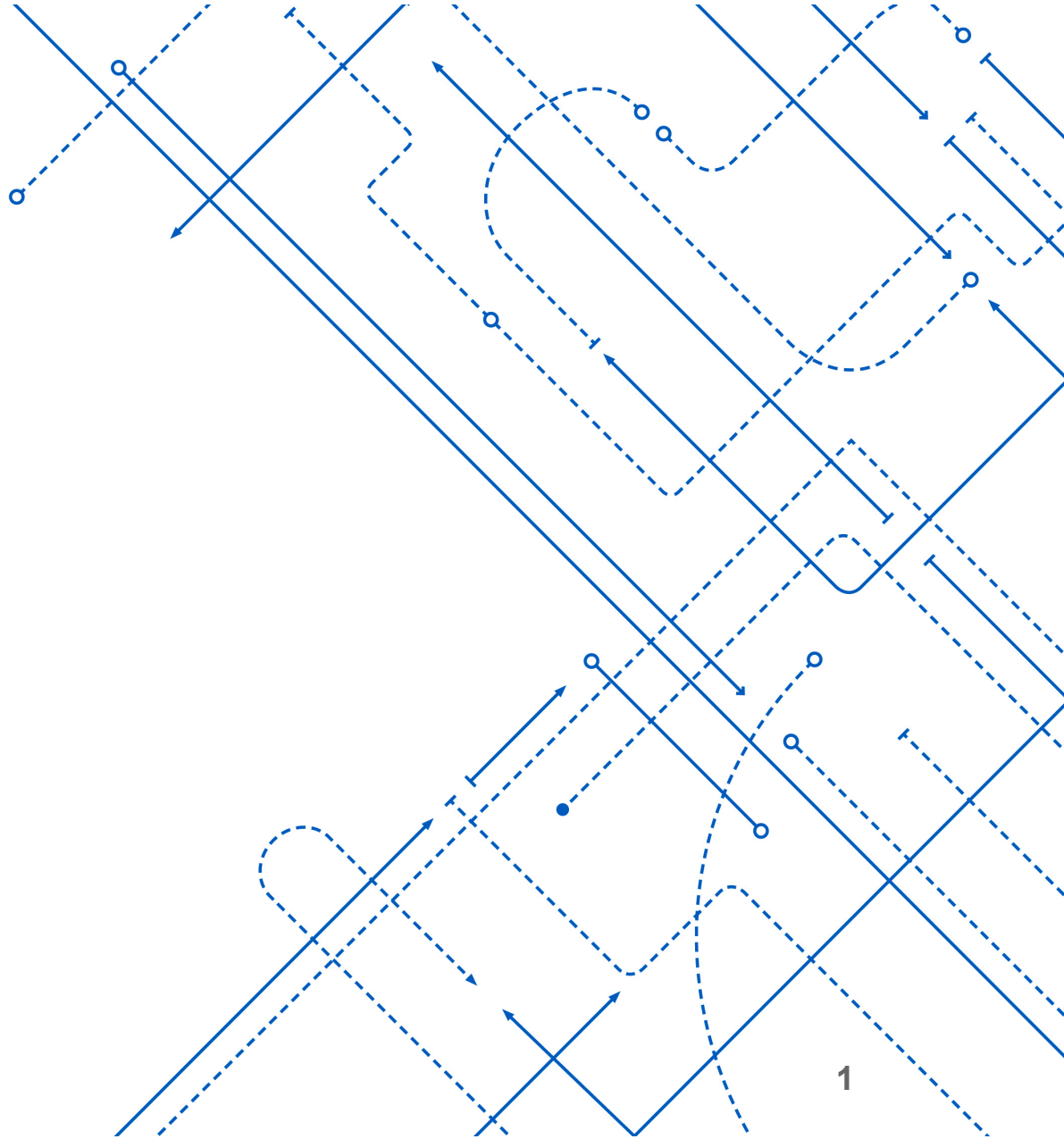
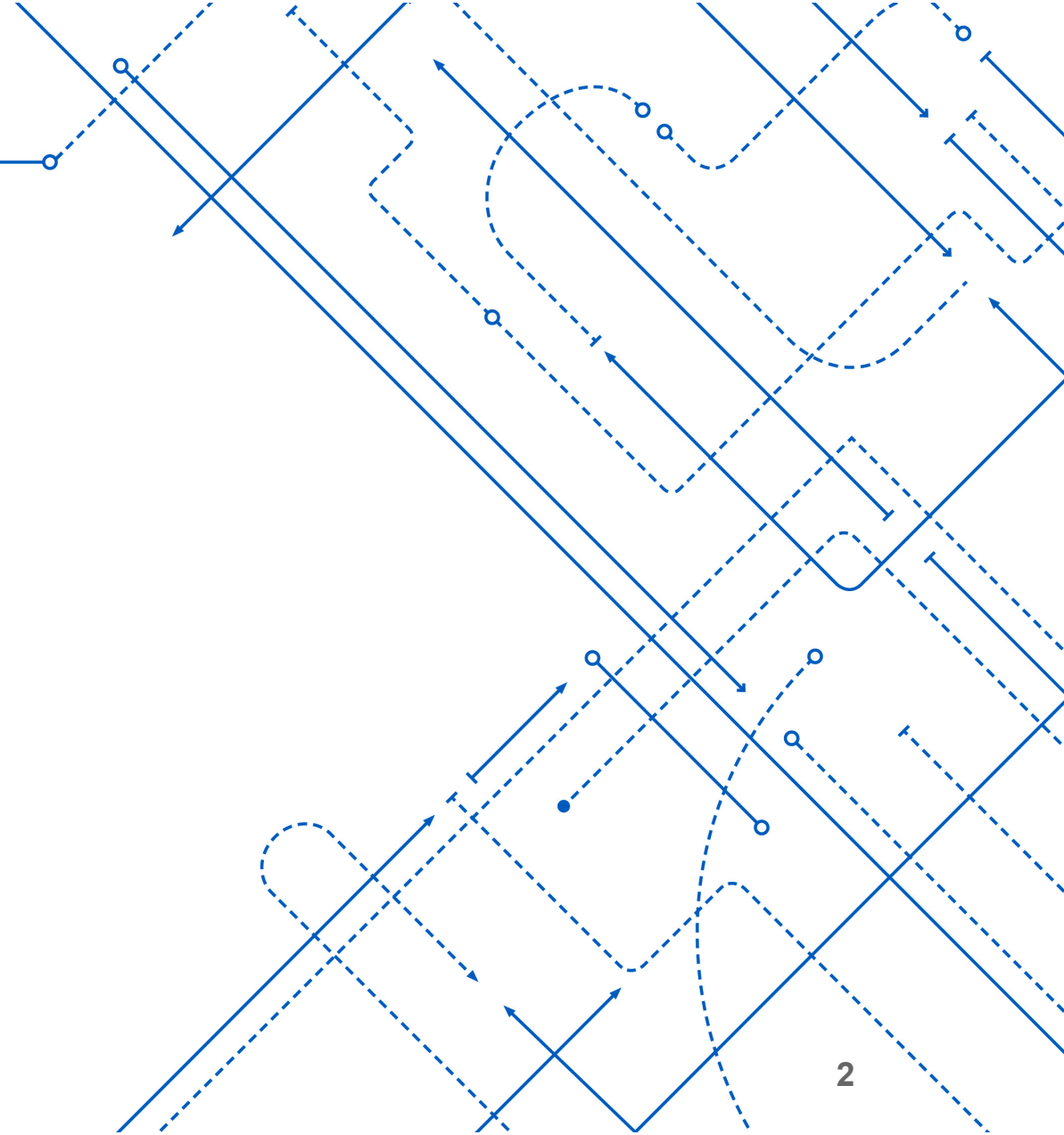
ATTENTION TO SHAKESPEARE !

