Supplementary material-SAMDIT: Systematic study of adding memory to divided input in transformer to process long documents

Arij Al Adel[1]

Moscow Institute of Physics and Technology, Dolgoprudny, Russia
arij.aladel@gmail.com

1 Positional encoding in transformers

There are two basic types of positional encoding:

1.1 Absolute

It was first introduced in [8], where representations of absolute positions were calculated by deterministic function and then added to the transformer inputs. The deterministic function was introduced as follows:

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{model}}) \tag{1}$$

$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{model}}) \tag{2}$$

The matrix of positional encoding of shape $(vocab\_size * d_{model})$, where the pos refers to the index of the token in the input, represents the row number of the matrix where $0 <= p < vocab\_size$, (i) is the variable to indicate the column index in the matrix $0 <= i < d/2$, d_{model} model capacity, and it is usually 512. This is a fixed map from the token position in a sequence. There is another kind of absolute position encoding which is learnable positional embedding, and it is called positional embedding [1] [4]. This kind of embedding has the same dimensions as the word embedding matrix in the model. So clearly, its limitation is the fixed maximum length of input even though it is built in a data-driven way and flexible more than manually designed sinusoidal encoding [8] and computational overhead to learn the embedding[3]. Both positional encoding and positional embedding are added to word embedding before attention block; they are available just in the beginning.

By adding the positional encoding, which we can call it $\mathbf{P}$, to the input embedding $\mathbf{E} \in R^{L*d_{model}}$ where $\mathbf{L}$ is the max input length before it is fed into attention, the attention matrix will be

$$\begin{aligned} A &= \frac{(\mathbf{E} + \mathbf{P})W^{(q)}W^{(k)T}(\mathbf{E} + \mathbf{P})^T}{\sqrt{d_{model}}} \\ &= \frac{\mathbf{E}W^{(q)}W^{(k)T}\mathbf{E}^T + \mathbf{P}W^{(q)}W^{(k)T}\mathbf{E}^T + \mathbf{E}W^{(q)}W^{(k)T}\mathbf{P}^T + \mathbf{P}W^{(q)}W^{(k)T}\mathbf{P}^T}{\sqrt{d_{model}}} \\ &\approx \mathbf{E}W^{(q)}W^{(k)T}\mathbf{E}^T + \mathbf{P}W^{(q)}W^{(k)T}\mathbf{E}^T + \mathbf{E}W^{(q)}W^{(k)T}\mathbf{P}^T + \mathbf{P}W^{(q)}W^{(k)T}\mathbf{P}^T \end{aligned} \tag{3}$$

where we can interpret each term as following:

- $\mathbf{E}W^{(q)}W^{(k)T}\mathbf{E}^T$: how much attention the tokens in the query need to be given for the tokens in the key.
- $\mathbf{P}W^{(q)}W^{(k)T}\mathbf{E}^T$: how much attention the position of the query tokens needs to be given for the key tokens.
- $\mathbf{E}W^{(q)}W^{(k)T}\mathbf{P}^T$: how much attention the query tokens need to be given the positions of keys tokens.
- $\mathbf{P}W^{(q)}W^{(k)T}\mathbf{P}^T$: how much attention the position of the query tokens needs to be given the positions of key tokens.

the division on d_{model} here is symbolic, in case of multi-head attention it can be d^k this is true for later annotation too. The output of the attention block before final linear layer will be ;

$$M = SoftMax(A)(E + P)W^{(v)} \tag{4}$$

Relative Relative positional encoding (tokens pairwise distances) was first introduced in [7] as an alternative to absolute position encoding proposed in [8]. This is a kind of learnable embedding, and authors drop the interaction between tokens that are more than k distance from each other, which reduces the computational cost [3]. They add the relation position information $\mathbf{R}$ to keys and values on the fly but not to the query and the final equation of the attention block will be like this:

$$\begin{aligned} A &= \frac{\mathbf{E}W^{(q)}(\mathbf{E}W^{(k)} + \mathbf{R}^{(k)})^T}{\sqrt{d_{model}}} \\ &= \frac{\mathbf{E}W^{(q)}W^{(k)T}\mathbf{E}^T + \mathbf{E}W^{(q)}\mathbf{R}^{(k)T}}{\sqrt{d_{model}}} \end{aligned} \tag{5}$$

$$M = SoftMax(A)(E + R^{(v)})W^{(v)} \tag{6}$$

where for each element in positional embedding the maximum distance is clipped as follows:

$$r_{ij}^{(k)} = w_{clip(j-i,k)}^{(k)} \tag{7}$$

$$r_{ij}^{(v)} = w_{clip(j-i,k)}^{(v)} \tag{8}$$

$$clip(j-i,k) = MAX(-k, MIN(k,x)) \tag{9}$$

and that means that all elements after clipping the edge will have the same valu,k or $-k$. We should note that authors of [7] compared adding the relative position embedding to keys but not values, values but not keys, to both of them, and finally without positional encoding at all. According to the comparison table in their paper, adding relative position embedding to the key is the same as adding it to both keys and values. Adding it just to values decreases the BLEU value and the results drastically drop without it. These results was just for the translation task.

