Parallel Attention Mechanisms in Neural Machine Translation

Julian Richard Medina
Computer Science
University of Colorado Colorado Springs
Colorado Springs, CO, USA
jmedina5@uccs.edu

Jugal Kalita
Computer Science
University of Colorado Colorado Springs
Colorado Springs, CO, USA
jkalita@uccs.edu

***Abstract*—Recent papers in neural machine translation have proposed the strict use of attention mechanisms over previous standards such as recurrent and convolutional neural networks (RNNs and CNNs). We propose that by running traditionally stacked encoding branches from encoder-decoder attention-focused architectures in parallel, that even more sequential operations can be removed from the model and thereby decrease training time. In particular, we modify the recently published attention-based architecture called Transformer by Google, by replacing sequential attention modules with parallel ones, reducing the amount of training time and substantially improving BLEU scores at the same time. Experiments over the English to German and English to French translation tasks show that our model establishes a new state of the art.**

***Index Terms*—machine translation, transformer, attention**

I. Introduction

Historically, statistical machine translation involved extensive work in the alignment of words and phrases developed by linguistic experts working with computer scientists [1]. Deep Learning surpasses these historically used methods and has primarily replaced these with the recent use of neural machine translation (NMT). The predominant design of the state of the art is the encoder-decoder model. The encoder takes sequential text, turning it into an internal representation. The decoder then takes this internal representation and generates a subsequent output. Since their emergence, attention mechanisms [2] have added to the effectiveness of the encoder decoder model and have been at the forefront of machine translation.

Attention mechanisms help the neural system focus on parts of the input, and possibly the output as it learns to translate. This concentration facilitates the capturing of dependencies between parts of the input and the output. After training the network, the attention mechanism enables the system to perform translations that can handle issues such as the movement of words and phrases, and fertility. However, even with these attention mechanisms, NMT models have their drawbacks, which include long training time and high computational requirements.

Recent papers [3], [4] in neural machine translation have proposed the strict use of attention mechanisms in networks such as the Transformer over previous approaches such as recurrent neural networks (RNNs) [5] and convolutional neural networks (CNNs) [6]. In other words, these approaches dispense with recurrences and convolutions entirely. In practice, attention mechanisms have mostly been used with recurrent architectures because removing the recurrent nature of the architecture makes the training more efficient by the removal of necessary sequential steps.

This paper contributes by continuing to pursue the removal of sequential operations within encoder-decoder models. These operations are removed through the parallelization of previously stacked encoder layers. This new parallelized model can obtain a new state of the art in machine translation after being trained on one NVIDIA GTX 1070 for as little as three hours.

The paper includes the following: a discussion of related work in the field of machine translation including encoder-decoder models and attention mechanisms; an explanation of the proposed novel architecture with motivations; and a description of the used methodology, along with evaluation including used data sets, hardware, hyper-parameters, and metrics. This paper concludes with results and possible avenues for future research.

II. Related Work

There has been a plethora of work in the past several years on end-to-end neural translation. ByteNet [7] uses CNNs with dilated convolutions for both encoding and decoding. Zhou et al. [8] use stacked interleaved bi-directional LSTM layers (up to 16 layers) with skipped connections; ensembling gives the best results. Google's earlier and path-breaking end-to-end translation approach [9] uses 16 LSTM layers with attention; once again, ensembling produces the best results. Facebook's end-to-end translation approach [10] depends entirely on CNNs with attention mechanism.

Our work reported in this paper is based on another translation work by Google. Google's Vaswani et al. [3] proposed the reduction in the sequential steps seen in CNNs and RNNs. The sole use of attention mechanisms and feed-forward networks within the common encoder-decoder sequential model replaces the necessity of deep convolutions for distant dependent relationships, and the memory and computation intensive operations required within recurrent networks. Original training and testing by Vaswani et al. were over both the WMT 2014 English-French (EN-FE) and English-German (EN-DE) data

sets, while this paper uses only the WMT 2014 EN-DE set and the IWSLT 2014 EN-DE and EN-FR data sets. This model is discussed later in the paper.