- Sambo Dutta (50318667) ,
- Soumita Das (50320170)

University at Buffalo
Department of Computer Science and Engineering
School of Engineering and Applied Sciences





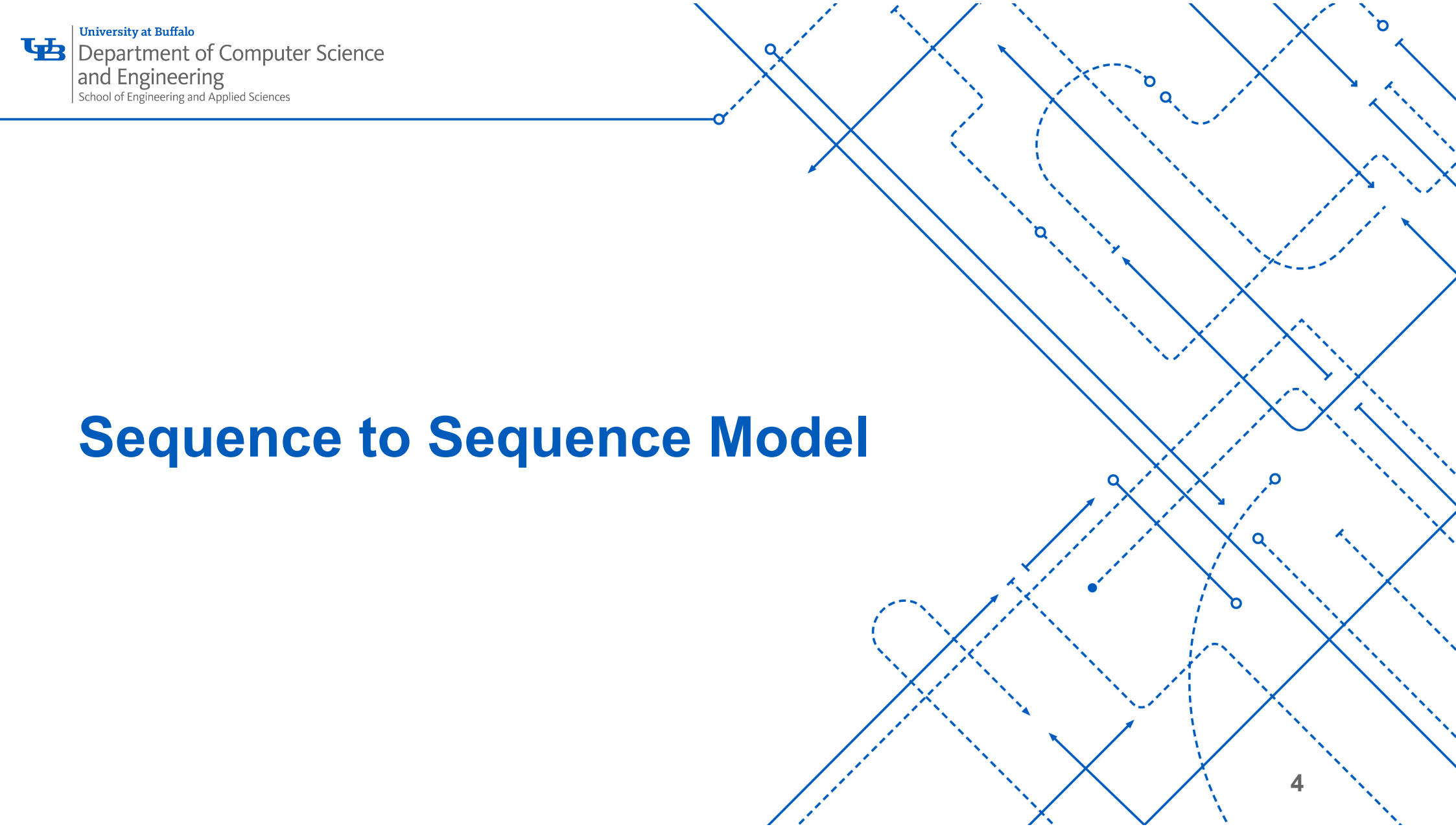
Introduction….

