LEARNING MONOTONIC ALIGNMENTS WITH SOURCE-AWARE GMM ATTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Transformers with soft attention have been widely adopted to various sequence-to-sequence tasks. While the soft attention is effective for learning semantic similarities between queries and keys based on their contents, it does not explicitly model the order of elements in sequences which are crucial for monotonic sequence-to-sequence tasks. Learning monotonic alignments between input and output sequences could be beneficial for long-form and online inference applications that are still challenging problem for the conventional soft attention algorithm. In this paper, we focus on monotonic sequence-to-sequence task and propose Source-Aware Gaussian Mixture Model (SAGMM) attention in which the attention scores are monotonically calculated considering both the content and order of the source sequence. We show that the proposed attention mechanism solves the online and long-form speech recognition problems without performance degradation in offline in-distribution speech recognition through experiments.

1 INTRODUCTION

For a recent few years, the transformer models with soft attention have been widely adopted to various sequence generation tasks (Raffel et al., 2019; Vaswani et al., 2017; Parmar et al., 2018; Karita et al., 2019). The soft attention does not explicitly model the order of elements in the sequence and attends all encoder outputs for each decoder step. However, the order of elements is crucial to understand monotonic sequence-to-sequence (Seq2Seq) tasks, e.g. automatic speech recognition (ASR), video analysis, and lip reading. Learning monotonic alignments helps model to attend subset of the encoder output without performance degradation in these tasks. In comparison, the soft attention is not suitable for streaming inference applications since the softmax operation needs to wait until all encoder outputs are produced. Furthermore, the soft attention has difficulty to generate long-form sequences which are not considered in training corpus.

The Gaussian Mixture Model (GMM) attention Graves (2013); Battenberg et al. (2020); Chiu et al. (2019) has been proposed to learn the monotonic mapping between encoder and decoder states for long-form sequence generation. The GMM attention is a pure location-aware algorithm in which the model selects attention windows cumulatively without considering source contents. This pure location-based property leads to mismatch between attention parameters from decoders and information distribution in encoder outputs.

In this paper, we adopt the GMM attention mechanism to modern transformer structure and propose Source-Aware Gaussian Mixture Model (SAGMM) attention which considers both contents and orders of source sequences. SAGMM is multi-modal attention that can also discard non-informative tokens in attention window. For online inference, we propose truncated SAGMM (SAGMM-tr) by discarding long-tail of the attention score in SAGMM. To the best of our knowledge, this is the first attempt to adopt the GMM-based attention to online sequence generation tasks. Learning accurate monotonic alignments also helps SAGMM-tr to attend subset of long sequences for each decoder step and improves the performance of model in long-form sequence generation tasks. The experiments on streaming and long-form ASR showed

substantial performance improvements compared to conventional algorithms without performance degradation on offline in-distribution ASR. We also tested SAGMM-tr in machine translation task and showed the performance of proposed algorithm in non-monotonic tasks.

2 SOURCE-AWARE GMM ATTENTION

2.1 SOFT ATTENTION

In this paper, we abbreviate the head index h during attention score calculation for simplicity. In the dot-product multi-head soft attention Vaswani et al. (2017) without relative positional encoding , the attention score from soft attention $Soft_\alpha$ is derived from the query matrix $Q \in \mathbb{R}^{I \times d}$ and key matrix $K \in \mathbb{R}^{J \times d}$ as follows:

$$Soft_\alpha = \text{softmax}\left(QK^{\intercal}/\sqrt{d}\right) \tag{1}$$

where d, I, and J are a feature dimension, decoder and encoder sequence length, respectively. The multi-head attention concatenates output vector from each head and apply a linear transform before a residual connection as

$$head^h = Soft_\alpha^h \, V^h \tag{2}$$

$$Multihead = \text{Concat}\left(head^1, ..., head^H\right) * W_O \tag{3}$$

where $Soft_\alpha^h$ denotes $Soft_\alpha$ from h-th head and $V^h \in \mathbb{R}^{J \times d}$ is value matrix for h-th head. In this paper, we adopt the relative positional encoding Shaw et al. (2018); Dai et al. (2019) for self-attention layers. The relative positional encoding gives a more strong baseline for the long-form sequence generation.

2.2 GMM ATTENTION

The previous studies on GMM attention Battenberg et al. (2020); Chiu et al. (2019) are based on early content-based attention (Cho et al., 2014; Sutskever et al., 2014). Li et al. (2020) adopt the GMM attention to the transformer framework, but they do not provide detailed descriptions. In this paper, we define the GMM attention as a variant of the multi-head attention in which each normal distribution corresponds to attention score of a single head.

Let us denote the i-th row of Q as $Q_i \in \mathbb{R}^{1 \times d}$. The normal distribution parameters for i-th step are given by

$$\Delta_i, \ \mathbf{\Sigma}_i, \ \phi_i = \zeta\left(Q_i W_\Delta\right), \ \ \zeta\left(Q_i W_{\mathbf{\Sigma}}\right), \ \ Q_i W_\phi \tag{4}$$

$$\mu_i = \Delta_i + \mu_{i-1} \tag{5}$$

where $\zeta(x)$ is a softplus function of x, and $W_\Delta \in \mathbb{R}^{d \times 1}$, $W_{\mathbf{\Sigma}} \in \mathbb{R}^{d \times 1}$, $W_\phi \in \mathbb{R}^{d \times 1}$. The GMM attention score for j-th encoder output in i-th decoder step $GMM\alpha_{i,j}$ is defined as follows:

$$N\left(j, \mu_i, \mathbf{\Sigma}_i\right) = \frac{1}{\sqrt{2\pi\mathbf{\Sigma}_i}} exp\left(-\frac{\left(j - \mu_i\right)^2}{2\mathbf{\Sigma}_i}\right) \tag{6}$$

$$GMM\alpha_{i,j} = N\left(j, \mu_i, \mathbf{\Sigma}_i\right) \tag{7}$$

$$head^h = \text{softmax}\left(\phi_i^h\right) \ \ GMM\alpha^h \ \ V^h. \tag{8}$$

In contrast to the early GMM Battenberg et al. (2020), the distinct value matrix V^h for each head in equation 8 helps jointly attending to information from different representation subspaces (Vaswani et al., 2017). The multi-head output $Multihead$ is calculated from equation 3 .

The conventional GMM attention mechanism is identical to integral over source sequences with an uniform key width as shown in Figure 1 (a). In this figure, each rectangle denotes the attention score of GMM attention with given Gaussian parameters. The GMM attention assumes that each encoder output is equally

important. However, this assumption is not satisfied for many input modalities, e.g. speech and videos from real environments. Moreover, the GMM attention score is uni-modal and cannot discard the non-informative tokens in attention window. To learn robust alignments for monotonic Seq2Seq tasks, we propose SAGMM which considers both contents and locations during attention mechanism.

