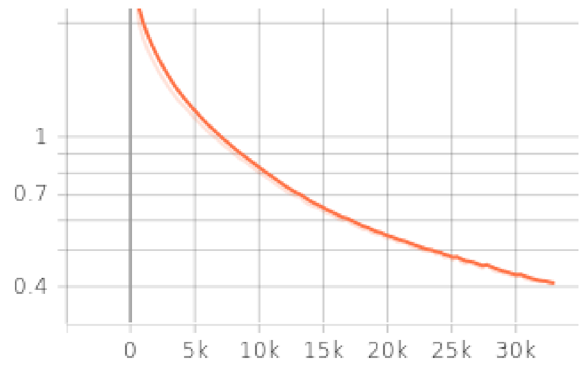
Asadullah Hill Galib
CSE 891-001: Deep Learning
Programming Assignment 3
November, 9

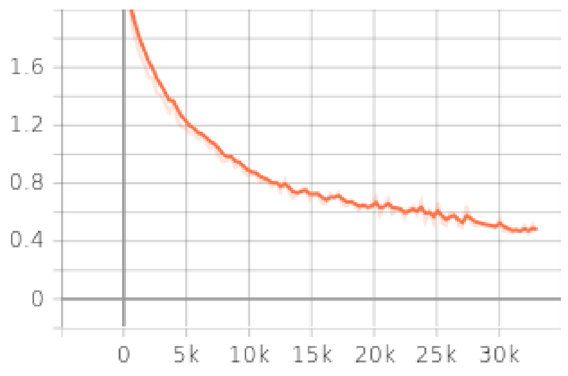
1 Gated Recurrent Unit

Training/Validation Loss Plots

Performance is better with the large dataset.

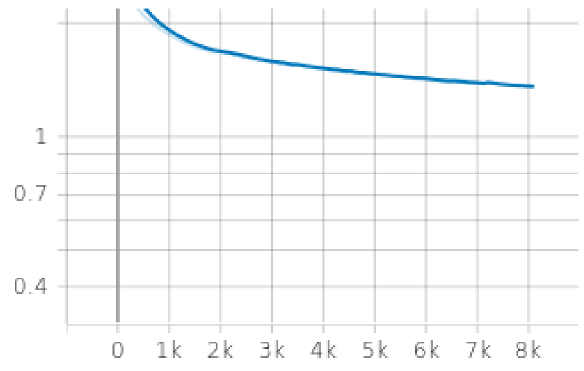


(a) Training Loss

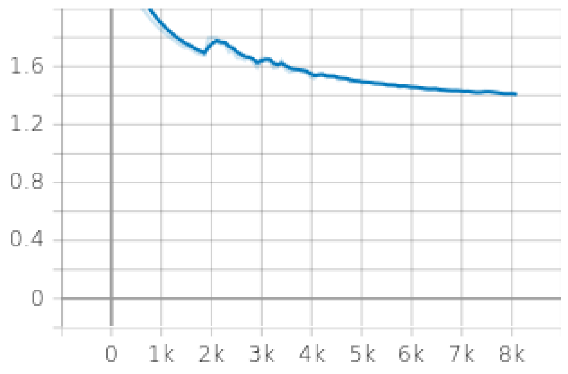


(b) Validation Loss

Figure 1: Training/Validation Loss Plots for GRU with large dataset



(a) Training Loss



(b) Validation Loss

Figure 2: Training/Validation Loss Plots for GRU with small dataset

Input Output Pairs with Failure Cases

source: the air conditioning is working
translated: ethay airway ondintiondfray isway orkingway

source: the quick brown fox
translated: ethay ucikicay onbray oxfay

source: deep learning is fun and mysterious
translated: eepway eanringlay isway unfay andway isteromentway

source: shining like the sun
translated: iningshay ikelay ethay unsay

source: breaking bad is a good one
translated: eakingbray adbay isway away oodgay oneway

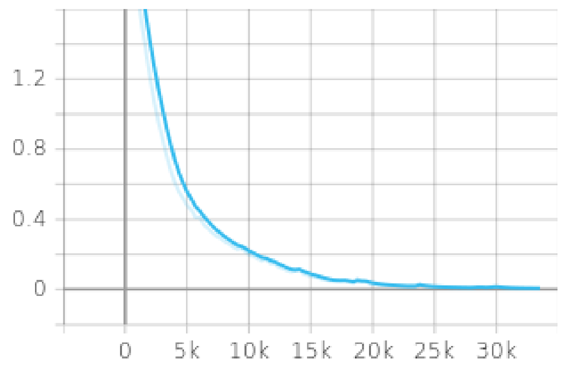
Failure Insights:

- The model can not translate long words properly: mysterious $\Longrightarrow$ isteromentway.
- It struggles with words starting with two consonants, it takes those starting consonants to the end: brown $\Longrightarrow$ onbray, the $\Longrightarrow$ ethay. (It might be a rule).
- It messes up character order, specially for long words: learning $\Longrightarrow$ eanringlay.