Works in the field of NMT recommend a particular focus on the encoder. Analysis by Domhan [11] poses two questions: what type of attention is needed, and where. In this analysis, self-attention had a higher correspondence with accuracy when placed in the encoder section of the architecture than the decoder, even claiming that the decoder, when replaced with a CNN or RNN, retained the same accuracy with little to no loss in robustness. Imamura, Fujita, and Sumita's [12] study shows that the current paradigm of using high-volume sets of parallel corpora are sufficient for decoders but are unreliable for the encoder. These conclusions encourage further research in the manipulation of position and design of the encoder and attention mechanisms within them.

III. Architecture

The Transformer architectures proposed by Vaswani et al. [3], seen in Figure 1, inspires this paper's work. We have made modifications to this architecture, to make it more efficient. However, our modifications can be applied to any encoder-decoder based model and is architecture-agnostic. These alterations follow from the following two hypotheses.

1) Reduction in the number of required sequential operations throughout the encoder section is likely to reduce training time without reducing performance.
2) Replacing the subsequent encoder attention stack is expected to result in discarding of inter-dependencies, and possibly incorrect, assumptions of encoder attention mechanisms and layers, improving performance.

For simplification, but without loss of generalization, this paper discusses the use and modification of Transformer based-models. The original Transformer model is composed of stacked self-attention layers. These self-attention mechanisms compare and relate multiple positions of one sequence in order to find a representation of itself. In Figure 1, we see such attention layers, one working on the input embedding, another on the output embedding, and the third on the both the input and the output embeddings. Each of these layers contains two main sub-layers including multi-head self attention, which feeds a simple feed-forward network, and a final layer of normalization. Around each of the main sub-layers, a skip or residual connection [13] is also used. This same structure is used in the decoder with an attention mask to avoid attending to subsequent positions.

The attention mechanism used by Vaswani et al. [3] can be thought of as a function that maps a query and set of key-value pairs to an output. The query, keys, values and output are all vectors. The output is obtained as a weighted sum of the values. The weight given to a value is learned by the system by considering how compatible the query is to the corresponding key. The particular form of attention used is called *scaled dot-product attention*. This is due to the mechanism being homologous to a scaled version of the multiplicative attention

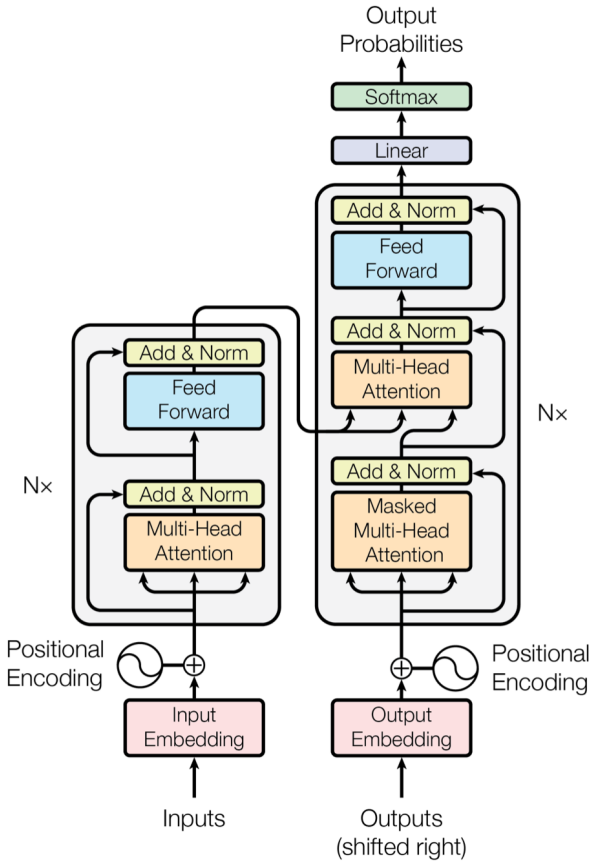

Fig. 1. Transformer model as proposed by Vaswani et. al [3].

proposed by Luong, Pham, and Manning [14]. Several attention layers used in parallel constitute what is called *multi-head attention*.