2.3 SOURCE-AWARE GMM ATTENTION

Similar to the GMM mechanism, the attention score of SAGMM is derived from the Gaussian probability distribution. However, we introduce non-uniform weights for encoder outputs to consider irregular information distribution over source sequences. The SAGMM can skip or focus on the sub-sequences according to their relative importance via the non-uniform key axis. This content-aware property of SAGMM helps the model learning stable monotonic alignments from the training corpus. Furthermore, the attention score of SAGMM is multi-modal that the model can discard non-informative tokens to aggregate information distributed over several remote frames.

In SAGMM, the normal distribution paramter $\Delta_i, \mathbf{\Sigma}_i, \phi_i$, and μ_i are derived from the same equation with equation 4 and equation 5. To encode the contents of source sequences, the weight for each encoder output δ_j is given from the j-th row of the key matrix K as follows:

$$\delta_j = \sigma\left(K_j W_\delta\right). \tag{9}$$

where $W_\delta \in \mathbb{R}^{d\times 1}$ and $\sigma(x)$ denotes a sigmoid function of x. Then, the probability of normal distribution $N(\nu_j; \mu_i, \mathbf{\Sigma}_i)$ is calculated from its cumulative sum ν_j given by

$$\nu_j = \delta_j + \nu_{j-1} \tag{10}$$

$$N\left(\nu_j; \mu_i, \mathbf{\Sigma}_i\right) = \frac{1}{\sqrt{2\pi\mathbf{\Sigma}_i}} exp\left(-\frac{\left(\nu_j - \mu_i\right)^2}{2\mathbf{\Sigma}_i}\right) \tag{11}$$

Finally, the SAGMM attention score $SAGMM\alpha_{i,j}$ is defined as follows:

$$SAGMM\alpha_{i,j} = \delta_j \; N\left(\nu_j; \mu_i, \mathbf{\Sigma}_i\right). \tag{12}$$

Figure 1 (b) shows the scheme of proposed SAGMM attention. In this figure, the area of rectangles are derived from both the normal distribution parameters and weights from the encoder outputs. The widths of rectangles δ_j and their cumulative sum ν_j are derived from the encoder outputs considering the contents and order of source sequences. In this view, the proposed algorithm is source-aware attention in contrast to pure location-aware GMM attention that adopts source token index j.

In the early stage of training, we introduce length penalty loss to help learning alignments between μ and ν defined as

$$L_{length} = \lambda_{length} * \left(\left(\mu_I - min\left(I, J\right)\right)^2 + \left(\nu_J - min\left(I, J\right)\right)^2\right). \tag{13}$$

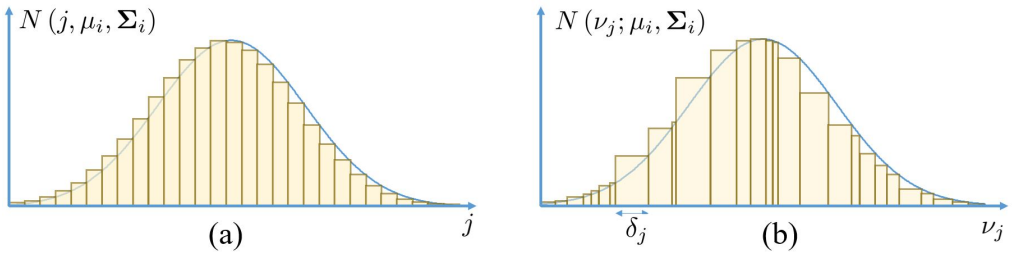

Figure 1: Scheme of (a) conventional GMM and (b) SAGMM attention score calculation.

with $\lambda_{length} = 0.0005$. We turned off L_{length} after 200K training steps in experiments. Finally, we modify 5 as $\mu_i = \mu_{i-1} + \text{relu3}(\Delta_i)$ to scale the μ_i and ν_j similar to token indices. This modification helps the interpretability of μ and ν by scaling them.

While SAGMM does not include the softmax operation over encoder outputs, following equation is satisfied for all i:

$$\lim_{j\to\infty} \lim_{\delta_j\to 0} \sum_j \delta_j \; N\left(\nu_j; \mu_i, \mathbf{\Sigma}_i\right) = 1. \tag{14}$$

In experiments, the model with SAGMM attention easily converged without tuning on the sum of attention scores. Since SAGMM does not rely on the softmax operation, we propose the SAGMM-tr for online inference applications by cropping the long-tail part of normal distribution.

2.4 SAGMM-TR FOR ONLINE INFERENCE

In SAGMM-tr, the normal distribution is truncated to limit past and future context as follows:

$$N_{tr}\left(\nu_j; \mu_i, \mathbf{\Sigma}_i\right) = \begin{cases} \frac{1}{\sqrt{2\pi\mathbf{\Sigma}_i}} exp\left(-\frac{(\nu_j-\mu_i)^2}{2\mathbf{\Sigma}_i}\right), & if\ abs\left(\nu_j - \mu_i\right) < 2\sqrt{\mathbf{\Sigma}_i} \\ 0, & else. \end{cases} \tag{15}$$

$$SAGMM\text{-}tr\ \alpha_{i,j} = \delta_j \; N_{tr}\left(\nu_j; \mu_i, \mathbf{\Sigma}_i\right). \tag{16}$$

Discarding the tokens with threshold $2\sqrt{\mathbf{\Sigma}_i}$ removes about 5% percent of attention score. We initialize the parameters of SAGMM-tr attention with those from SAGMM model and fine-tuned them until the performance converges. We found that the models with SAGMM-tr showed slightly better performance than SAGMM and mainly reported the results with SAGMM-tr model.

Algorithm 1: Decoding process of SAGMM-tr attention mechanism for a single head. The multi-headed layer can be calculated by running single-head algorithm in parallel.