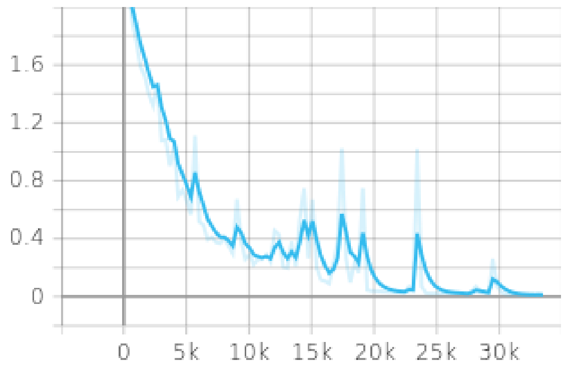
2 Additive Attention

Training/Validation Loss Plots

Additive Attention significantly outperforms the GRU-based encoder.

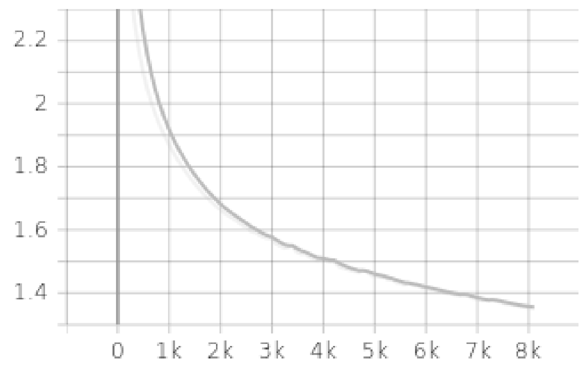


(a) Training Loss

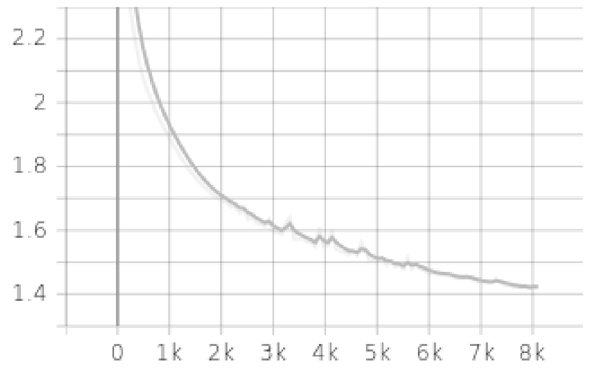


(b) Validation Loss

Figure 3: Training/Validation Loss Plots for Additive Attention with large dataset



(a) Training Loss



(b) Validation Loss

Figure 4: Training/Validation Loss Plots for Additive Attention with small dataset

Input Output Pairs with Failure Cases

source: the air conditioning is working
translated: ethay airway onditioningcay isway orkingway

source: the quick brown fox
translated: ethay uickqay ownbray oxfay

source: deep learning is fun and mysterious
translated: eepday earninglay isway unfay andway ysteriousmay

source: shining like the sun
translated: iningshay ikelay ethay unsay

source: breaking bad is a good one
translated: eakingbray adbay isway away oodgay oneway

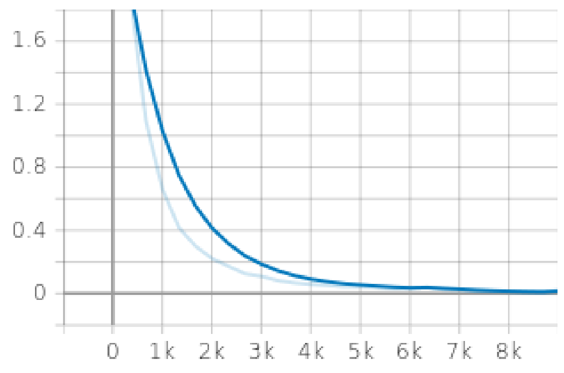
Failure Insights:

- It sometimes struggles with words starting with two consonants, it takes those starting consonants to the end: brown $\Longrightarrow$ onbray, the $\Longrightarrow$ ethay. (It might be a rule).