A brief description the proposed modifications of this architecture is discussed below.

A. Parallel Encoding Branches

A motivation for creating the Transformer model was the sluggish training and generation times of other common sequence-to-sequence models such as RNNs and CNNs [3]. This was done by simplifying and limiting sequential operations and computational requirements while also increasing the model's ability to exploit current hardware architecture. This paper proposes that removal of the previously stacked branches of the encoder (there is a stack of N encoder and other blocks on the left side of Figure 1), parallelizing these separate encoder 'trees', and incorporating their learned results for the decoder, will further eliminate sequential steps and accelerate learning within current sequence-to-sequence models. The architectures discussed are modeled in Figure 2.

Alterations to this parallel Transformer model were made and the following models were trained, tested, and are discussed in this paper:

- Additive Parallel Attention (APA),

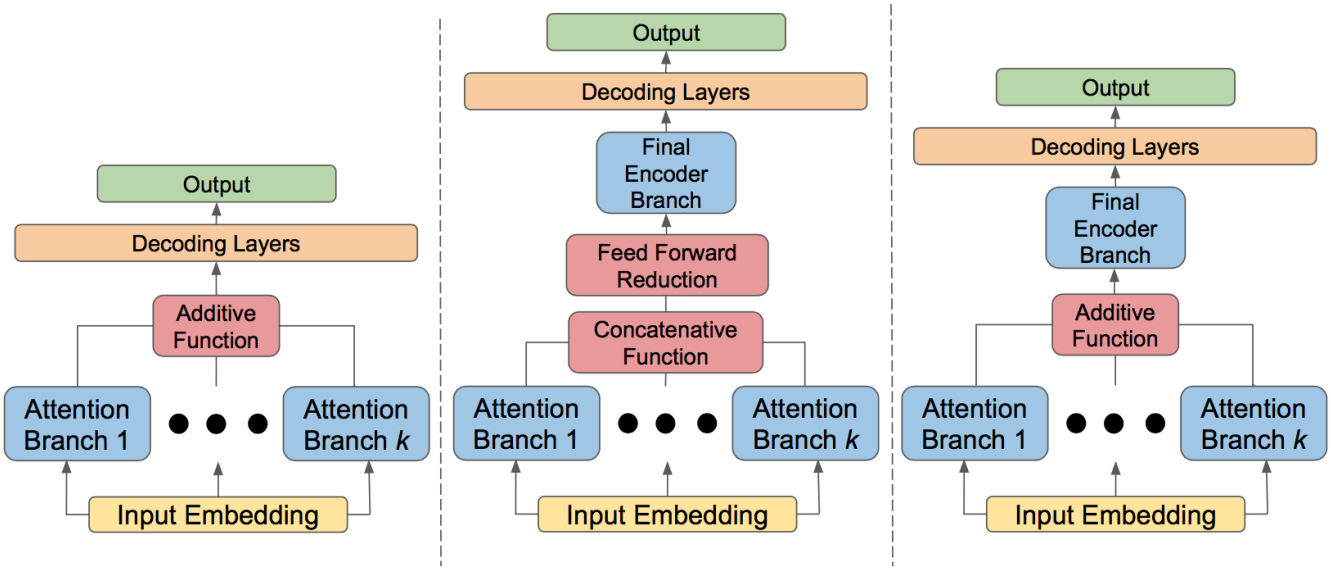

Fig. 2. From left to right we present three models. 1) APA: Parallel encoded Transformer that uses homologous stacks of encoding trees with random initialization, and addition of their learned attention. 2) ACPA: Attended Parallel encoding where the branches concatenate learned results, a feed-forward network reduces dimensionality, and a final encoder branch encodes the results. 3) AAPA: Attended Parallel Encoding Branches where a final encoding attention branch attends to the added learned results.