Data: encoder key and value vectors K_j, V_j for $j \in \{1, 2, ...J\}$, $i = 1$, $\mu_0 = \nu_0 = 0$, $y_0 = SOS$
while $y_{i-1} \neq EOS$ **do**
  $Q_i = \text{feedforward}\left(y_{i-1}\right)$ `/* Lower decoder layers */`
  $\Delta_i, \mathbf{\Sigma}_i, \phi_i = \text{GMMparam}\left(Q_i\right), \quad \mu_i = \mu_{i-1} + \text{relu3}\left(\Delta_i\right)$ `/* equation 4 */`
  $pos_s, = \mu_i - 2\sqrt{\mathbf{\Sigma}_i}, \;\; pos_e = \mu_i + 2\sqrt{\mathbf{\Sigma}_i}, Attend = \mathbf{0}$
  for $j = 1$ *to* J **do**
    $\delta_j = \sigma\left(K_j W_\delta\right), \quad \nu_j = \nu_{j-1} + \delta_j$
    if $\nu_j > pos_e$ **then**
      Break
    else if $\nu_j > pos_s$ **then**
      $SAGMM\text{-}tr\ \alpha_{i,j} = \delta_j \; N_{tr}\left(\nu_j; \mu_i, \mathbf{\Sigma}_i\right)$ `/* equation 15, 16 */`
      $Attend = Attend + SAGMM\text{-}tr\ \alpha_{i,j} V_j$
  $y_i = \text{Output}\left(\text{Decoder}\left(Attend\right)\right), \quad i = i + 1$

The fixed length window with hyperparameters can be also adopted to truncate a Gaussian distribution. We tested the window width $c = 15$ centered by μ_i and compared the recognition performance of the models to adaptive window width in experiments.

For the online speech recognition with SAGMM-tr, we randomly concatenate the **1**-vector after the end of source sequence with a probability p_{eos} in the training stage and trained to emit the end-of-sequence token EOS only with the utterances containing **1**-vector. The **1**-vector concatenation helps to suppress the EOS token in a long silence part. Finally, we build the uni-directional encoder whose maximum latency is similar

to Zhang et al. (2020); Dong et al. (2019) by adopting the block-wise mask on self-attention layers. The detailed explanation for the block-wise masking is given in Appendix. Algorithm 1 shows a psuedo code for the SAGMM-tr attention in the inference stage.

3 RELATED WORKS

Chiu & Raffel (2018) propose monotonic chunkwise attention (MoChA) based on hard monotonic sampling over attention window (Raffel et al., 2017b). The monotonic multi-head attention Ma et al. (2019); Inaguma et al. (2020) and monotonic infinite lookback attention Arivazhagan et al. (2019) are based on the concept similar to (Chiu & Raffel, 2018). MoChA replaces hard sampling operation to probability distribution over memory in the training stage. In contrast, SAGMM uses the same operations for training and inference. Also, MoChA relies on numerically unstable cumulative product which requires additional efforts for training (Miao et al., 2019; Raffel et al., 2017a). In contrast, SAGMM adopts the stable cumulative sum to describe monotonic alignments. Finally, MoCha is not effective for long-form speech recognition compared to GMM attention Chiu et al. (2019) while our model outperformed GMM in "test-long" set in experiments.

Dong & Xu (2020) introduce the encoder output weight for ASR task. Compared to their work, our attention mechanism considers the decoder content during attention score calculation. SAGMM also allows the overlap between successive attention windows which is helpful to generate semantics that span several decoder tokens. Lee et al. (2020) propose the attention mechanism which enables early inference without softmax operation. Their approach does not explicitly model monotonic alignment and the update gate for online inference in Lee et al. (2020) is uni-modal similar to conventional GMM attention.

The non-attentive neural network-based approaches which do not rely on encoder-decoder attention have been studied by Connectionist Temporal Classification (CTC) Graves et al. (2006); Liptchinsky et al. (2017); Li et al. (2019), and RNN-transducer (RNN-T) Graves (2012), Transformer-transducer (Transformer-T) Zhang et al. (2020); Yeh et al. (2019), and Imputer Chan et al. (2020). We focus on the Seq2Seq approach which do not require an assumption on sequence lengths and uses the simple beam search algorithm for the inference. Finally, the block-wise inference has been widely studied with manually defined decoding blocks Jaitly et al. (2015); Tsunoo et al. (2020) or joint decoding with CTC model Moritz et al. (2020). We do not use the joint CTC decoding Kim et al. (2016) nor human supervision for the blocked inference.

4 EXPERIMENT

We conducted experiments in the speech command Warden (2018) and LibriSpeech Panayotov et al. (2015) datasets to compare the performance of our model in standard, online and long-form ASR tasks. We only compare the end-to-end speech recognition model without the external language model rescoring. We also performed preliminary experiments on the translation task to show the performance of the SAGMM-tr on non-monotonic tasks.

4.1 EXPERIMENTS ON CONCATENATED SPEECH COMMAND DATASET

First, we performed a speech recognition experiment with limited vocabulary to show that the common soft attention algorithm has difficulty to decode utterances with unseen sequence lengths. The speech command dataset consists of 1 seconds speeches from various speakers saying single words from 30 vocabulary. We built 100K utterances of training corpusWarden (2018) by concatenating randomly chosen 5-9 utterances in the training data. 500 test utterances were also built by concatenating $\{3, 7, 10, 15, 20\}$ randomly chosen words in the test set. We cropped both the start and end of chosen utterances randomly from 0.05s to 0.15s before concatenating them to prevent the SAGMM from finding a trivial solution.

We compared the word error rate (WER) of transformers with soft and SAGMM attentions on test sets. Detailed hyperparameters for the experiment are described in Appendix. In table 1, the soft attention mechanism tends to memorize the sequence length distribution from the training corpus and fails to decode cor-

rectly in the unseen sequence length. On the other hand, the proposed SAGMM algorithm was robust to sequence length mismatches. The decoded examples for soft attention and SAGMM are shown in Appendix.

Table 1: The word error rates (%) on the concatenated speech command dataset with various number of concatenated words.

	3	**7**	**10**	**15**	**20**
soft attention	40.00	**5.71**	13.20	39.80	56.05
SAGMM	**4.67**	6.29	**5.50**	**5.40**	**6.45**

4.2 EXPERIMENTS ON LIBRISPEECH DATASET

We conducted experiments on the LibriSpeech whose training dataset consists of 960 hours of read audio books (Panayotov et al., 2015). We validated the performance of models on dev sets and reported WERs in tests sets with the fixed beam width 4.

First, we compared the performance of SAGMM-tr with conventional algorithms in the offline speech recognition. We trained the multi-head transformers similar to the previous experiments. The encoder consisted of 10 layers of self-attention block with 768 hidden nodes and 3,072 filters. The decoder was constructed by stacking 4 self-attention blocks with the same node and filter sizes. We adopt the CTC loss on the encoder output with $\lambda_{ctc} = 0.1$ for fair comparison with the previous attention-based models Lee et al. (2020); Dong & Xu (2020); Kim et al. (2019). The detail hyperparameters for the experiment and performance of SAGMM-tr without CTC loss are shown in Appendix.

