Chen. Conv-transformer transducer: Low latency, low frame rate, streamable end-to-end speech recognition. *arXiv preprint arXiv:2008.05750*, 2020.

Janez Kaiser, Bogomir Horvat, and Zdravko Kačič. Overall risk criterion estimation of hidden markov model parameters. *Speech Communication*, 38(3-4):383–398, 2002.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

Matthew Gibson. *Minimum Bayes risk acoustic model estimation and adaptation*. PhD thesis, Citeseer, 2008.

Karel Veselỳ, Arnab Ghoshal, Lukás Burget, and Daniel Povey. Sequence-discriminative training of deep neural networks. In *Interspeech*, volume 2013, pages 2345–2349, 2013.

Alex Graves and Navdeep Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1764–1772, 2014.

Matt Shannon. Optimizing expected word error rate via sampling for speech recognition. *arXiv preprint arXiv:1706.02776*, 2017.

Rohit Prabhavalkar, Tara N Sainath, Yonghui Wu, Patrick Nguyen, Zhifeng Chen, Chung-Cheng Chiu, and Anjuli Kannan. Minimum word error rate training for attention-based sequence-to-sequence models. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4839–4843. IEEE, 2018.

Michael C Fu. Gradient estimation. *Handbooks in operations research and management science*, 13: 575–616, 2006.

Shakir Mohamed, Mihaela Rosca, Michael Figurnov, and Andriy Mnih. Monte carlo gradient estimation in machine learning. *Journal of Machine Learning Research*, 21(132):1–62, 2020.

Jiatao Gu, Graham Neubig, Kyunghyun Cho, and Victor OK Li. Learning to translate in real-time with neural machine translation. *arXiv preprint arXiv:1610.00388*, 2016.

Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*, 2016.

D Bahdanau, Philemon Brakel, R Lowe, J Pineau, K Xu, A Goyal, Aaron Courville, and Yoshua Bengio. An actor-critic algorithm for structured prediction. *Proc. of ICLR*, 2017.

Yingbo Zhou, Caiming Xiong, and Richard Socher. Improving end-to-end speech recognition with policy learning. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5819–5823. IEEE, 2018.

Irwan Bello, Hieu Pham, Quoc V Le, Mohammad Norouzi, and Samy Bengio. Neural combinatorial optimization with reinforcement learning. *arXiv preprint arXiv:1611.09940*, 2016.

Emil Julius Gumbel. *Statistical theory of extreme values and some practical applications: a series of lectures*, volume 33. US Government Printing Office, 1948.

Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.