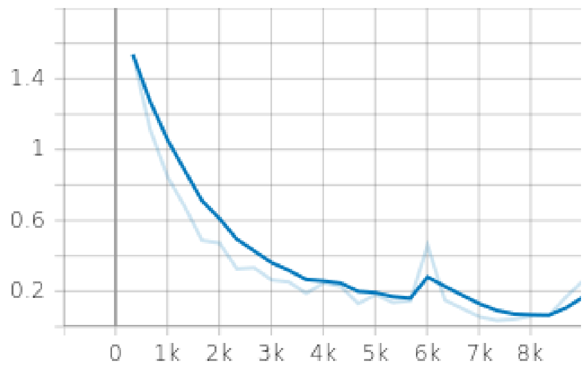
3 Scaled Dot Product Attention

I have applied layer normalization as mentioned in the paper.

Training/Validation Loss Plots

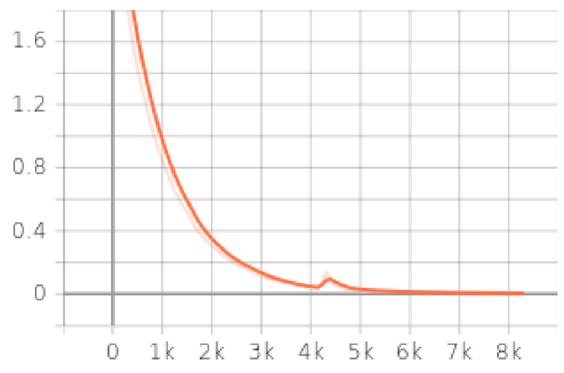


(a) Training Loss

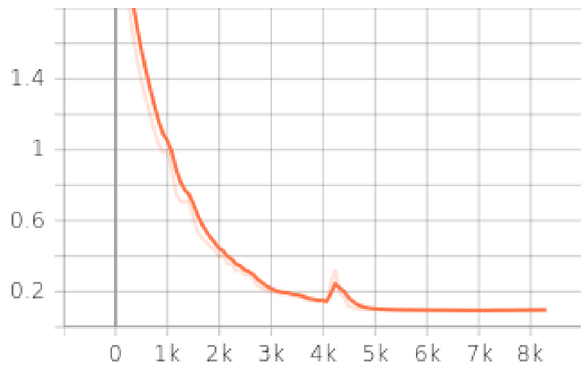


(b) Validation Loss

Figure 5: Training/Validation Loss Plots for Transformer with large dataset



(a) Training Loss



(b) Validation Loss

Figure 6: Training/Validation Loss Plots for Transformer with small dataset

3.5

The translation is far better than the GRU decoder, and very close to the additive attention modeling (val losses: GRU: 0.4851, Additive: 0.012, Transformer: 0.109). In terms of qualitative comparison, the additive attention decoder seems like more accurately translate words than the Transformer (after seeing the Test Sentences). Additive attention decoder almost translate all the words exactly, except words starting with two consonants. But, transformer struggles with other cases too: long words, ordering of word, words starting with vowels, etc.

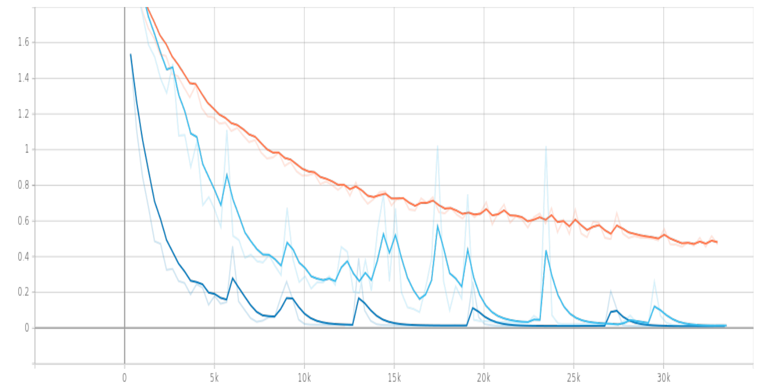
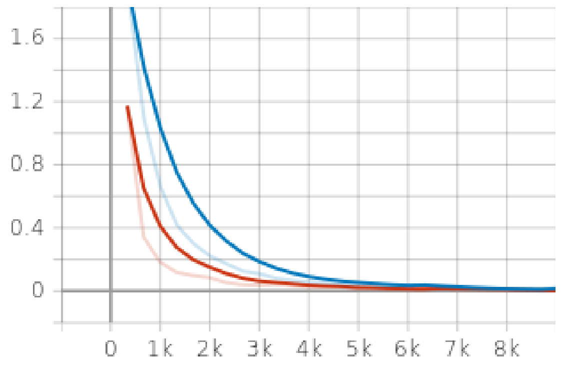


Figure 7: Comparison of decoders (GRU: orange, Additive Attention: Light Blue, Transformer: Blue)

3.6

Effect of model size: According to Figure 8, the more the model size, the better the generalization (as validation loss is smaller for hidden size of 64).