Table 2 shows the performance of various non-streamable models. In non-streamable models, our baseline transformer showed similar performance with previous studies considering the number of parameters. We also compared the performance of GMM and SAGMM-tr models with the bi-directional encoder to other streamable models. The performance of SAGMM-tr attention matched our transformer model and showed better performance than other streamable Seq2Seq models. From the results, SAGMM-tr successfully described the monotonic alignments between speech and transcription without attending all encoder outputs.

Next, we tested the performance of SAGMM-tr in online ASR task. We trained the single-head and multi-head SAGMM-tr models with the block-wise masked encoder. The truncated GMM model was also tested. However, the model attended to the future context as shown in A.3 and failed to learn the correct alignments. In the multi-head SAGMM-tr, several heads in the first decoder layer were not trained and softmax(ϕ_i^h) went to zero. We simply pruned these heads after the performance on dev set converged and fine-tuned for additional 30K steps. Figure 2 shows the attention plots of the single-head SAGMM-tr which successfully captures alignments between speech and transcription for an utterance with a long silence. Attention plots for soft attention, GMM, and multi-head SAGMM-tr models are in Appendix A.3.

Table 3 shows the performance of models with various online inference approaches. From the table, the single-head SAGMM-tr outperformed other CTC and attention-based algorithms. The multi-head mechanism further improved the performance of SAGMM-tr. Compared to the similar model with the bi-directional encoder, the block-wise mask on encoder self-attention layers increased the WER on test-clean set by 0.31%. Our model showed slightly worse performance than state-of-the-art algorithm with transformer-T (Zhang et al., 2020). Subword regularization Kudo (2018); Hannun et al. (2019) and

Figure 2: Attention plots for single-head SAGMM-tr model with a speech utterance with a long silence.

Table 2: The word error rates (%) of offline speech recognition models on LibriSpeech (with number of parameters when available).

		test-clean	**test-other**
Non-streamable	LAS (Zeyer et al., 2018)	4.70	15.20
	LAS (180M) (Park et al., 2019)	3.4	8.3
	LAS (361M) (Park et al., 2019)	**2.8**	**6.8**
	Transformer (270M) (Synnaeve et al., 2020)	2.89	6.98
	Transformer (121M) (Ours)	**3.20**	**7.96**
Streamable model, bi-directional enc.	RNN-T (130M) (Zhang et al., 2020)	3.2	7.8
	Transformer-T (45.7M) (Yeh et al., 2019)	6.08	13.89
	Transformer-T (139M) (Zhang et al., 2020)	**2.4**	**5.6**
	DecGRC (Lee et al., 2020)	4.83	14.90
	Mocha (Lee et al., 2020)	4.95	15.32
	CIF (Dong & Xu, 2020)	3.41	12.62
	GMM (121M) (Ours)	3.71	8.52
	SAGMM-tr (121M) (Ours)	**3.21**	**7.99**

Table 3: The word error rates (%) of online speech recognition models on LibriSpeech (with number of parameters when available).

		test-clean	**test-other**
CTC variants	Gated ConvNet (Liptchinsky et al., 2017)	6.7	-
	CTC (Chan et al., 2020)	4.6	13.0
	Imputer(DP)(Chan et al., 2020)	4.0	**11.1**
	Jasper (333M) (Li et al., 2019)	**3.86**	11.95
Transducers	Transformer-T (45.7M) (Yeh et al., 2019)	6.47	15.79
	ConvT-T (67M) (Huang et al., 2020)	3.5	8.3
	Transformer-T (139M) (Zhang et al., 2020)	**3.0**	**7.7**
Streamable Seq2Seq	Triggered Attention (Moritz et al., 2019)	7.4	19.2
	Mocha (Kim et al., 2019)	6.30	18.41
	Mocha (Lee et al., 2020)	5.15	16.45
	DecGRC (Lee et al., 2020)	5.87	17.04
	CIF (Dong & Xu, 2020)	3.96	11.19
	SAGMM-tr (121M) (single-head)	3.73	9.92
	SAGMM-tr (121M) (multi-head)	**3.52**	**9.29**

sequence-level loss Sabour et al. (2019); Prabhavalkar et al. (2018) can help to match the performance of SAGMM-tr with the best transducer models. The SAGMM-tr can also be adopted to various monotonic Seq2Seq tasks easily since it can be inferred via simple beam search without probability marginalization in transducers and does not require assumption that the target should not be longer than the source.

We also tested SAGMM-tr models with fixed attention window width. We started with the SAGMM-tr models trained for the online ASR experiment and measured the WERs before and after fine-tuning with the constant window width ($c = 15$ frames). From the results in table 4 (a), SAGMM-tr attention mechanism can be also adopted to the models with fixed attention window width. The attention plots for SAGMM-tr with the fixed window width are shown in Appendix.

Finally, we concatenated utterances from the same speaker in "test-clean" set to build "test-long" set for the long-form speech recognition. The average length of utterances in the test-long set was 54 seconds. In the SAGMM-tr model, the beam search algorithm was slightly modified from Algorithm 1 to suppress the EOS token until the attention window covers $\nu_{h,J}$. The performance of GMM attention did not improved by the same modification since GMM attention attends to $\nu_{h,J}$ before the end of sentence.

Table 4: The experiment results on (a) SAGMM-tr with the fixed window length and (b) long-form speech recognition on "test-long" set. (sh) and (mh) denote the single-head and multi-head attention, respectively.

(a)

	test-clean	**test-other**
SAGMM-tr (sh)	3.73	9.92
SAGMM-tr (mh)	**3.52**	**9.29**
SAGMM-tr (sh,fixed)	3.67	9.76
SAGMM-tr (mh,fixed)	4.45	10.34
+fine-tuning	3.60	9.42

(b)

	test-clean	**test-long**
Transformer	**3.20**	32.77
GMM	3.71	24.71
SAGMM-tr (sh)	3.73	**5.60**
SAGMM-tr(mh)	3.52	8.13

From the table 4 (b), SAGMM-tr outperformed the conventional soft and GMM attention models in long-form speech recognition. Comparing the SAGMM-tr models, the multi-head SAGMM-tr showed worse performance than the single-head model. The performance degradation of multi-head mechanism on "test-long" set might come from accumulated attention windows range mismatches. Future works should explore techniques to regularize the attention window ranges in multi-headed models.