- In this project we address the task of using Neural Machine Translation to apply artistic style transfer.

- *Seq2Seq* model leveraging RNN or LSTM has been used previously to convert modern english language to *Shakespearean* style.

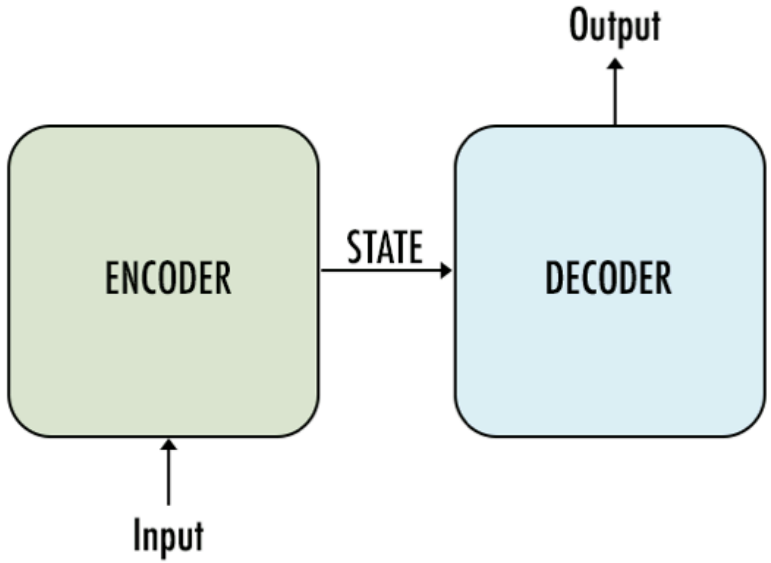
- We aim to further improve the speed and quality of this style transfer using *Transformer* model.

- Lastly we present a slightly modified version of Transformer that we call the 'Pizza' structure and discuss its advantages.