Later work [2] took this comparison note and used it for music data sets, adding relative positional embedding just to the key so the final representation of the relative attention will be:

$$M = SoftMax(A)(E)W^{(v)} \tag{10}$$

This final option was used for positional encoding in [6]. The authors of [9] analysed just the learnable positional encoding that was used in the transformers to see if they could catch the position meaning. The authors did that using two types of transformers, encoder-based transformer(BERT)[1] and (RoBERTa) [4], and Decoder-based transformer(GPT-2)[5] using machine translation, language modeling and text classification tasks. As a result, this work reveals that GPT-2 can learn the absolute positions, while BERT cannot. The same trend was kept for the relative position. Nevertheless, they discovered that GPT-2 could catch the learned positional embedding in GPT-2 and could represent the relative positions better than the sinusoidal function, which is supposed to represent the relative positions as declared in landmark paper [8].

References

1. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: NAACL (2019)
2. Huang, C.Z.A., Vaswani, A., Uszkoreit, J., Simon, I., Hawthorne, C., Shazeer, N.M., Dai, A.M., Hoffman, M.D., Dinculescu, M., Eck, D.: Music transformer: Generating music with long-term structure. In: ICLR (2019)
3. Liu, X., Yu, H.F., Dhillon, I.S., Hsieh, C.J.: Learning to encode position for transformer with continuous dynamical model. ArXiv abs/2003.09229 (2020)
4. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. ArXiv abs/1907.11692 (2019)
5. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog 1(8), 9 (2019)
6. Raffel, C., Shazeer, N.M., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. ArXiv abs/1910.10683 (2020)

7. Shaw, P., Uszkoreit, J., Vaswani, A.: Self-attention with relative position representations. In: NAACL (2018)
8. Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NIPS (2017)
9. Wang, Y.A., Chen, Y.N.: What do position embeddings learn? an empirical study of pre-trained language model positional encoding. In: EMNLP (2020)

A Memory content and attention maps in encoder

Giving an example from SAMSum data set sample 115, we will display the text before tokenization, full tokenized text, then how it is divided into chunks each chunk of 384 tokens length before encoding and after encoding to see the effect of using memory slots, then we will display the content of memory slots so we will have a closer look into the encoder outputs, memory capacity is equal to block length divided by 32 i.e. $384/32 = 12$ memory tokens in each slot:

- raw text input: Kyle: Who wants to go out for a drink? Megan: No, sorry, I'm cleaning the house today. Roseanne: You've always loved cleaning, haven't you? I remember how angry you used to get with your brother for leaving a mess in the kitchen. Vince: Yeah, she'd always yell at me, even though I was the one in charge when our parents were away. Kyle: I don't get why it matters so much whether I clean my flat once a week or once a month. No one died from a bit of dust. Megan: Remind me to never stay at your place :P Roseanne: I'm somewhere in between. My house is always a mess, but I hate it when it's dirty. Vince: What's the difference? Roseanne: I can't stand when there's dust, crumbs on the floor, etc., so I clean regularly. But I love it when all my stuff just lies around. When I put everything back on shelves or into cabinets, I keep on getting mad that I have to look for it, get up and take it out, and generally lose so much time. I like to have everything within my reach. Vince: There's no better wardrobe than the armchair, right? XD Kyle: I prefer the floor. There's more space on it :D Megan: Just reading this gives me the creeps. Kyle: Fine, go back to cleaning, we'll think of you while sharing a beer :D Roseanne: Hey, I never said I can come. My hubby's away this weekend, so I have no one to leave the kids with ;(Vince: Take them to Megan's place, they can help her clean :D Megan: You know what? I don't care. I can watch your kids, Roseanne. Just bring them over. Roseanne: But are you serious or just mad at us? Megan: No, I'm serious. I like your kids. And they're nice enough that if I ask them to do something, they actually do it :D Roseanne: LOL, okay. Actually, I think that when I tell them to go and help aunt Meg with cleaning, they'll think it's fun. Vince: And when you tell them to clean their own rooms, they say you're a monster? :D Roseanne: Exactly. But Bill for example always cleans the windows at my parents house. They don't even need to ask anymore, he simply knows it's his job. Kyle: Do they pay him? :P Roseanne: Well, sometimes. But still I think he likes doing it. I don't know, maybe it's because no one checks if he did it well, he's just left alone to do his thing, so he feels like a man in the house in a way? Megan: Tell them I can pay with cookies if they help me. Roseanne: The famous chocolate cookies? Megan: Why not? I’m feeling generous today :D Roseanne: Guys, I think I’m gonna change my mind about that beer. Meg, would you accept one more helper? :D Megan: You’re always welcome! But sorry boys, it’s going to be a “girls only” event (except for little Billy). Otherwise my house will never get cleaned today :D