4.3 Experiments on a non-monotonic task

The proposed algorithm focuses on strict monotonic tasks which do not allow reordering. However, the self-attention layers in the encoder could reorder the semantics in input sequences before monotonic alignments from the SAGMM-tr. We conducted machine translation experiments on the WMT EN-DE environment to show the performance of SAGMM-tr attention in non-monotonic tasks. We trained the encoder-decoder model similar to the transformer-base in Vaswani et al. (2017) except the head size 4 of the encoder-decoder and decoder self-attention. We trained model with 4.5M pairs of sentences and measured the BLEU score on the newstest2013 set after 100K training steps. For uni-directional encoder, the block-wise mask with $M = 5$ is applied to SAGMM-tr model with bi-directional encoder and fine-tuned for additional 40K steps.

Table 5 shows the performance of soft and SAGMM-tr attention models on the newstest2013 set. As expected, the performance of SAGMM-tr did not match that of the conventional transformer models with soft attention. The uni-directional encoder also deteriorates the performance of SAGMM-tr similar to speech recognition experiments. Since SAGMM-tr does not consider the local reordering, the proposed algorithm has difficulty in learning locally non-monotonic functions. Our future works will improve the performance of SAGMM-tr on local reordering tasks by positive constraint relaxation on Δ_i.

Table 5: The BLEU scores on the newstest2013 set for the WMT EN-DE environment.

	EN-DE
Transformer (Vaswani et al., 2017)	25.8
Transformer (Ours)	25.7
SAGMM-tr (bi-enc)	24.0
SAGMM-tr (uni-enc)	23.7

5 Conclusion

We have proposed the SAGMM which attends the subset of source sequences according to the normal distribution on the weighted encoder output axis, considering both contents and order of elements in sequences. We have also proposed SAGMM-tr for online/real-time inference applications. From the results on various speech recognition tasks, the proposed attention mechanism is shown to be able to learn the monotonic alignments between source and target sequences without a human supervision. The performance of transformer with SAGMM-tr improved for online and long-form speech recognition without performance degradation in the standard offline ASR task. In future works, we would like to adopt the SAGMM-based attention mechanisms for natural language processing tasks which allow local reordering during the sequence generation. We are also interested in latency minimization for online inference by controlling μ and $\mathbf{\Sigma}$ in both the training and inference stages.

REFERENCES

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. *ACL*, 2019.

Eric Battenberg, RJ Skerry-Ryan, Soroosh Mariooryad, Daisy Stanton, David Kao, Matt Shannon, and Tom Bagby. Generating sequences with recurrent neural networks. *ICASSP*, 2020.

William Chan, Chitwan Saharia, Geoffrey Hinton, Mohammad Norouzi, and Navdeep Jaitly. Imputer: Sequence modelling via imputation and dynamic programming. *ICML*, 2020.

Chung-Cheng Chiu and Colin Raffel. Monotonic chunkwise attention. *ICLR*, 2018.

Chung-Cheng Chiu, Wei Han, Yu Zhang, Ruoming Pang, Sergey Kishchenko, Patrick Nguyen, Arun Narayanan, Hank Liao, Shuyuan Zhang, Anjuli Kannan, Rohit Prabhavalkar, Zhifeng Chen, Tara Sainath, and Yonghui Wu. A comparison of end-to-end models for long-form speech recognition. *ASRU*, 2019.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *EMNLP*, 2014.

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

Linhao Dong and Bo Xu. Cif: Continuous integrate-and-fire for end-to-end speech recognition. *ICASSP*, 2020.

Linhao Dong, Shuang Xu, and Bo Xu. Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. *ICASSP*, 2018.

Linhao Dong, Feng Wang, and Bo Xu. Self-attention aligner: A latency-control end-to-end model for asr using self-attention network and chunk-hopping. *ICASSP*, 2019.

Daniel Duckworth, Arvind Neelakantan, Ben Goodrich, Lukasz Kaiser, and Samy Bengio. Parallel scheduled samplings. *ArXiv*, abs/1906.04331, 2019.

Alex Graves. Sequence transduction with recurrent neural networks. *ArXiv*, abs/1211.3711, 2012.

Alex Graves. Generating sequences with recurrent neural networks. *ArXiv*, abs/1308.0850, 2013.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. onnectionist tem-poral classification: labelling unsegmented sequence data with recurrent neural networks. *ICML*, 2006.

Awni Hannun, Ann Lee, Qiantong Xu, and Ronan Collobert. Sequence-to-sequence speech recognition with time-depth separable convolutions. *ArXiv*, abs/1904.02619, 2019.

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

Hirofumi Inaguma, Masato Mimura, and Tatsuya Kawahara. Enhancing monotonic multihead attention for streaming asr. *ArXiv*, abs/2005.09394, 2020.

Navdeep Jaitly, David Sussillo, Quoc V. Le, Oriol Vinyals, Ilya Sutskever, and Samy Bengio. A neural transducer. *ArXiv*, abs/1511.04868, 2015.

Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, Ryuichi Yamamoto, Xiaofei Wang, Shinji Watanabe, Takenori Yoshimura, and Wangyou Zhang. A comparative study on transformer vs rnn in speech applications. *ASRU*, 2019.

Kwangyoun Kim, Kyungmin Lee, Dhananjaya Gowda, Junmo Park, Sungsoo Kim, Sichen Jin, Young-Yoon Lee, Jinsu Yeo, Daehyun Kim, Seokyeong Jung, Jungin Lee, Myoungji Han, and Chanwoo Kim. Attention based on-device streaming speech recognition with large speech corpus. *ASRU*, 2019.

Suyoun Kim, Takaaki Hori, and Shinji Watanabe. Joint ctc-attention based end-to-end speech recognition using multi-task learning. *ArXiv*, abs/1609.06773, 2016.

Taku Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates. *ACL*, 2018.

Hyeonseung Lee, Woo Hyun Kang, Sung Jun Cheon, Hyeongju Kim, and Nam Soo Kim. Gated recurrent context: Softmax-free attention for online encoder-decoder speech recognition. *ArXiv*, abs/2007.05214, 2020.

Jason Li, Vitaly Lavrukhin, Boris Ginsburg, Ryan Leary, Oleksii Kuchaiev, Jonathan M Cohen, Huyen Nguyen, and Ravi Teja Gadde. Jasper: An end-to-end convolutional neural acoustic model. *Interspeech*, 2019.

Naihan Li, Yanqing Liu, Yu Wu, Shujie Liu, Sheng Zhao, and Ming Liu. Robutrans: A robust transformer-based text-to-speech model. *AAAI*, 2020.

Vitaliy Liptchinsky, Gabriel Synnaeve, and Ronan Collobert. Letter-based speech recognition with gated convnets. *ArXiv*, abs/1712.09444, 2017.