	No. Training Sentence Pairs		No. Testing Sentence Pairs	
Data Set	EN-DE	EN-FR	EN-DE	EN-FR
IWSLT [15]	197K	220K	628	622
WMT	4.5M	36M	3000	3000

TABLE I
DATA SETS USED FOR TRAINING AND TESTING FOR THE TRANSLATION TASKS OF ENGLISH-GERMAN AND ENGLISH-FRENCH. THE ENGLISH-FRENCH STATISTICS WERE INCLUDED FOR THE WMT DATA SET ALTHOUGH IT IS NOT DIRECTLY USED IN THE PAPER, AS IT WILL BE INCLUDED IN FUTURE WORK.

- Attended Concatenated Parallel Attention (ACPA), and
- Attended Additive Parallel Attention (AAPA).

B. Model Variations

Additive Parallel Attention (APA): We replace the entire stack of (multi-head attention, add and normalize, feed forward, add and normalize) repeated N times on the original Transformer architecture on the left column, on the input side. We instead have several such attention sub-networks in parallel. The output layers of these networks contain attention embeddings for the input. The values at the output layers among the stacks are added. This model is seen to the left in Fig. 2.

Attended Concatenated Parallel Attention (ACPA): This approach is similar to APA and AAPA, but the values at the output layers of the attention sub-networks are concatenated instead of being added. This model is seen in the middle of Fig. 2.

Attended Additive Parallel Attention (AAPA): This model is built similarly to the APA model. However, it removes one of the parallel stacks and uses it as a final sequential attention mechanism over the additive results. This model is seen to the right in Fig. 2.

When incorporating the results of the parallel encoding branches, two models of thought are pursued: additive and concatenation. The APA and AAPA models directly add the results of all encoding branches, whereas the ACPA models concatenate all encoding results and use a simple non-linear layer to learn a dimension-reduction among all attention branches. The attended parts of both the ACPA and AAPA models incorporate a final attention layer over all encoding branches before they are sent to the decoding layers.

IV. Experiments and Evaluation

All proposed architectures including the base Transformer model [3] are trained over the International Workshop on Spoken Language Translation (IWSLT) 2016 corpus and tested similarly over the IWSLT 2014 test corpus [15]. The training corpus includes over 200,000 parallel sentence pairs, and 4 million tokens for each language. The testing set contains 1,250 sentences, and 20-30 thousand tokens for French and German. This paper also performed experiments over the larger WMT data set including 4.5 and 36 million training sentence pairs for the EN-DE and EN-FR tasks respectively. The testing set for these experiments was the standard Newstest 2014 test set including around 3000 sentence pairs for each language task. These statistics are noted in Table I. The sentence pairs range in length from one to sixty tokens to get

Model	BLEU		Single GPU Run-Time (s)	
	EN-DE	EN-FR	EN-DE	EN-FR
Transformer as proposed by Vaswani et al. [3]	47.57 ± 4.97	56.15 ± 0.42	8052.19	9480.70
Attended Additive Parallel Attention 5 Parallel Branches (AAPA)	**57.05** ± 0.45	**63.26** ± 0.43	8158.26	9596.08
Attended Additive Parallel Attention 4 Branches	56.22 ± 0.63	62.68 ± 0.25	7805.73	9114.84
Attended Additive Parallel Attention 3 Branches	56.68 ± 0.47	62.75 ± 0.35	7412.81	8686.92
Attended Additive Parallel Attention 2 Branches	55.94 ± 0.01	61.24 ± 0.53	**6998.18**	**8228.14**
Attended Concatenated Parallel Attention (ACPA)	48.67 ± 4.47	62.31 ± 0.21	8186.77	9710.70

TABLE II

MODEL COMPARISON FOR TEST RESULTS OVER THE **IWSLT 2014 TEST SET**. THE BLEU SCORE IS GIVEN AS AN AVERAGE OF THE FINAL EPOCH OVER MULTIPLE RUNS WHERE ALSO A STANDARD DEVIATION (SD) IS ALSO GIVEN. BY REDUCING THE NUMBER OF PARALLEL BRANCHES IN THE ENCODER, THE MODEL CAN MAINTAIN HIGH ACCURACY AND REDUCE RUN-TIME. ALL OF THESE MODELS WERE DEVELOPED IN IN THE OPENNMT TOOLKIT [18].