- tokenized input: number of tokens is 761 so this input will be chunked into two chunks each chunk of length 384 tokens last chunk will be padded as we will see later

  [’Kyle’, ’:’, ’Who’, ’wants’, ’to’, ’go’, ’out’, ’for’, ”, ’a’, ’drink’, ’?’, ’Mega’, ’n’, ’:’, ’No’, ’,’, ’sorry’, ’,’, ’I’, ”’”, ’m’, ’cleaning’, ’the’, ’house’, ’today’, ’.’, ’Rose’, ’anne’, ’:’, ’You’, ”’”, ’ve’, ’always’, ’loved’, ’cleaning’, ’,’, ’have’, ’n’, ”’”, ’t’, ’you’, ’?’, ’I’, ’remember’, ’how’, ’angry’, ’you’, ’used’, ’to’, ’get’, ’with’, ’your’, ’brother’, ’for’, ’leaving’, ”, ’a’, ’mess’, ’in’, ’the’, ’kitchen’, ’.’, ’Vince’, ’:’, ’Yeah’, ’,’, ’she’, ”’”, ’d’, ’always’, ”, ’y’, ’ell’, ’at’, ’me’, ’,’, ’even’, ’though’, ’I’, ’was’, ’the’, ’one’, ’in’, ’charge’, ’when’, ’our’, ’parents’, ’were’, ’away’, ’.’, ’Kyle’, ’:’, ’I’, ’don’, ”’”, ’t’, ’get’, ’why’, ’it’, ’matters’, ’so’, ’much’, ’whether’, ’I’, ’clean’, ’my’, ’flat’, ’once’, ”, ’a’, ’week’, ’or’, ’once’, ”, ’a’, ’month’, ’.’, ’No’, ’one’, ’died’, ’from’, ”, ’a’, ’bit’, ’of’, ’dust’, ’.’, ’Mega’, ’n’, ’:’, ’Re’, ’mind’, ’me’, ’to’, ’never’, ’stay’, ’at’, ’your’, ’place’, ”, ’:’, ’P’, ’Rose’, ’anne’, ’:’, ’I’, ”’”, ’m’, ’somewhere’, ’in’, ’between’, ’.’, ’My’, ’house’, ’is’, ’always’, ”, ’a’, ’mess’, ’,’, ’but’, ’I’, ’hate’, ’it’, ’when’, ’it’, ”’”, ’s’, ’dirty’, ’.’, ’Vince’, ’:’, ’What’, ”’”, ’s’, ’the’, ’difference’, ’?’, ’Rose’, ’anne’, ’:’, ’I’, ’can’, ”’”, ’t’, ’stand’, ’when’, ’there’, ”’”, ’s’, ’dust’, ’,’, ”, ’crumbs’, ’on’, ’the’, ’floor’, ’etc’, ’.’, ’,’, ’so’, ’I’, ’clean’, ’regularly’, ’.’, ’But’, ’I’, ’love’, ’it’, ’when’, ’all’, ’my’, ’stuff’, ’just’, ’lies’, ’around’, ’.’, ’When’, ’I’, ’put’, ’everything’, ’back’, ’on’, ’shelves’, ’or’, ’into’, ’cabinets’, ’,’, ’I’, ’keep’, ’on’, ’getting’, ’mad’, ’that’, ’I’, ’have’, ’to’, ’look’, ’for’, ’it’, ’,’, ’get’, ’up’, ’and’, ’take’, ’it’, ’out’, ’,’, ’and’, ’generally’, ’lose’, ’so’, ’much’, ’time’, ’.’, ’I’, ’like’, ’to’, ’have’, ’everything’, ’within’, ’my’, ’reach’, ’.’, ’Vince’, ’:’, ’There’, ”’”, ’s’, ’no’, ’better’, ’wardrobe’, ’than’, ’the’, ’arm’, ’chair’, ’,’, ’right’, ’?’, ”, ’X’, ’D’, ’Kyle’, ’:’, ’I’, ’prefer’, ’the’, ’floor’, ’.’, ’There’, ”’”, ’s’, ’more’, ’space’, ’on’, ’it’, ”, ’:’, ’D’, ’Mega’, ’n’, ’:’, ’Just’, ’reading’, ’this’, ’gives’, ’me’, ’the’, ’creep’, ’s’, ’.’, ’Kyle’, ’:’, ’Fine’, ’,’, ’go’, ’back’, ’to’, ’cleaning’, ’,’, ’we’, ”’”, ’ll’, ’think’, ’of’, ’you’, ’while’, ’sharing’, ”, ’a’, ’beer’, ”, ’:’, ’D’, ’Rose’, ’anne’, ’:’, ’Hey’, ’,’, ’I’, ’never’, ’said’, ’I’, ’can’, ’come’, ’.’, ’My’, ’hub’, ’by’, ”’”, ’s’, ’away’, ’this’, ’weekend’, ’,’, ’so’, ’I’, ’have’, ’no’, ’one’, ’to’, ’leave’, ’the’, ’kids’, ’with’, ”, ’;’, ’(’, ’Vince’, ’:’, ’Take’, ’them’, ’to’, ’Mega’, ’n’, ”’”, ’s’, ’place’, ’,’, ’they’, ’can’, ’help’, ’her’, ’clean’, ”, ’:’, ’D’, ’Mega’, ’n’, ’:’, ’You’, ’know’, ’what’, ’?’, ’I’, ’don’, ”’”, ’t’, ’care’, ’.’, ’I’, ’can’, ’watch’, ’your’, ’kids’, ’,’, ’Rose’, ’anne’, ’.’, ’Just’, ’bring’, ’them’, ’over’, ’.’, ’Rose’, ’anne’, ’:’, ’But’, ’are’, ’you’, ’serious’, ’or’, ’just’, ’mad’, ’at’, ’us’, ’?’, ’Mega’, ’n’, ’:’, ’No’, ’,’, ’I’, ”’”, ’m’, ’serious’, ’.’, ’I’, ’like’, ’your’, ’kids’, ’.’, ’And’, ’they’, ”’”, ’re’, ’nice’, ’enough’, ’that’, ”, ’if’, ’I’, ’ask’, ’them’, ’to’, ’do’, ’something’, ’,’, ’they’, ’actually’, ’do’, ’it’, ”, ’:’, ’D’, ’Rose’, ’anne’, ’:’, ’LOL’, ’,’, ’okay’, ’.’, ’Actually’, ’,’, ’I’, ’think’, ’that’, ’when’, ’I’, ’tell’,