Xutai Ma, Juan Pino, James Cross, Liezl Puzon, and Jiatao Gu. Monotonic multihead attention. *ArXiv*, abs/1909.12406, 2019.

Haoran Miao, Gaofeng Cheng, Pengyuan Zhang, Ta Li, and Yonghong Yan. Online hybrid ctc/attention architecture for end-to-end speech recognition. *Interspeech*, 2019.

Niko Moritz, Takaaki Hori, and Jonathan Le Roux. Triggered attention for end-to-end speech recognition. *ICASSP*, 2019.

Niko Moritz, Takaaki Hori, and Jonathan Le Roux. Streaming automatic speech recognition with the transformer model. *ArXiv*, abs/2001.02674, 2020.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. *ICASSP*, 2015.

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. *Interspeech*, 2019.

Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Łukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. *ICML*, 2018.

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. *ICASSP*, 2018.

Colin Raffel, Minh-Thang Luong, Peter J. Liu, Ron J. Weiss, and Douglas Eck. Online and linear-time attention by enforcing monotonic alignments. *ICML*, 2017a.

Colin Raffel, Minh-Thang Luong, Peter J. Liu, Ron J. Weiss, and Douglas Eck. Online and linear-time attention by enforcing monotonic alignments. *ICLR*, 2017b.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv*, abs/1910.10683, 2019.

Sara Sabour, William Chan, and Mohammad Norouzi. Optimal completion distillation for sequence learning. *ICLR*, 2019.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *NAACL*, 2018.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. *NIPS*, 2014.

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. *Workshop on Self-supervision in Audio and Speech,*, 2020.

Emiru Tsunoo, Yosuke Kashiwagi, and Shinji Watanabe. Streaming transformer asr with blockwise synchronous inference. *ArXiv*, abs/2006.149411, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Neurips*, 2017.

Pete Warden. Speech commands: A dataset for limited-vocabulary speech recognition. *ArXiv*, abs/1804.03209, 2018.

Ching-Feng Yeh, Jay Mahadeokar, Kaustubh Kalgaonkar, Yongqiang Wang, Duc Le, Mahaveer Jain, Kjell Schubert, Christian Fuegen, and Michael L. Seltzer. Transformer-transducer: End-to-end speech recognition with self-attention. *ArXiv*, abs/1910.12977, 2019.

Albert Zeyer, André Merboldt, Ralf Schlüter, and Hermann Ney. A comprehensive analysis on attention models. *NIPS: Workshop IRASL*, 2018.

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. *ArXiv*, abs/2002.02562, 2020.

A APPENDIX

A.1 BLOCK-WISE MASK FOR ENCODER SELF-ATTENTION LAYER

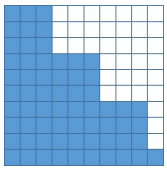

Figure 3: Encoder self-attention mask between query (vertical axis) and key (horizontal axis) in $M = 3$ case.

The block-wise uni-directional mask was adopted for encoder self-attention layers to block future context more than M tokens. 3 shows the block-wise masking with $M = 3$ case. In the experiment with SAGMM-tr, we chose $M = 30$ whose gmaximum latency is comparable to other studies with uni-directional encoders (Zhang et al., 2020; Dong et al., 2019). The blcok-wise masking is different from lookahead operation since encoder outputs from the block-wise masking are not guaranteed to look future M token depends on their position in block while the lookahead operation provides maximum future context for all encoder outputs.

A.2 EXPERIMENT CONFIGURATIONS

For all speech recognition experiments, the 80-dimensional log-Mel filterbank features were extracted with a 25ms window and shifted every 10ms. Each input vector stacked 3 log-Mel feature vectors, downsampled to 30ms frame rate.

In the speech command dataset experiment, We trained the multi-headed transformer Vaswani et al. (2017); Dong et al. (2018) with soft attention and SAGMM attention for encoder-decoder alignments. We adopt the relative positional encoding for both the encoder and decoder self-attention (Dai et al., 2019). The encoder consisted of 6 layers of self-attention block with 512 nodes and 2,048 filter size, respectively. The decoder was constructed by stacking 3 self-attention block with the same node and filter sizes. The number of heads for encoder self attention, encoder-decoder attention, and decoder self attention were $\{8, 4, 4\}$, respectively. We employed label smoothing of value $\epsilon_{ls} = 0.15$, and applied parallel scheduled sampling with probability 0.2 after 100k (Duckworth et al., 2019). We used the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 10^{-9}$). The learning rate was linearly increased to 0.1 until 16k steps and set constant in following training steps. We used the tensorflow framework and trained the model until 500K step with 4 m40 GPUs. We averaged 5 last checkpoints before the inference. The dropout ratio for attention weights, rectified linear units, output of sub-layers, and neural network input after positional encoding were 0.2, 0.1, 0.2, and 0.2, respectively. The Specaug algorithm Park et al. (2019) was also applied during training. The width of beam was set to 4 in all experiments. The output token segments are grepheme of the reference.

In the LibriSpeech experiment, The learning rate was linearly increased to 0.1 until 16k steps and set constant until the model converges on dev-clean dataset. Then, we lowered learning rate to 0.02 and fine-tuned until the model converges. We used the tensorflow framework and the model is trained from 7 to 10 days on 8 p40 GPUs. We averaged $\{5, 10, 15\}$ last checkpoints before the performance evaluation. The 1K wordpiece vocabularies extracted from training dataset is used for output token segmentation. Other parameters are the same with those in the speech command dataset experiment.

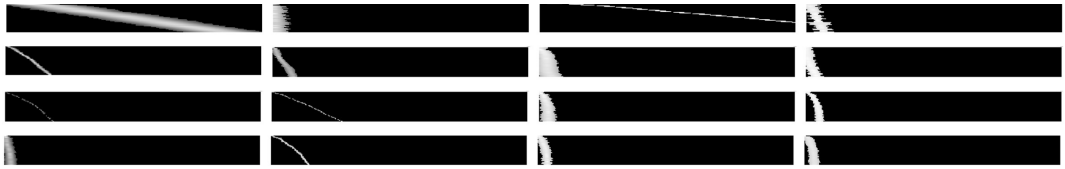

Figure 4: Attention alignments plots from multi-head truncated GMM model.

A.3 ATTENTION PLOT FOR VARIOUS ENCODER-DECODER MODELS

To show the SAGMM-tr is able to learn alignments between encoder and decoders, we plot the attention score for single and multi-head models. We used the utterance "Chapter eleven ¡long pause¿ the morrow brought a very sober looking morning the sun making only a few efforts to appear and Catherine augured from it everything most favourable to her wishes" in LibriSpeech.