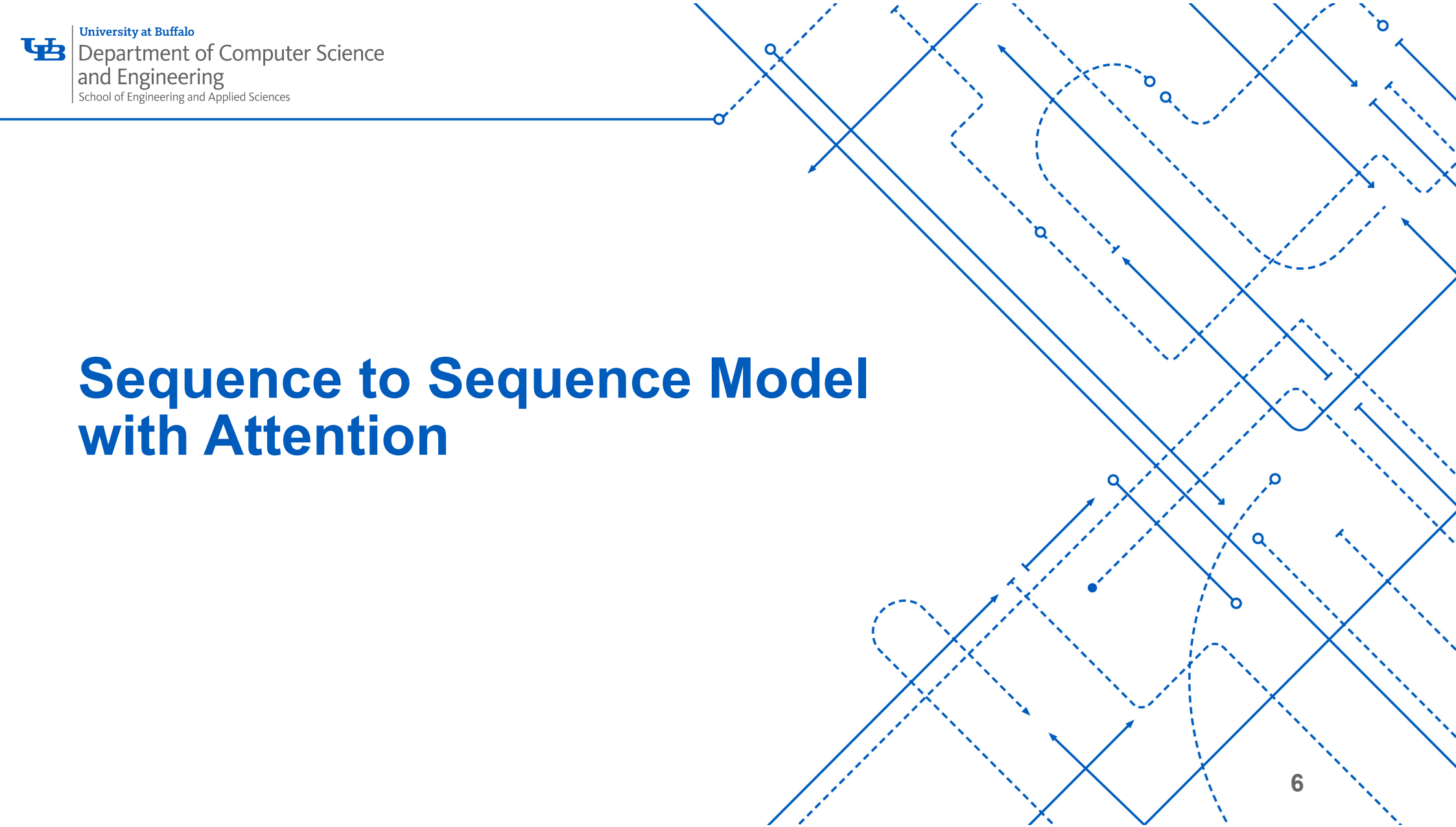


Sequence to Sequence Model

- An encoder component comprising of RNNs like LSTM or GRU encodes the entire input sequence to a hidden state vector.
- This vector aims to encapsulate the information for all input elements.
- The decoder component similarly accepts a hidden state from the previous unit and produces output as well as its own hidden state for the next word.

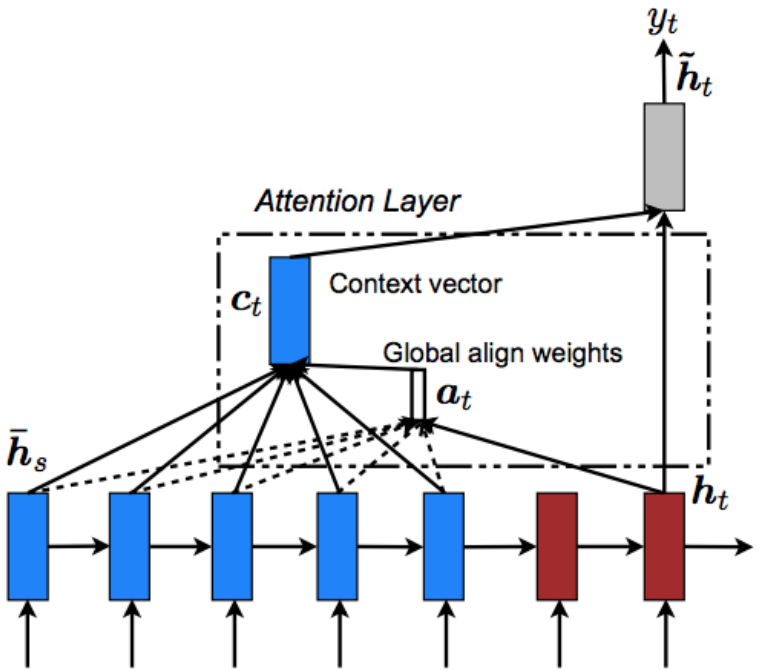


University at Buffalo
Department of Computer Science and Engineering
School of Engineering and Applied Sciences



Sequence to Sequence Model with Attention

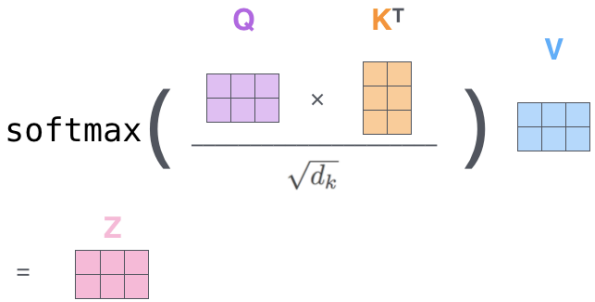
- When decoding a sentence, we want the model to pay *attention* to only a part of the sentence and not the entire one.
- **Soft** Attention: the alignment weights are learned and placed "softly" over all patches in the input.
- **Hard** Attention: only selects one patch of the input to attend to at a time.