’them’, ’to’, ’go’, ’and’, ’help’, ’aunt’, ’Me’, ’g’, ’with’, ’cleaning’, ’,’, ’they’, ”’”, ’ll’, ’think’, ’it’, ”’”, ’s’, ’fun’, ’.’, ’Vince’, ’:’, ’And’, ’when’, ’you’, ’tell’, ’them’, ’to’, ’clean’, ’their’, ’own’, ’rooms’, ’,’, ’they’, ’say’, ’you’, ”’”, ’re’, ”, ’a’, ’monster’, ’?’, ”, ’:’, ’D’, ’Rose’, ’anne’, ’:’, ”, ’Exactly’, ’.’, ’But’, ’Bill’, ’for’, ’example’, ’always’, ’clean’, ’s’, ’the’, ’windows’, ’at’, ’my’, ’parents’, ’house’, ’.’, ’They’, ’don’, ”’”, ’t’, ’even’, ’need’, ’to’, ’ask’, ’anymore’, ’,’, ”, ’he’, ’simply’, ’knows’, ’it’, ”’”, ’s’, ’his’, ’job’, ’.’, ’Kyle’, ’:’, ’Do’, ’they’, ’pay’, ’him’, ’?’, ”, ’:’, ’P’, ’Rose’, ’anne’, ’:’, ’Well’, ’,’, ’sometimes’, ’.’, ’But’, ’still’, ’I’, ’think’, ”, ’he’, ’like’, ’s’, ’doing’, ’it’, ’.’, ’I’, ’don’, ”’”, ’t’, ’know’, ’,’, ’maybe’, ’it’, ”’”, ’s’, ’because’, ’no’, ’one’, ’checks’, ”, ’if’, ”, ’he’, ’did’, ’it’, ’well’, ’,’, ”, ’he’, ”’”, ’s’, ’just’, ’left’, ’alone’, ’to’, ’do’, ’his’, ’thing’, ’,’, ’so’, ”, ’he’, ’feels’, ’like’, ”, ’a’, ’man’, ’in’, ’the’, ’house’, ’in’, ”, ’a’, ’way’, ’?’, ’Mega’, ’n’, ’:’, ’Tell’, ’them’, ’I’, ’can’, ’pay’, ’with’, ’cookies’, ”, ’if’, ’they’, ’help’, ’me’, ’.’, ’Rose’, ’anne’, ’:’, ’The’, ’famous’, ’chocolate’, ’cookies’, ’?’, ’Mega’, ’n’, ’:’, ’Why’, ’not’, ’?’, ’I’, ”’”, ’m’, ’feeling’, ’generous’, ’today’, ”, ’:’, ’D’, ’Rose’, ’anne’, ’:’, ’Guys’, ’,’, ’I’, ’think’, ’I’, ”’”, ’m’, ”, ’gonna’, ’change’, ’my’, ’mind’, ’about’, ’that’, ’beer’, ’.’, ’Me’, ’g’, ’,’, ’would’, ’you’, ’accept’, ’one’, ’more’, ’help’, ’er’, ’?’, ”, ’:’, ’D’, ’Mega’, ’n’, ’:’, ’You’, ”’”, ’re’, ’always’, ’welcome’, ’!’, ’But’, ’sorry’, ’boys’, ’,’, ’it’, ”’”, ’s’, ’going’, ’to’, ’be’, ”, ’a’, ’“’, ’girl’, ’s’, ’only’, ’”’, ’event’, ’(’, ’except’, ’for’, ’little’, ’Billy’, ’).’, ’Otherwise’, ’my’, ’house’, ’will’, ’never’, ’get’, ’cleaned’, ’today’, ”, ’:’, ’D’, ’¡/s¿’]