The truncated GMM model fails to learn the true alignment attended several distinct windows in 4. Surprisingly, the performance of truncated GMM was 4.03% and 10.16% for test-clean and test-other, respectively. The decent performance of truncated GMM implies that the GMM learns to attend as many frames as possible instead of learning alignments between sequences. Therefore, the truncated GMM attention is not suitable for streaming inference.

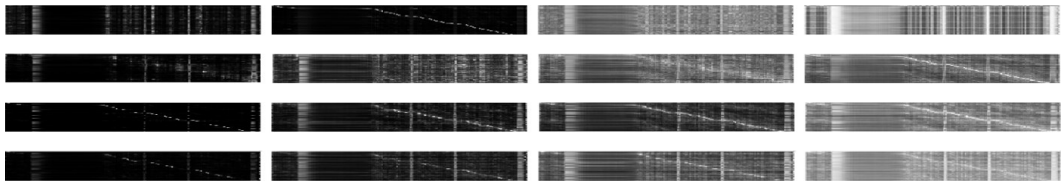

Figure 5: Attention alignments plots from multi-head soft attention model.

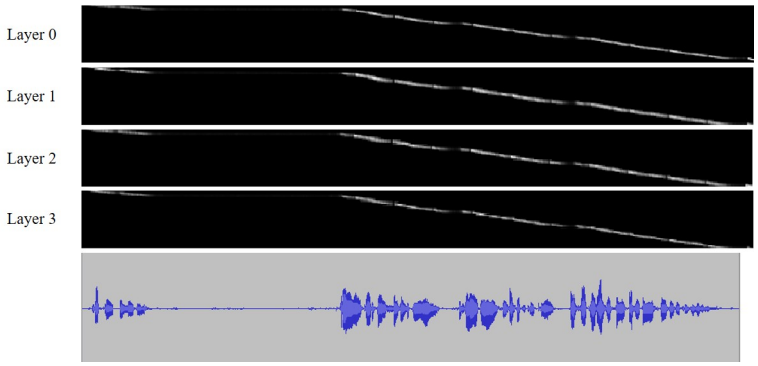

Figure 6: Attention alignments plots for single-head SAGMM-tr for the utterance "Chapter eleven ¡long pause¿ the morrow brought a very sober looking morning the sun making only a few efforts to appear and Catherine augured from it everything most favourable to her wishes" in LibriSpeech. The attention plot in time domain longer than the original wave due to **1**-vector concatenation for EOS 2.4

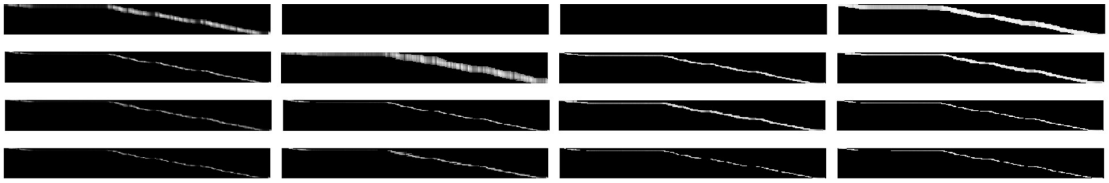

Figure 7: Attention alignments plots from multi-head SAGMM-tr model with adaptive window width. The empty plots denote the pruned head in the first layer

Figure 8: Attention alignments plots from multi-head SAGMM-tr attention model with fixed window width. The empty plots denote the pruned head in the first layer

In the soft attention model, the several heads learns monotonic alignments between inputs and outputs as shown in Figure 5. However, other heads attend to the whole speech presence or absence intervals. We assume these heads help utterance-level normalization. While a few heads show monotonicity, the soft attention is not suitable for online inference due to softmax operation over all frames.

The attention alignment figure for single and multi-headed SAGMM-tr model is shown in Figures 6, 7 and 8. From the figures, both models learn the optimal alignments between input wave and transcription and distinguish the word boundaries without relying on human knowledge. The width of attention windows in multi-head SAGMM-tr with adaptive window width are longer than those in the single-head model. The additional objective function for regularizing $\mathbf{\Sigma}_i$ would be helpful to reduce the window length in figure 7. In the multi-head model with fixed window length, the model learns the multi-head monotonic alignment without increasing the window width. Considering the attention window width as a hyperparameter from external knowledge, it is shown that additional information could improve the model latency in SAGMM-tr while the model can also be trained without human knowledge.

A.4 Ablation studies

We tested the three ablation experiments on SAGMM-tr attention in Figure 6. First, we tested the model without CTC loss on encoder output. The CTC loss improves the word error rate performances of model similar to previous studies, but the model showed decent performance without the CTC loss. Second, we tested the SAGMM-tr model with uni-directional encoder without initialization from the bi-directional encoder. The result shows slightly worse performance compared to the original SAGMM-tr. Finally, we train the SAGMM model in the training and decoded with the SAGMM-tr decoding scheme without fine-tuning. The performance degradation in 6 shows that the mismatch between training and test stages should be removed by the fine-tuning with truncated normal distribution.

Table 6: The word error rate (%) of ablation experiments for SAGMM-tr attention mechanism in LibriSpeech.

	test-clean	**test-other**
SAGMM-tr	3.52	9.29
SAGMM-tr w/o CTC loss	3.84	9.86
SAGMM-tr uni-encoder from scratch	3.69	9.79
SAGMM-tr w/o truncated attention fine-tuning	3.76	9.96

A.5 ATTENTION PLOTS FOR SAGMM-TR IN MACHINE TRANSLATION EXPERIMENTS

We plot the first three decoder layers of the SAGMM-tr model with bi-directional encoder for an example utterance "The patient really needs to be made to understand the degree of risk of his cancer, by offering him the options available, not necessarily treating prostate cancers that are not long-term life threatening, and opting instead, in such cases, for active monitoring of the disease." and model output "Der Patient muss wirklich verstanden werden, um den Grad des Risikos seines Krebses zu verstehen, indem er ihm die Möglichkeiten zur Verfügung stellt, nicht notwendigerweise mit Prostatakrebs umzugehen, die nicht langfristiges Leben bedrohen und sich stattdessen in solchen Fällen für eine aktive Überwachung der Krankheit entscheiden.".