Model	BLEU	Single GPU Run-Time (s)
Transformer Large [3]	60.95	168,806.61
Attended Additive Parallel Attention Large 7 Parallel Branches (AAPA)	61.98	173,163.03
Transformer	61.00	138,032.33
Attended Additive Parallel Attention 5 Parallel Branches	62.69	141,041.74
Attended Additive Parallel Attention 4 Branches	**62.77**	133,374.33
Attended Additive Parallel Attention 3 Branches	62.07	123,929.10
Attended Additive Parallel Attention 2 Branches	62.59	**116,450.75**
Attended Concatenated Parallel Attention (ACPA)	60.32	142,363.06

TABLE III

MODEL COMPARISON FOR TEST RESULTS OVER THE LARGER NMT ENGLISH-GERMAN TEST SET. ALL OF THESE MODELS WERE DEVELOPED IN IN THE OPENNMT TOOLKIT [18].

Model	BLEU	
	Cased	Uncased
Transformer Large [3]	**24.20** ± 0.081	**23.72** ± 0.005
Attended Additive Parallel Attention 3 Branches	23.90 ± 0.04	23.406 ± 0.04
Attended Additive Parallel Attention 2 Branches	23.794 ± 0.28	23.494 ± 0.02

TABLE IV

MODEL COMPARISON OVER THE **WMT 2016** ENGLISH-GERMAN TRANSLATION TASK WITH OUR MODELS IMPLEMENTED IN THE TENSOR2TENSOR [19] LIBRARY BY GOOGLE. EACH MODEL WAS TRAINED TO 250K TRAINING STEPS. ALTHOUGH OUR MODELS ARE COMPARABLE TO THE TRANSFORMER MODEL FOR THE WMT EN-DE TASK, THEY SURPASS TRANSFORMER FOR THE IWSLT 2014 TEST SET.

a full measure of the tested models and robustness to both short and long input.

Across all models, a greedy-decoding function for both training and testing time, the Kullback-Leibler divergence loss function, the Adam optimizer [16], and the number of training epochs (10) were kept constant. The training and testing were done using the NMT task of English to German (EN-DE) and IWSLT English to French and English to German translation and each network was trained using one graphics processing unit (GPU). The utilized machine GPU configuration was one NVIDIA GTX 1070.

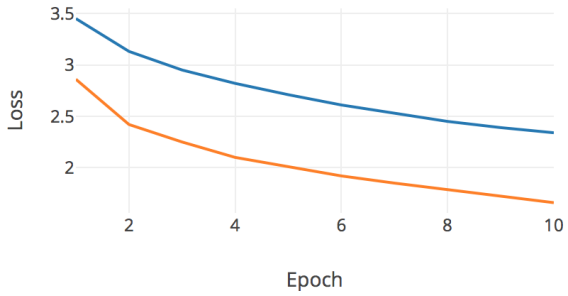

Fig. 3. This plot shows validation loss for both the Transformer model (blue) and our modified model (orange) over the IWSLT EN-DE task. The parallel encoder shows a consistently lower starting and end-training loss.

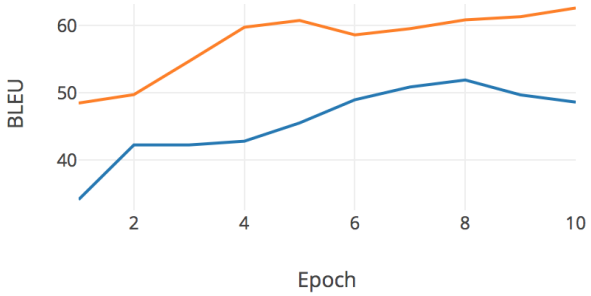

Fig. 4. This plot shows validation BLEU metric score for both the Transformer model (blue) and our modified model (orange) over the IWSLT EN-DE task. The parallel encoder shows a consistently higher BLEU score and shows linear increase while the Transformer shows some plateauing in later epochs.