(a) Training Loss

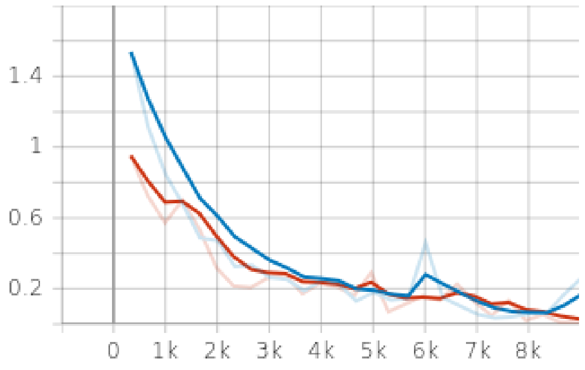
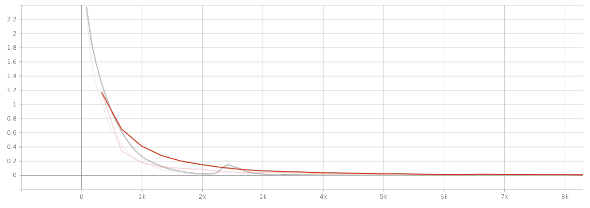

(b) Validation Loss

Figure 8: Effect of model size with the large dataset: hidden size 32: **blue**, hidden size: 64: **red**


(a) Training Loss

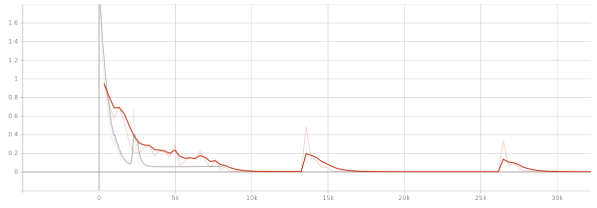

(b) Validation Loss

Figure 9: Effect of dataset size: small dataset: **gray**, large dataset: **orange**

Effect of dataset size: According to Figure 9, both sizes of dataset give fairly close generalization.

Plots of the four runs in Figure. 10 and Figure. 11:

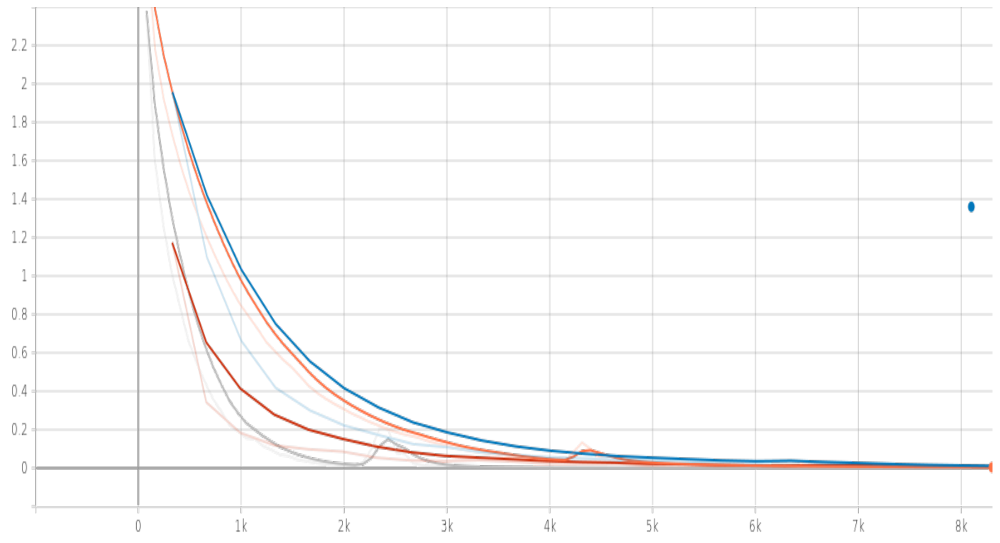


Figure 10: Comparison of training losses for the four runs(small dataset with hidden size of 32 : **orange**, small dataset with hidden size of 64 : **gray**, large dataset with hidden size of 32 : **blue**, small dataset with hidden size of 32 : **red**)