Transformer !

Key Points

- Dispenses any sequential neural networks like RNN or CNN.
- Multi-layered encoder and decoder.
- Consists only of Self-Attention and Feed-Forward Neural Network.
- Multi-headed Attention: the one with 8 heads !
- The *Masked* multi-headed attention prevents the decoder to look at the future input tokens when decoding a certain word.
- Self-attention is the method the Transformer uses to bake the understanding of other relevant words into the one being currently processed.



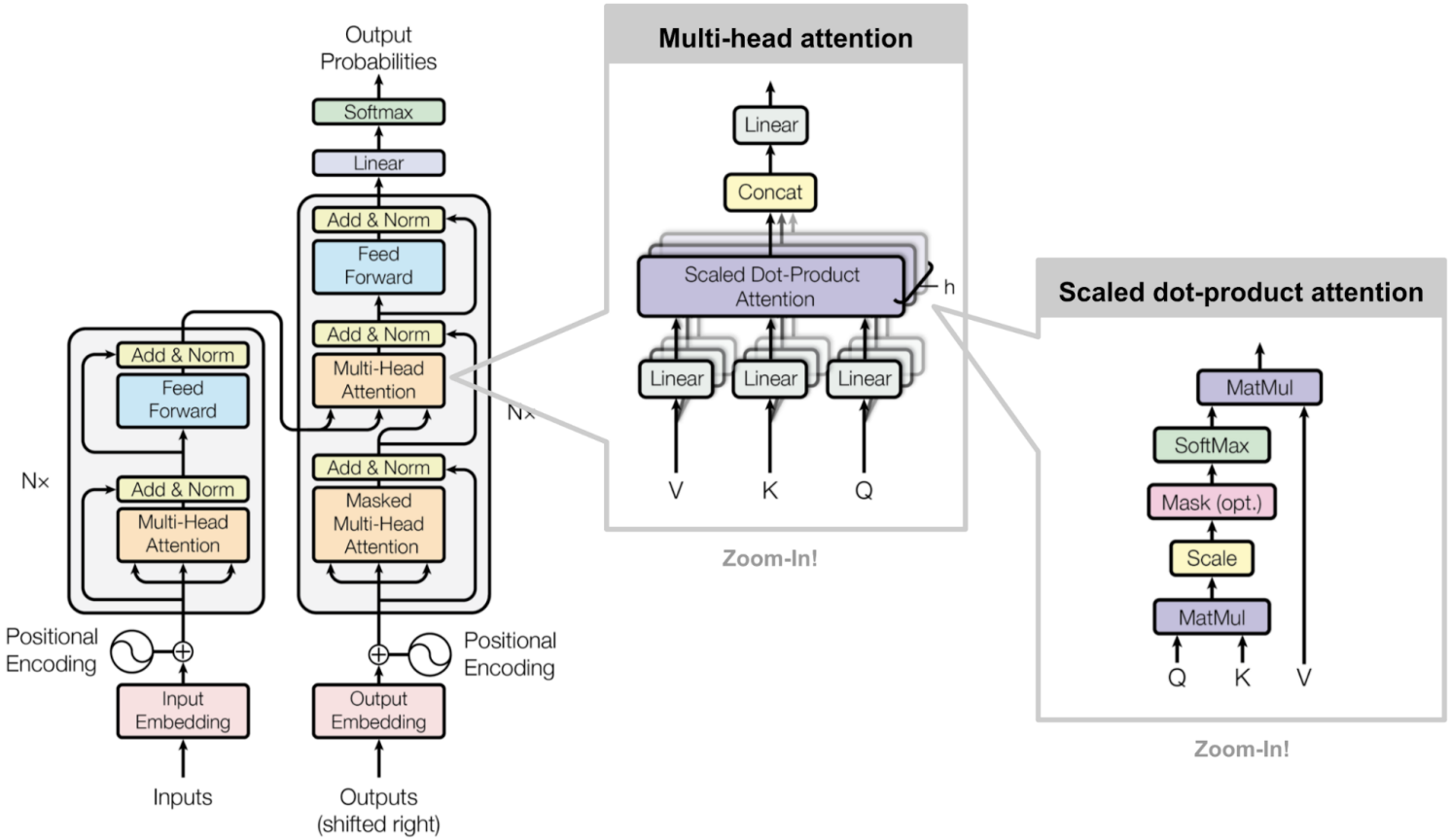
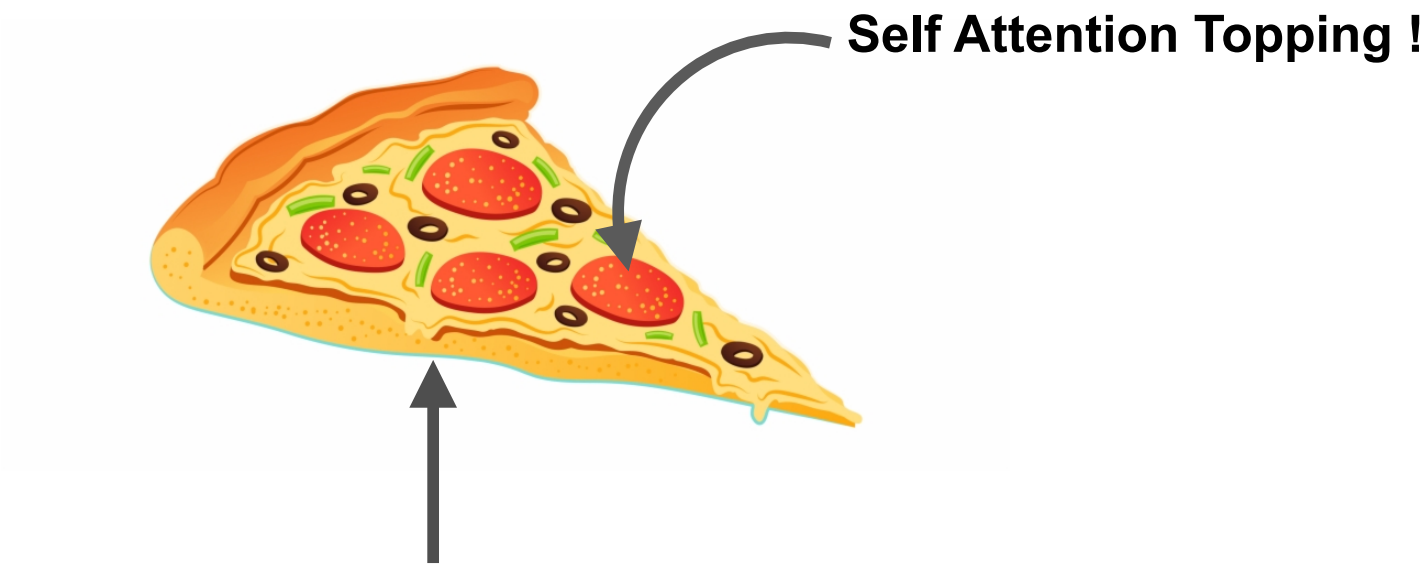