For the assessment of each model and translation task this paper uses the bilingual evaluation understudy (BLEU) metric [17]. This is a modified precision calculation using n-grams such as unigram, grouped unigrams, and bigrams. The BLEU metric claims to have a high correlation to translation quality judgments made by humans. BLEU computes scores for individual sentences by comparing them with good quality reference translations. The individual scores are averaged over the the entire corpus, without taking intelligibility or grammatical correctness into account.

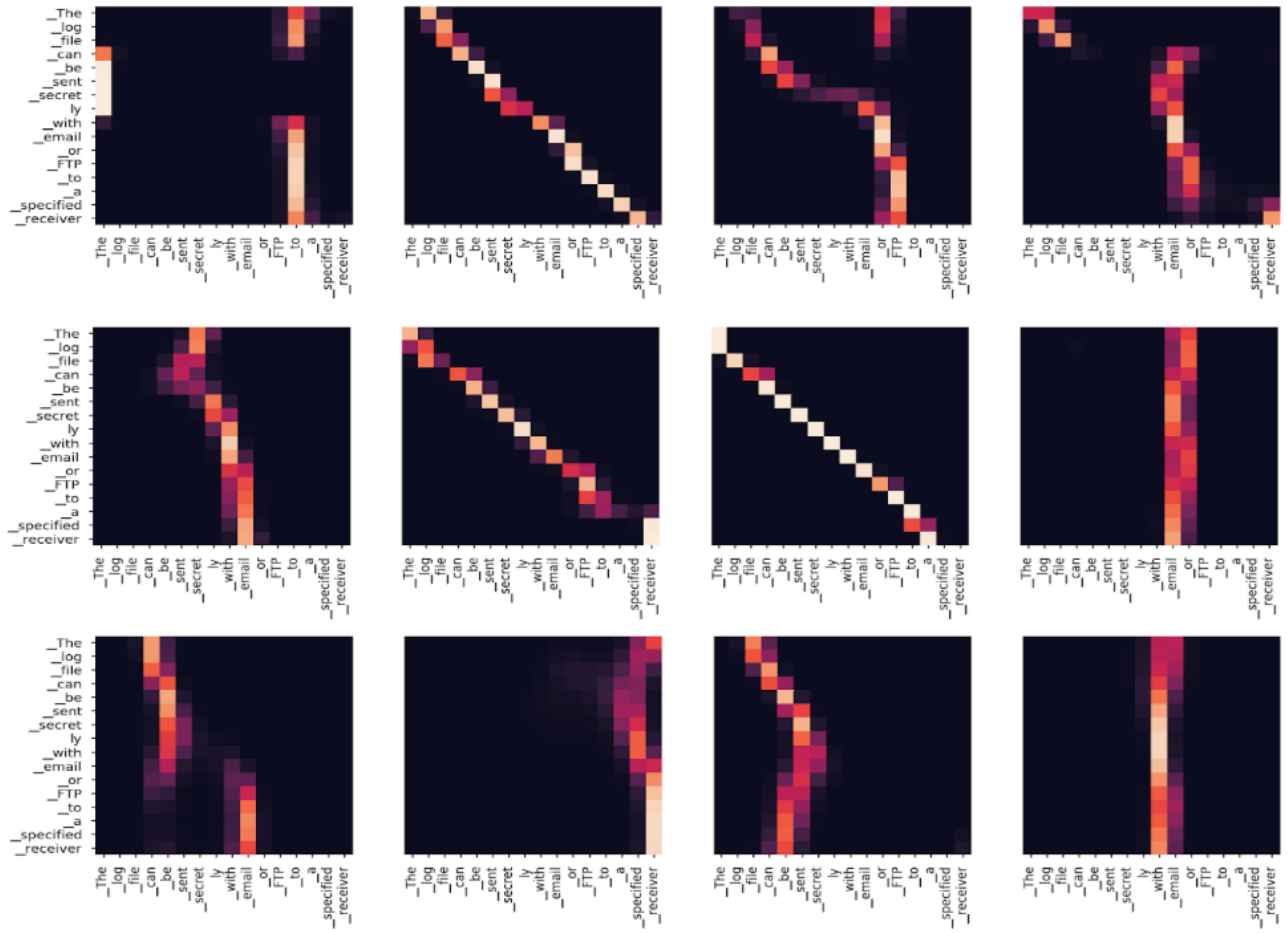

Fig. 5. Visualization of multi-head attention weights in encoder branches 0, 2, and 4. Although each receives the same input embedding, through random initialization, each learns different focuses.

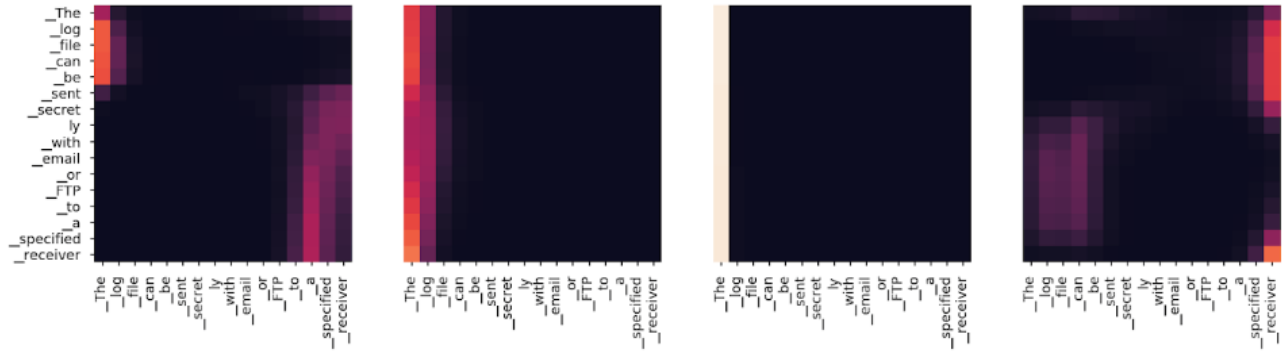

Fig. 6. Visualization of the weights for the final encoder that attends over all other encoding branches. This encoder's weights are relatively light, abstract, and have less obvious patterns when compared to the individual encoding branches.

V. Results

A. Attention Visualization

One concern during early hypothesis testing was that if each attention branch looks at the same input, that each one would learn to focus on the same properties of the original embedding. However, through visualization of each attention layer, it is obvious that regardless of the same input, the encoder branches through random initialization learn different focuses as seen in Figure 5. The final branch for the attended models however would learn very light to no attention weights as seen in Figure 6. This is one area of research this group wishes to pursue in the future.

B. Machine Translation

Table II shows that the AAPA model consistently performed on average nearly ten points higher in the BLEU metric on the English to German translation task on the IWSLT 2014 test set. It also performed very well on the English to French translation task.

On the much larger WMT English-German test set, all our models achieve better results then Vaswani et al. [3]. Our model with five parallel encoding branches has a BLEU score of 62.69 compared to 60.95 and 61.00 for the two Transformers shown in Table III. Our approach also takes considerably less time than the large Transformer model with a stack of eight encoder attention heads, although it is a little slower than the smaller Transformer model reported by Vaswani et al. [3]. In terms of the BLEU metric, we establish state-of-the-art performance for both EN-DE and EN-FR translation considering the IWSLT 2014, and comparable results for the WMT data sets. Since our results came up very good, surpassing state of the art for the IWSLT 2014 dataset, we ran our experiments multiple times to ensure the results are correct.