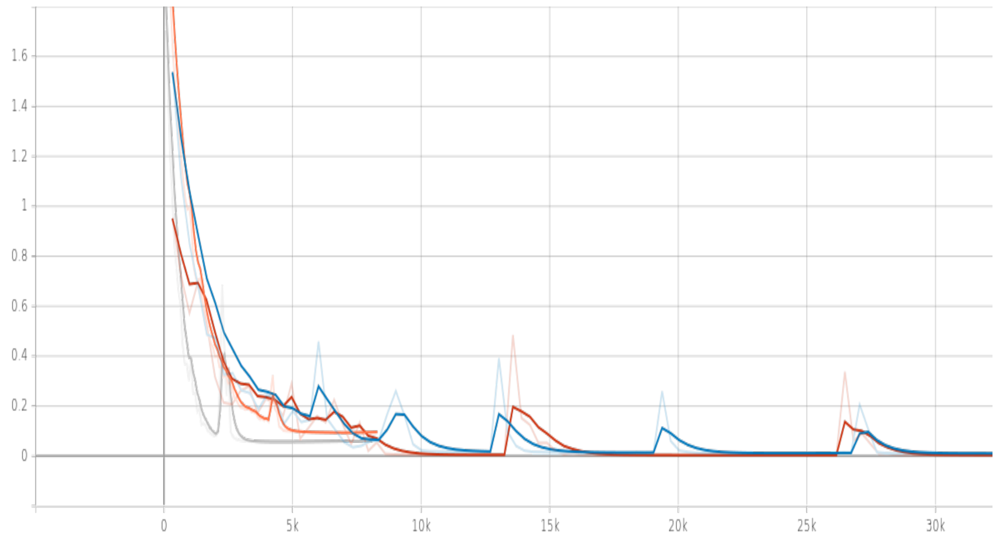


Figure 11: Comparison of validation losses for the four runs(small dataset with hidden size of 32 : **orange**, small dataset with hidden size of 64 : **gray**, large dataset with hidden size of 32 : **blue**, small dataset with hidden size of 32 : **red**)

3.7 Non-causal Attention

Though the non-causal attention yields very good quantitative results (i.e., near 0 losses), it generalizes very poorly. It is producing kind of garbage translations (like, the $\implies$ eeeott).

Here, it uses information from future time steps as the future time steps are not masked. So, it learns quite well while training as it uses future information. But, it can not translate in the test phase as it does not learn the mapping from input to output.

3.8 Advantages and disadvantages

Additive attention:

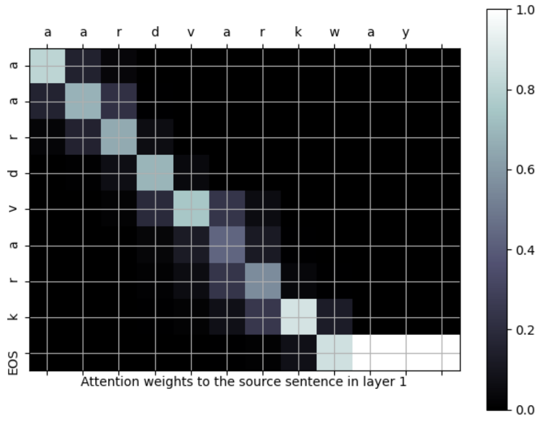
- Advantage: It induces non-linearity by adding keys and queries. This non-linearity would capture non-linear relationship.
- Disadvantage: Computationally complex: one multiplication of the input vector by a matrix, then by another matrix, and then the softmax computation.

Scaled-dot attention:

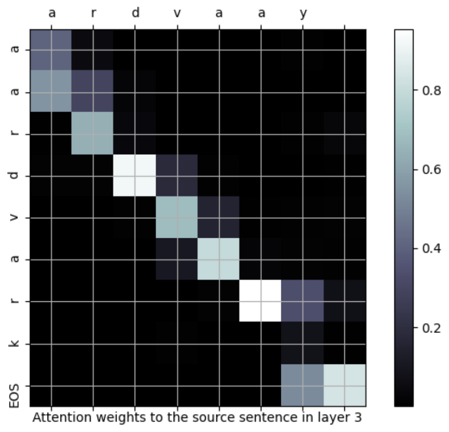
- Advantage: Computationally less complex as it has optimized matrix calculation while projecting keys, values, queries.
- Disadvantage: its dot product is a kind of linear transformation which would not capture non linear relationship. It may not fit that well as additive attention.