- First block:

  Before encoding:

  [’Kyle’, ’:’, ’Who’, ’wants’, ’to’, ’go’, ’out’, ’for’, ”, ’a’, ’drink’, ’?’, ’Mega’, ’n’, ’:’, ’No’, ’,’, ’sorry’, ’,’, ’I’, ”’”, ’m’, ’cleaning’, ’the’, ’house’, ’today’, ’.’, ’Rose’, ’anne’, ’:’, ’You’, ”’”, ’ve’, ’always’, ’loved’, ’cleaning’, ’,’, ’have’, ’n’, ”’”, ’t’, ’you’, ’?’, ’I’, ’remember’, ’how’, ’angry’, ’you’, ’used’, ’to’, ’get’, ’with’, ’your’, ’brother’, ’for’, ’leaving’, ”, ’a’, ’mess’, ’in’, ’the’, ’kitchen’, ’.’, ’Vince’, ’:’, ’Yeah’, ’,’, ’she’, ”’”, ’d’, ’always’, ”, ’y’, ’ell’, ’at’, ’me’, ’,’, ’even’, ’though’, ’I’, ’was’, ’the’, ’one’, ’in’, ’charge’, ’when’, ’our’, ’parents’, ’were’, ’away’, ’.’, ’Kyle’, ’:’, ’I’, ’don’, ”’”, ’t’, ’get’, ’why’, ’it’, ’matters’, ’so’, ’much’, ’whether’, ’I’, ’clean’, ’my’, ’flat’, ’once’, ”, ’a’, ’week’, ’or’, ’once’, ”, ’a’, ’month’, ’.’, ’No’, ’one’, ’died’, ’from’, ”, ’a’, ’bit’, ’of’, ’dust’, ’.’, ’Mega’, ’n’, ’:’, ’Re’, ’mind’, ’me’, ’to’, ’never’, ’stay’, ’at’, ’your’, ’place’, ”, ’:’, ’P’, ’Rose’, ’anne’, ’:’, ’I’, ”’”, ’m’, ’somewhere’, ’in’, ’between’, ’.’, ’My’, ’house’, ’is’, ’always’, ”, ’a’, ’mess’, ’,’, ’but’, ’I’, ’hate’, ’it’, ’when’, ’it’, ”’”, ’s’, ’dirty’, ’.’, ’Vince’, ’:’, ’What’, ”’”, ’s’, ’the’, ’difference’, ’?’, ’Rose’, ’anne’, ’:’, ’I’, ’can’, ”’”, ’t’, ’stand’, ’when’, ’there’, ”’”, ’s’, ’dust’, ’,’, ”, ’crumbs’, ’on’, ’the’, ’floor’, ’etc’, ’.’, ’,’, ’so’, ’I’, ’clean’, ’regularly’, ’.’, ’But’, ’I’, ’love’, ’it’, ’when’, ’all’, ’my’, ’stuff’, ’just’, ’lies’, ’around’, ’.’, ’When’, ’I’, ’put’, ’everything’, ’back’, ’on’, ’shelves’, ’or’, ’into’, ’cabinets’, ’,’, ’I’, ’keep’, ’on’, ’getting’, ’mad’, ’that’, ’I’, ’have’, ’to’, ’look’, ’for’, ’it’, ’,’, ’get’, ’up’, ’and’, ’take’, ’it’, ’out’, ’,’, ’and’, ’generally’, ’lose’, ’so’, ’much’, ’time’, ’.’, ’I’, ’like’, ’to’, ’have’, ’everything’, ’within’, ’my’, ’reach’, ’.’, ’Vince’, ’:’, ’There’, ”’”, ’s’, ’no’, ’better’, ’wardrobe’, ’than’, ’the’, ’arm’, ’chair’, ’,’, ’right’, ’?’, ”, ’X’,