During the Transformer and attended parallel model's training lifetime, it can be seen that loss was consistently lower for our modified parallel model with five parallel stacks as seen in Figure 3. In this task, loss doesn't always correspond to a higher metric, in this case our model also shows a continuous higher score in the BLEU metric over the validation set while the Transformer shows signs of plateauing early on Figure 4.

However, our parallelized model did have a slightly higher training time over a single GPU. One final experiment conducted to improve this drawback, also seen in the same table, is the reduction of number of parallel branches in the encoder. By reducing the number incrementally, our BLEU score stays equivalent to higher perplexity layers, but linearly reduces the run-time.

VI. Conclusions

In step with the goals of the original Transformer, this work continued to pursue the removal of sequential operations within attention-based translation models. Although dependent on choice of tool-kit implementation as shown in Table IV, this new parallelized Transformer model reaches a new state-of-the-art in machine translation and provides multiple new directions for future research. It also shows through random initialization that attention mechanisms can learn different focuses and that by eliminating possibly negative inter-dependencies among them, superior results can be obtained.

VII. Acknowledgement

This material is based upon work supported by the National Science Foundation under Grant No. 1659788. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] D. Jurafsky, "Speech & Language Processing," Pearson Education, 2000.
[2] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," CoRR abs/1409.0473, 2014.
[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," In Advances in Neural Information Processing Systems, 2017.
[4] K. Ahmed, N. S. Keskar, and R. Socher, "Weighted Transformer Network for Machine Translation," CoRR abs/1711.02132, 2017.
[5] J. L. Elman, "Finding structure in time." Cognitive Science 14, no. 2 179-211, 1990.
[6] LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P., "Gradient-based learning applied to document recognition," Proceedings of the IEEE, 86(11), pp.2278-2324, 1998.
[7] N. Kalchbrenner, L. Espeholt, K. Simonyan, A. van den Oord, A. Graves, and K. Kavukcuoglu. Neural machine translation in linear time. arXiv preprint arXiv:1610.10099v2, 2017.
[8] J. Zhou, Y. Cao, X. Wang, P. Li, and W. Xu. Deep recurrent models with fast-forward connections for neural machine translation. CoRR, 2016.
[9] Y. Wu, M. Schuster, Z. Chen et al. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144, 2016.
[10] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin. Convolutional sequence to sequence learning. arXiv preprint arXiv:1705.03122v2, 2017.
[11] T. Domhan, "How Much Attention Do You Need? A Granular Analysis of Neural Machine Translation Architectures." In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol. 1, pp. 1799-1808. 2018.
[12] K. Imamura, A. Fujita, and E. Sumita. "Enhancement of Encoder and Attention Using Target Monolingual Corpora in Neural Machine Translation." In Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, pp. 55-63. 2018.
[13] H. Kaiming, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778, 2016.
[14] T. Luong, P. Hieu, and D. M. Christopher, "Effective Approaches to Attention-based Neural Machine Translation." In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1412-1421. 2015.
[15] M. Cettolo, C. Girardi, and M. Federico, "WIT3: Web Inventory of Transcribed and Translated Talks," In Proc. of European Association for Machine Translation, pp. 261-268, Trento, Italy, 2012.
[16] D. P. Kingma, J. Ba, "Adam: A Method for Stochastic Optimization," In proceedings of the 3rd International Conference for Learning Representations, San Diego, 2014.
[17] K. Papineni, s. Roukos, T. Ward, and W. J. Zhu, "BLEU: a method for automatic evaluation of machine translation," In Proceedings of the 40th annual meeting on Association for Computational Linguistics, 311-318, 2002.
[18] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush. "OpenNMT: Open-Source Toolkit for Neural Machine Translation." Proceedings of ACL 2017, System Demonstrations (2017): 67-72.
[19] A. Vaswani, S. Bengio, E. Brevdo, F. Chollet, A. N. Gomez, S. Gouws, L. Jones et al. "Tensor2Tensor for Neural Machine Translation." Vol. 1: MT Researchers Track (2018): 193.