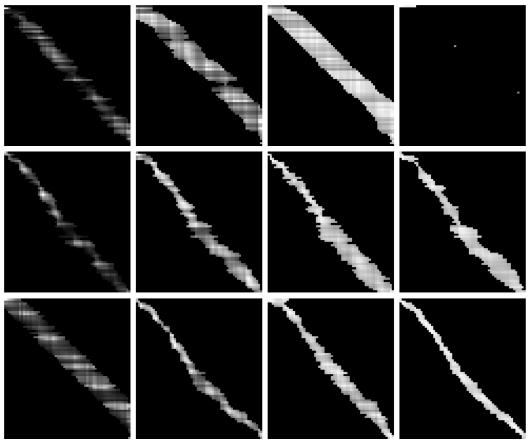

Figure 9: Encoder-decoder attention alignments plots from multi-head SAGMM-tr model with bi-directional encoder in machine translation experiment.

We also plot the first three encoder-decoder attention and encoder-self attention layers of the SAGMM-tr model with uni-directional encoder for an example utterance "The patient really needs to be made to understand the degree of risk of his cancer, by offering him the options available, not necessarily treating prostate cancers that are not long-term life threatening, and opting instead, in such cases, for active monitoring of the disease." and model output "Der Patient muss wirklich verstanden werden, um das Risiko seines Krebses zu verstehen, indem er ihm die Möglichkeiten zur Verfügung stellt, nicht notwendigerweise mit Prostatakrebs umzugehen, die nicht langfristiges Leben bedrohen, und stattdessen in solchen Fällen für eine aktive Überwachung der Krankheit zu entscheiden.".

From the figures 9, 10 and 11, the transformers with SAGMM-tr approximates the machine translation task as a strict monotonic task and finds optimal alignments under the assumption. Since the machine translation

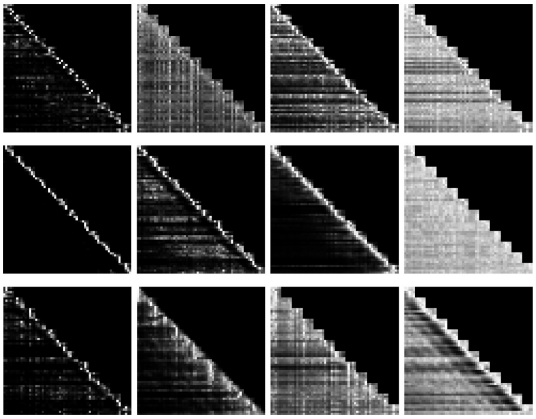

Figure 10: Encoder self-attention alignments plots from multi-head SAGMM-tr model with uni-directional encoder in machine translation experiment.

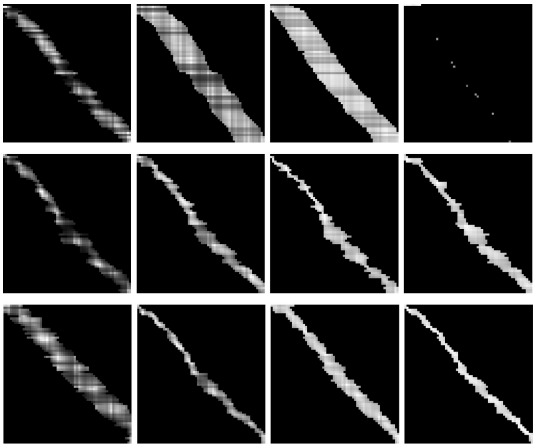

Figure 11: Encoder-decoder attention alignments plots from multi-head SAGMM-tr model with uni-directional encoder in machine translation experiment.

is not strictly monotonic task, the performance of the SAGMM-tr attention deteriorates compared to the soft attention. However, the transformer with SAGMM-tr attention and uni-directional encoder enables early inference for simultaneous translation.

A.6 Decoded examples of soft attention and SAGMM models on speech command and LibriSpeech datasets

Table 7 and 8 show the decoded examples from the models with unseen sequence lengths. From the tables, soft attention has difficulty to generate the sequences that are significantly shorter or longer than the training corpus distribution. In contrast, the SAGMM-tr is robust to the sequence length mismatch and generates the transcription to arbitrary length.

Table 7: The example results for attention-based models in concatenated speech command dataset.

	Transcription
Reference	seven three stop
Soft attention	seven three stop stop
SAGMM	seven three stop
Reference	two right two no cat bird right left tree four tree one wow marvin left
Soft attention	two right two no cat bird right left marvin left
SAGMM	two right two no cat bird right left tree four tree one wow marvin left

Table 8: The example results for attention-based models in test-clean-long set from LibriSpeech dataset. (sh) and (mh) denote the single-head and multi-head attention, respectively.

	Transcription
Reference	it was idle for him to move himself to be generous towards them to tell himself that if he ever came to their gates stripped of his pride beaten and in beggar's weeds that they would be generous towards him loving him as themselves idle and embittering finally to argue against his own dispassionate certitude that the commandment of love bade us not to love our neighbour as ourselves with the same amount and intensity of love but to love him as ourselves with the same kind of love the phrase and the day and the scene harmonized in a chord words was it their colours they were voyaging across the deserts of the sky a host of nomads on the march voyaging high over ireland westward bound the europe they had come from lay out there beyond the irish sea europe of strange tongues and valleyed and woodbegirt and citadelled and of entrenched and marshalled races
Soft attention	it was idle for him to move himself to be generous towards them to tell himself that if he ever came to their gates stripped of his pride beaten and in beggars weeds that they would be generous towards him loving him as themselves to love him as ourselves with the same amount and intensity of love the europe they had come from lay out there beyond the irish sea europe of strange tongues and valleyed and martialed races
SAGMM-tr(sh)	it was idle for him to move himself to be generous towards them to tell himself that if he ever came to their gates stripped of his pride beaten and in beggars weeds that they would be generous towards him loving him as themselves idle and embittering finally to argue against his own dispassionate certitude that the commandment of love bade us not to love our neighbour as ourselves with the same amount and intensity of love but to love him as ourselves with the same kind of love the phrase and the day and the scene harmonized in accord words was it their colors they were voyaging across the deserts of the sky a host of no mads on the march voyaging high over ireland westward bound the europe they had come from lay out there beyond the irish sea europe of strange tongues and valleys and wood be girt and set and of intrenched and marshalled races
SAGMM-tr(mh)	it was idle for him to move himself to be generous towards them to tell himself that if he ever came to their gates stripped of his pride beaten and in beggar's weeds that they would be generous towards him loving him as themselves idle and embittering finally to argue against his own dispassionate certitude that the command of love bade us not to love our neighbour as ourselves with the same amount and intensity of love but to him as ourselves with the same kind of love the phrase and the day and the scene harmonized in accord words was it their colors they were voyaging across the deserts of the sky a host of no mads on the march voyaging high over ireland westward bound the europe they had come from lay there beyond the irish sea europe of strange tongues and valleyed and wood girt and citadel and of intrenched and marched and races