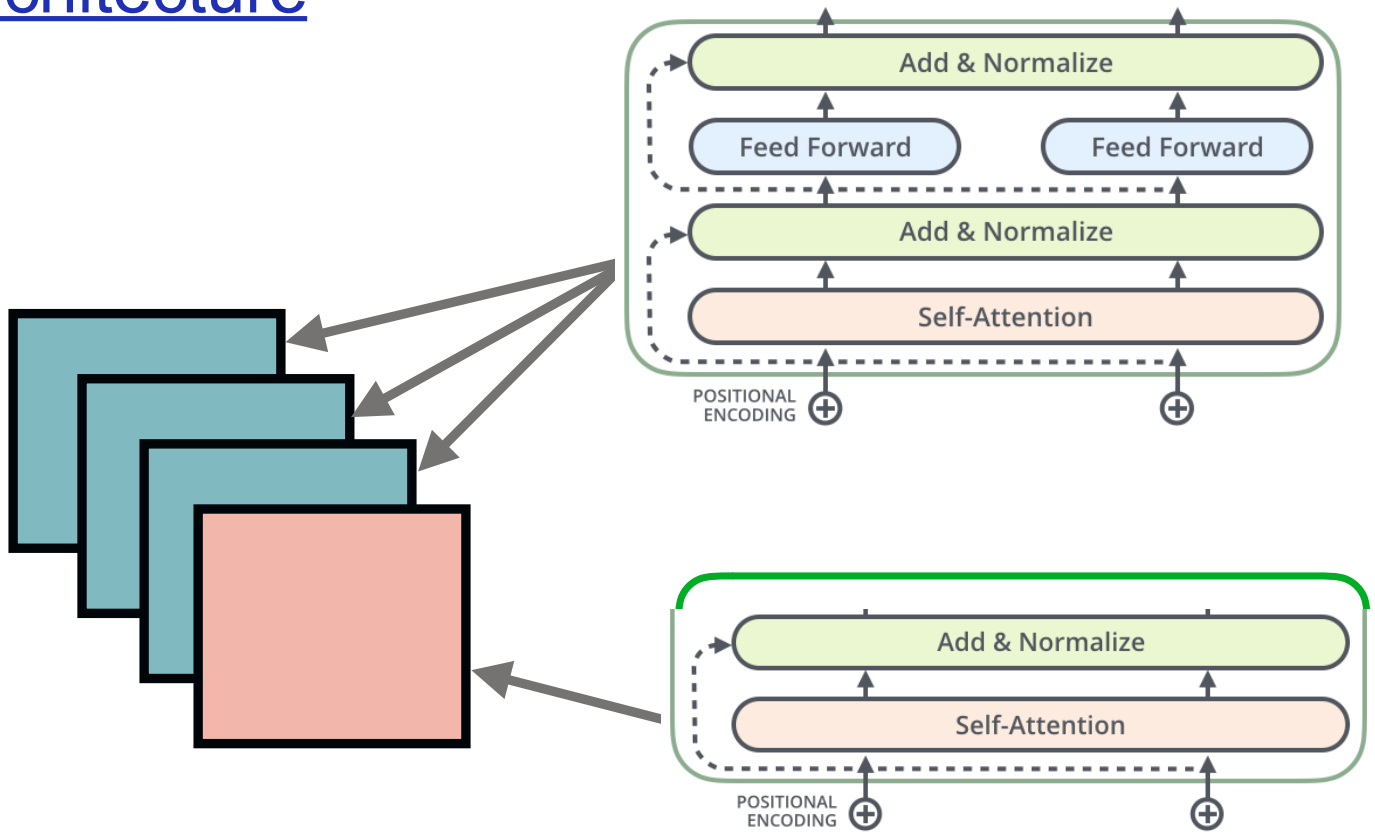
Figure 1: Architecture of Transformer (Image source: Fig 1 & 2 in Vaswani, et al., 2017.)