’D’, ’Kyle’, ’:’, ’I’, ’prefer’, ’the’, ’floor’, ’.’, ’There’, ”’”, ’s’, ’more’, ’space’, ’on’, ’it’, ”, ’:’, ’D’, ’Mega’, ’n’, ’:’, ’Just’, ’reading’, ’this’, ’gives’, ’me’, ’the’, ’creep’, ’s’, ’.’, ’Kyle’, ’:’, ’Fine’, ’,’, ’go’, ’back’, ’to’, ’cleaning’, ’,’, ’we’, ”’”, ’ll’, ’think’, ’of’, ’you’, ’while’, ’sharing’, ”, ’a’, ’beer’, ”, ’:’, ’D’, ’Rose’, ’anne’, ’:’, ’Hey’, ’,’, ’I’, ’never’, ’said’, ’I’, ’can’, ’come’, ’.’, ’My’, ’hub’, ’by’, ”’”, ’s’, ’away’, ’this’, ’weekend’, ’,’, ’so’, ’I’, ’have’, ’no’, ’one’, ’to’, ’leave’, ’the’, ’kids’, ’with’, ”, ’;’, ’(’, ’Vince’, ’:’, ’Take’, ’them’, ’to’, ’Mega’, ’n’, ”’”, ’s’, ’place’, ’,’, ’they’, ’can’, ’help’, ’her’]

After encoding:

[’Kyle’, ’comenzi’, ’Who’, ’wants’, ’sarbatori’, ’go’, ’Out’, ’FOR’, ’sticla’, ’drinks’, ’drinks’, ’cumpara’, ’Mega’, ’Alicia’, ’cît’, ’No’, ’coloan’, ’sorry’, ’datorita’, ’I’, ’datorita’, ’Oricum’, ’cleaning’, ’illage’, ’house’, ’today’, ’ceapa’, ’Rose’, ’anne’, ’cerinte’, ’you’, ’datorita’, ’ve’, ’always’, ’loves’, ’cleaning’, ’nebun’, ’Has’, ’bian’, ’Hannover’, ’constient’, ’cladiri’, ’?’, ’tang’, ’remember’, ’how’, ’angry’, ’you’, ’ancienne’, ’nature’, ’Get’, ’tră’, ’your’, ’brother’, ’for’, ’Leaving’, ’incep’, ’gradini’, ’mess’, ’gradini’, ’corridor’, ’kitchen’, ’band’, ’Vince’, ’:’, ’Yeah’, ’train’, ’she’, ’căt’, ’would’, ’always’, ’cerinte’, ’shout’, ’ell’, ’offensive’, ’me’, ’datorita’, ’even’, ’obwohl’, ’vest’, ’rolul’, ’varf’, ’plate’, ’guard’, ’responsables’, ’when’, ’dra’, ’parents’, ’were’, ’away’, ’gasi’, ’Kyle’, ’:’, ’vietii’, ’don’, ’functie’, ’None’, ’understand’, ’why’, ’culoarea’, ’Matter’, ’so’, ’piata’, ’whether’, ’pacate’, ’cleaning’, ’My’, ’flat’, ’once’, ’incep’, ’weekly’, ’weekly’, ’reteta’, ’once’, ’incep’, ’yearly’, ’month’, ’oricine’, ’No’, ’nobody’, ’died’, ’proaspat’, ’incep’, ’sarbatori’, ’bit’, ’sarbatori’, ’dust’, ’.’, ’Mega’, ’căt’, ’cerinte’, ’Re’, ’mind’, ’me’, ’prohibition’, ’Never’, ’stay’, ’villa’, ’aluat’, ’place’, ’incep’, ’:’, ’P’, ’Rose’, ’anne’, ’:’, ’simti’, ”’”, ’metru’, ’somewhere’, ’meciul’, ’between’, ’aparțin’, ’My’, ’house’, ’psiho’, ’always’, ’noc’, ’apariti’, ’mess’, ’pielea’, ’but’, ’prefer’, ’hate’, ’pielea’, ’when’, ’it’, ”’”, ’sanatos’, ’dirty’, ’rier’, ’Vince’, ’questions’, ’What’, ’functie’, ’mânt’, ’differences’, ’difference’, ’simti’, ’Rose’, ’anne’, ’:’, ’sustine’, ’Can’, ’functie’, ’cannot’, ’withstand’, ’when’, ’there’, ’functie’, ’judeţul’, ’dust’, ’flüssig’, ’noc’, ’crumbs’, ’tău’, ’tră’, ’floor’, ’etc’, ’...).’, ’ssen’, ’so’, ’terio’, ’cleaning’, ’regularly’, ’win’, ’But’, ’préfér’, ’love’, ’it’, ’when’, ’All’, ’My’, ’stuff’, ’just’, ’lying’, ’around’, ’mig’, ’When’, ’dging’, ’put’, ’everything’, ’Back’, ’apariti’, ’shelves’, ’cumva’, ’Into’, ’cabinets’, ’congestion’, ’heft’, ’Keep’, ’auf’, ’getting’, ’mad’, ’heft’, ’tri’, ’obligation’, ’manually’, ’look’, ’search’, ’heft’, ’jung’, ’get’, ’Up’, ’ündig’, ’take’, ’it’, ’out’, ’logistic’, ’holz’, ’generally’, ’lose’, ’tellement’, ’Much’, ’time’, ’”)’, ’prefer’, ’prefer’, ’preferably’, ’Have’, ’everything’, ’adica’, ’my’, ’reach’, ’aille’, ’Vince’, ’cerinte’, ’lexic’, ’simti’, ’sufletul’, ’no’, ’meilleures’, ’wardrobe’, ’than’, ’nisip’, ’arm’, ’chair’, ’net’, ’huh’, ’?’, ’incep’, ’X’, ’D’, ’Kyle’, ’aparatul’, ’preference’, ’prefer’, ’Rücken’, ’floor’, ’cumva’, ’expunere’, ’multumesc’, ’cresterea’, ’fewer’, ’space’, ’on’, ’cartofi’, ’gama’, ’:’, ’D’, ’Mega’, ’Hannah’, ’:’, ’Just’, ’reading’, ’Archived’, ’gives’, ’me’, ’tră’, ’creep’, ’s’, ’randul’, ’Kyle’, ’cît’, ’Fine’, ’cladiri’, ’Go’, ’înapoi’, ’ivitate’, ’cleaning’, ’graphic’, ’jointly’, ’honneur’, ’ll’, ’Think’, ’sarbatori’, ’reteta’, ’while’, ’sharing’, ’sticla’, ’drinks’, ’Beer’, ’ner’, ’:’, ’D’, ’Rose’, ’anne’, ’:’, ’Hey’, ’censor’, ’sustinut’, ’never’, ’gardinen’, ’căt’, ’Can’, ’come’, ’constraints’, ’My’, ’hub’, ’by’, ’Dimensiuni’, ’angle’, ’away’, ’ubi’, ’weekend’, ’incat’, ’cadou’, ’iocese’, ’lack’, ’no’, ’nobody’, ’județ’, ’leave’, ’copilul’, ’kids’, ’with’, ’incep’, ’;’, ’dispoziti’, ’Vince’, ’prezinta’, ’Take’, ’youngsters’, ’dispoziti’, ’Mega’, ’Hannah’, ”’”, ’aluat’, ’place’, ’gold’, ’they’, ’Can’, ’help’, ’herself’]

Memory of block1:

[’Spitalul’, ’Spitalul’, ’Spitalul’, ’Spitalul’, ’Spitalul’, ’Spitalul’, ’Spitalul’, ’Spitalul’, ’Spitalul’, ’Spitalul’, ’Spitalul’, ’Spitalul’]