4 Visualizing Attention

Words beginning with a vowel: With additive attention it is a success case, while causal transformer fails to fix vowel at the start. Causal transformer mostly attends to the second letter while producing the first letter. Probably it can not capture the pattern of fixing vowel at the start with its nearly linear projection, while additive attention may succeeds to differentiate the pattern with its profound non-linearity. (See Figure. 13)



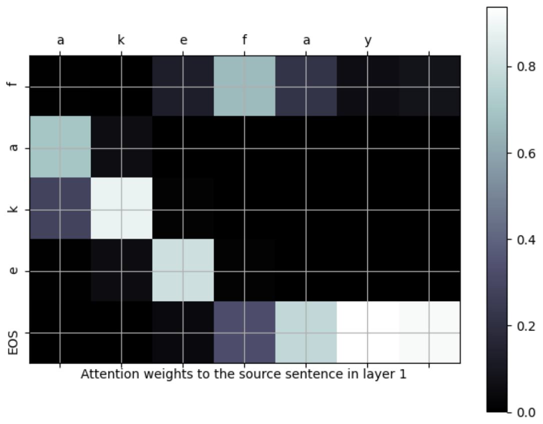
(a) With Additive Attention (Success)



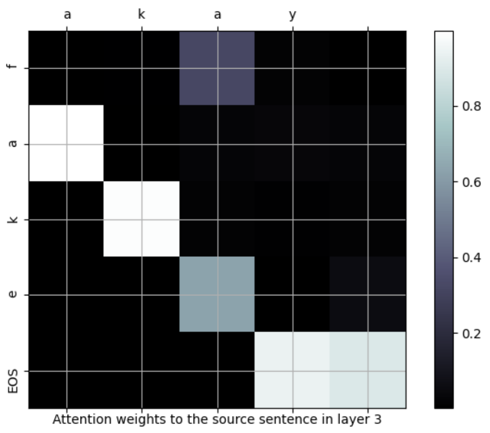
(b) With Causal Transformer (Failure partially)

Figure 12: Attention Maps for ”aardvark”

Words beginning with a single consonant: With additive attention it is a success case, while causal transformer fails. Causal transformer mostly attends to the diagonal elements from the second letter. It can not move first letter to last as it is attending little to the first letter while additive attention puts lot of attention to the first letter in moving the first letter to the end. The possible reason for the failure might be the word is small enough for causal transformer to attend different positions. Also, there might be high non-linearity in this pattern, which is hard to capture by causal transformer. (See Figure. 13)



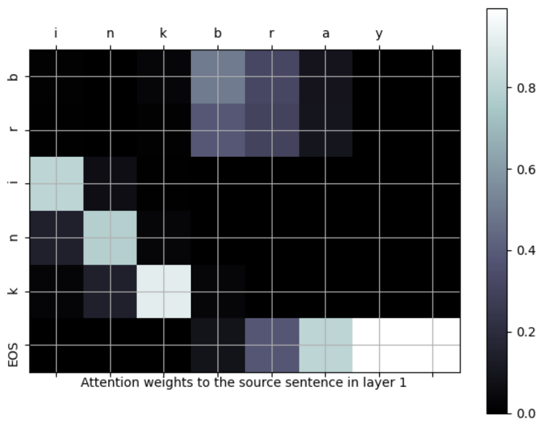
(a) With Additive Attention (Success)



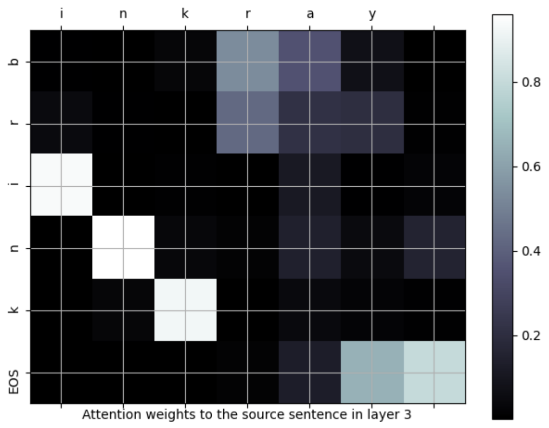
(b) With Causal Transformer (Failure)

Figure 13: Attention Maps for ”fake”

Words beginning with two or more consonants: With additive attention it is a success case, while causal transformer almost succeeds to translate it except moving the first letter to the last. Causal transformer works well in this case than before because the word is not too small. But, with words beginning with more than two consonants, both model struggles as the rule is more complex in that case (i.e., moving two or more words to the end). (See Figure. 14 and 15)