The Modified Transformer

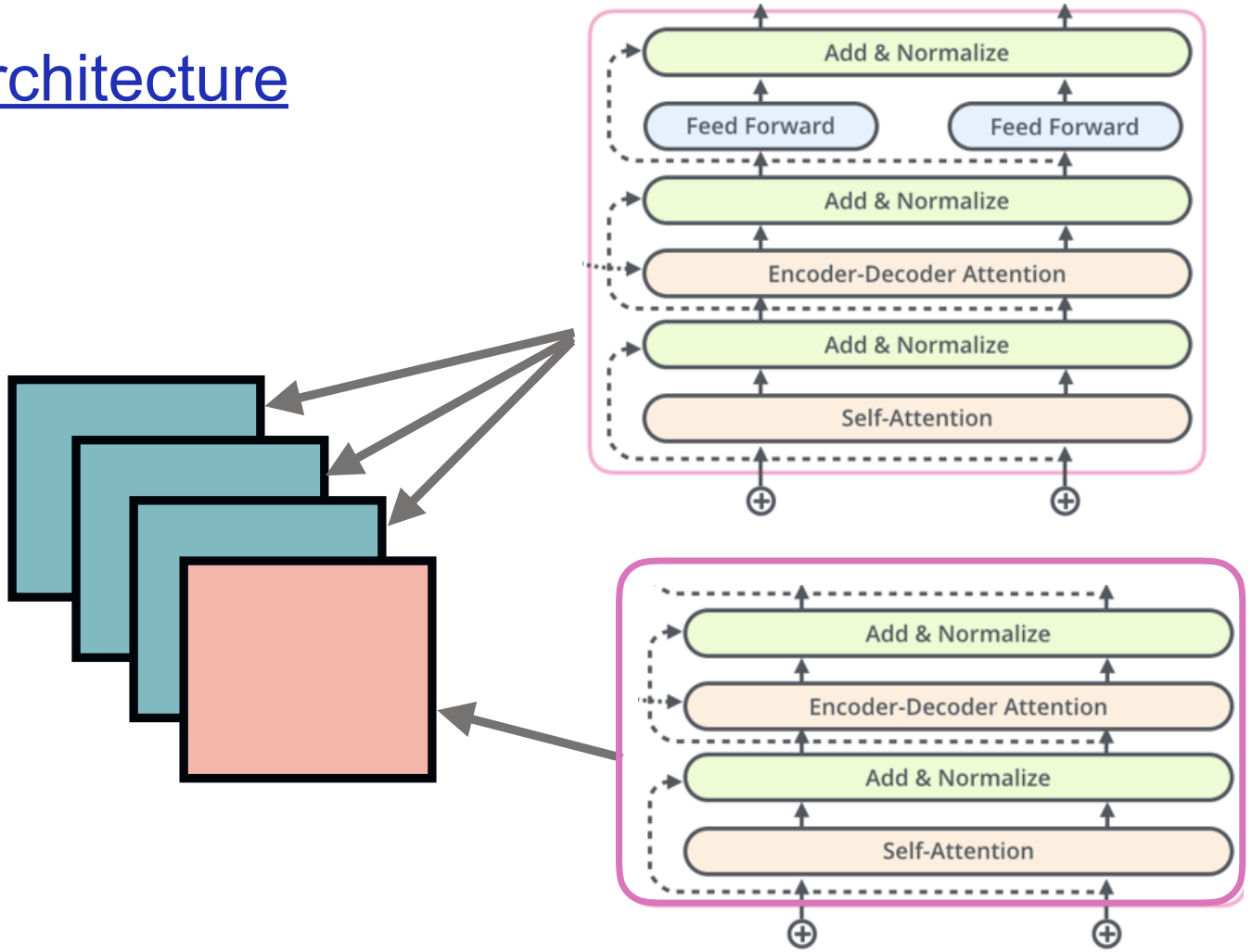
What is the ‘Pizza’ Structure ?



Encoder Architecture

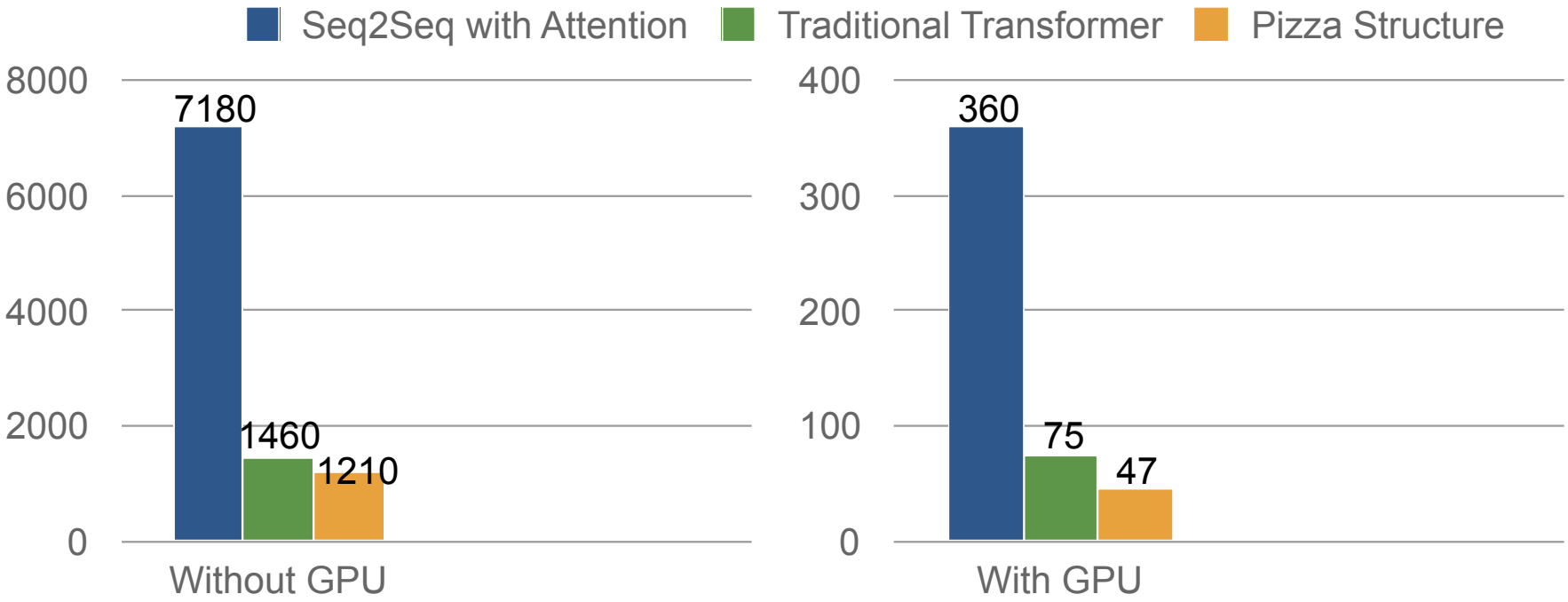


Decoder Architecture



Advantages:

- The topping layer has 3 times less parameters than the traditional stack layers.
- The training time reduces significantly.



Results :

Original Text	Seq2Seq with Attention	Transformer	Pizza Transformer
have you killed tybalt ?	hast thou slain tybalt ?	hast thou slain tybalt ?	hast thou slain tybalt ?
i will hit you !	i will , beat thee !	i will strike thee !	i will strike thee !
What is wrong with you ?	what is amiss you ?	what is your passion ?	what is thy wrong with thee ?
what fight was here ?	what fray was here ?	what fray was here ?	what fray was here ?
what do you want ?	what would you ?	what would you ?	what s your will ?
i ll bite you by the ear for that joke .	i will tell thee by the joyful office for that jest .	i ll bite you by the ear of that jest .	i ll bite thee for that kind of jest .
BLEU Score:	**0.497**	**0.508**	**0.517**

Conclusion:

- How attention plays a major role.
- Style Transfer and Translation tasks performs better using only Self Attention.
- Transformers make the task more parallelizable .
- Introduction to the Pizza structure.
- Advantages over traditional transformer:
    1. Less parameters
    2. Faster training speed
    3. Better accuracy

References:

[1] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'14). MIT Press, Cambridge, MA, USA, 3104–3112.

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. NIPS.

[3] N. Kalchbrenner and P. Blunsom. Recurrent continuous translation models. In EMNLP, 2013.

[4] Bahdanau, Dzmitry & Cho, Kyunghyun & Bengio, Y (2014). Neural Machine Translation by Jointly Learning to Align and Translate. ArXiv. 1409.

[5] Luong, Minh-Thang & Pham, Hieu & Manning, Christoper. (2015). Effective Approaches to Attention- based Neural Machine Translation.

[6] Weng, Lilian. Attention? Attention! (2018). https://lilianweng.github.io/lil-log/2018/06/24/attention- attention.html

[7] Alammar, Jay. The Illustrated Transformer. (2018) http://jalammar.github.io/illustrated-transformer/ [8] Xu, W., Ritter, A., Dolan, W.B., Grishman, R., Cherry, C. (2012). Paraphrasing for Style. COLING.

Thank You !