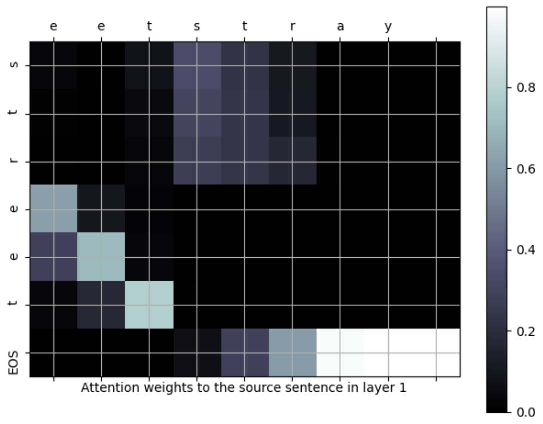


(a) With Additive Attention (Success)

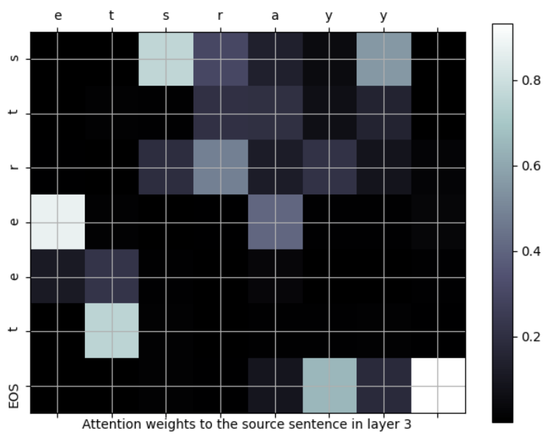


(b) With Causal Transformer (Almost Success)

Figure 14: Attention Maps for ”brink”



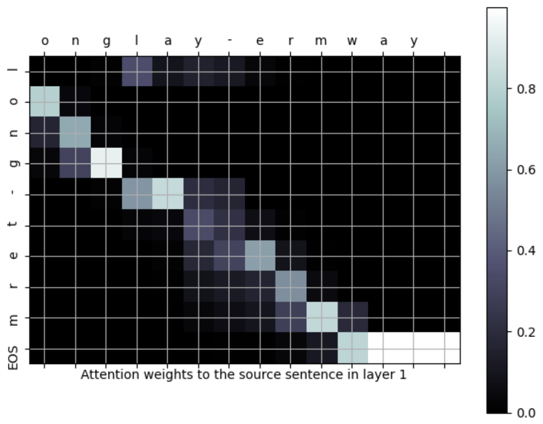
(a) With Additive Attention (Almost Success)



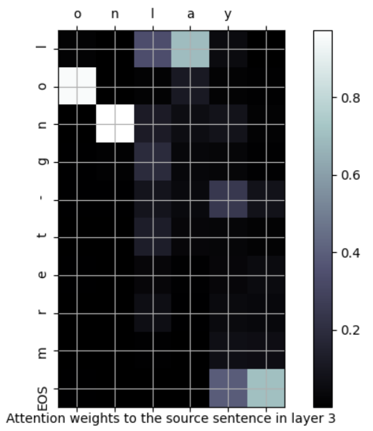
(b) With Causal Transformer (Almost Success)

Figure 15: Attention Maps for ”street”

Compound words consisting of two words separated by a dash: With additive attention it is a complete success case, it can capture the rule that each part of the word must be translated separately, and stuck back together with a dash. It attends to appropriate positions more frequently, specially near the ”dash”, it attends around the ”dash”: left and right of the ”dash”. These precise attentions help to succeeds for additive attention. On the other hand, causal transformer messes up everything by attending letters arbitrary. (See Figure. 16)



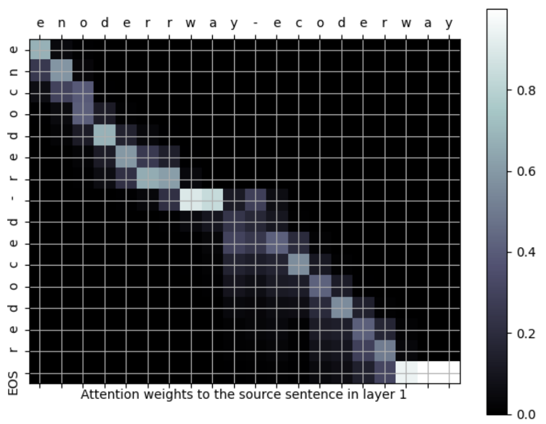
(a) With Additive Attention (Success)



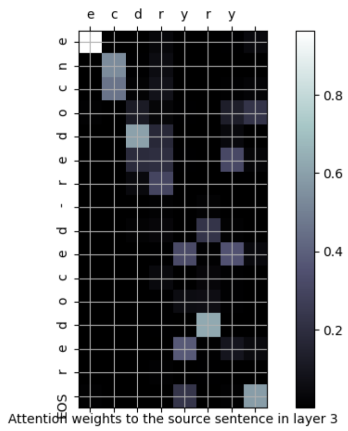
(b) With Causal Transformer (Failure)

Figure 16: Attention Maps for ”long-term”

Long-words: With additive attention it is a near success case, it can capture the whole word pattern except missing some letters (eg. ’c’ of encoder and first ’d’ of decoder). It succeeds as it is attending to the proper positions most of the time. In the failure letters, it fails to attend the corresponding letter (’c’ of encoder and first ’d’ of decoder). On the other hand, causal transformer messes up everything by attending letters arbitrary. (See Figure. 17)



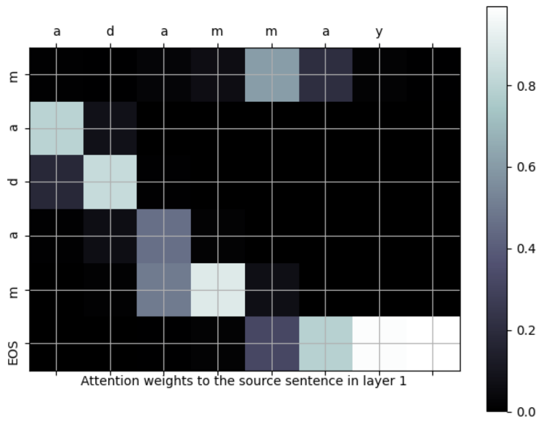
(a) With Additive Attention (Almost Success)



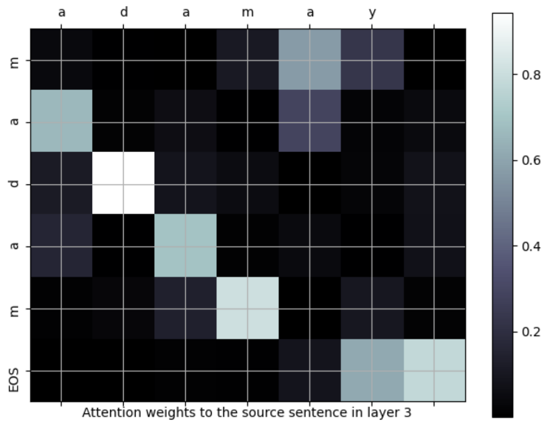
(b) With Causal Transformer (Failure)

Figure 17: Attention Maps for "encoder-decoder"

Palindrome words: Here, both additive attention and causal transformer work well. Because, both of them attend to the appropriate positions. However, causal transform struggles to move the first letter to the end. (See Figure. 18)



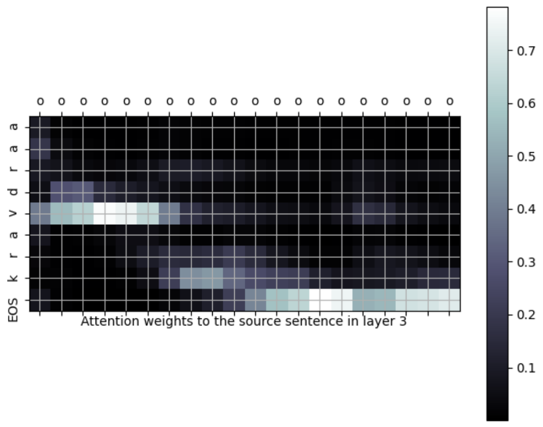
(a) With Additive Attention (Success)



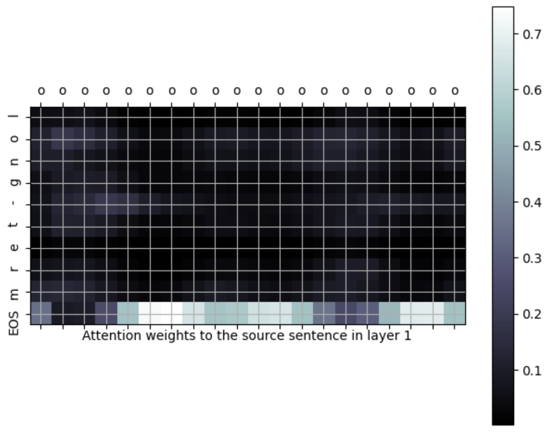
(b) With Causal Transformer (Success)

Figure 18: Attention Maps for "madam"

With Non-causal Transformer: It fails in all cases. It attends mostly to "EOS" and can not capture any relationship at all. As it uses future information while training, it actually does not learn anything and fails completely in testing. (See Figure. 19)



(a) With Additive Attention (Success)



(b) With Causal Transformer (Success)

Figure 19: Attention Maps for ”aardvark” and ”long-term”