- Second block:

Before encoding since the remain text length is 377 tokens then it needs to be padded with 7 pad tokens to get chunk with length 384 tokens:

[’clean’, ”, ’:’, ’D’, ’Mega’, ’n’, ’:’, ’You’, ’know’, ’what’, ’?’, ’I’, ’don’, ”’”, ’t’, ’care’, ’.’, ’I’, ’can’, ’watch’, ’your’, ’kids’, ’,’, ’Rose’, ’anne’, ’.’, ’Just’, ’bring’, ’them’, ’over’, ’.’, ’Rose’, ’anne’, ’:’, ’But’, ’are’, ’you’, ’serious’, ’or’, ’just’, ’mad’, ’at’, ’us’, ’?’, ’Mega’, ’n’, ’:’, ’No’, ’,’, ’I’, ”’”, ’m’, ’serious’, ’.’, ’I’, ’like’, ’your’, ’kids’, ’.’, ’And’, ’they’, ”’”, ’re’, ’nice’, ’enough’, ’that’, ”, ’if’, ’I’, ’ask’, ’them’, ’to’, ’do’, ’something’, ’,’, ’they’, ’actually’, ’do’, ’it’, ”, ’:’, ’D’, ’Rose’, ’anne’, ’:’, ’LOL’, ’,’, ’okay’, ’.’, ’Actually’, ’,’, ’I’, ’think’, ’that’, ’when’, ’I’, ’tell’, ’them’, ’to’, ’go’, ’and’, ’help’, ’aunt’, ’Me’, ’g’, ’with’, ’cleaning’, ’,’, ’they’, ”’”, ’ll’, ’think’, ’it’, ”’”, ’s’, ’fun’, ’.’, ’Vince’, ’:’, ’And’, ’when’, ’you’, ’tell’, ’them’, ’to’, ’clean’, ’their’, ’own’, ’rooms’, ’,’, ’they’, ’say’, ’you’, ”’”, ’re’, ”, ’a’, ’monster’, ’?’, ”, ’:’, ’D’, ’Rose’, ’anne’, ’:’, ”, ’Exactly’, ’.’, ’But’, ’Bill’, ’for’, ’example’, ’always’, ’clean’, ’s’, ’the’, ’windows’, ’at’, ’my’, ’parents’, ’house’, ’.’, ’They’, ’don’, ”’”, ’t’, ’even’, ’need’, ’to’, ’ask’, ’anymore’, ’,’, ”, ’he’, ’simply’, ’knows’, ’it’, ”’”, ’s’, ’his’, ’job’, ’.’, ’Kyle’, ’:’, ’Do’, ’they’, ’pay’, ’him’, ’?’, ”, ’:’, ’P’, ’Rose’, ’anne’, ’:’, ’Well’, ’,’, ’sometimes’, ’.’, ’But’, ’still’, ’I’, ’think’, ”, ’he’, ’like’, ’s’, ’doing’, ’it’, ’.’, ’I’, ’don’, ”’”, ’t’, ’know’, ’,’, ’maybe’, ’it’, ”’”, ’s’, ’because’, ’no’, ’one’, ’checks’, ”, ’if’, ”, ’he’, ’did’, ’it’, ’well’, ’,’, ”, ’he’, ”’”, ’s’, ’just’, ’left’, ’alone’, ’to’, ’do’, ’his’, ’thing’, ’,’, ’so’, ”, ’he’, ’feels’, ’like’, ”, ’a’, ’man’, ’in’, ’the’, ’house’, ’in’, ”, ’a’, ’way’, ’?’, ’Mega’, ’n’, ’:’, ’Tell’, ’them’, ’I’, ’can’, ’pay’, ’with’, ’cookies’, ”, ’if’, ’they’, ’help’, ’me’, ’.’, ’Rose’, ’anne’, ’:’, ’The’, ’famous’, ’chocolate’, ’cookies’, ’?’, ’Mega’, ’n’, ’:’, ’Why’, ’not’, ’?’, ’I’, ”’”, ’m’, ’feeling’, ’generous’, ’today’, ”, ’:’, ’D’, ’Rose’, ’anne’, ’:’, ’Guys’, ’,’, ’I’, ’think’, ’I’, ”’”, ’m’, ”, ’gonna’, ’change’, ’my’, ’mind’, ’about’, ’that’, ’beer’, ’.’, ’Me’, ’g’, ’,’, ’would’, ’you’, ’accept’, ’one’, ’more’, ’help’, ’er’, ’?’, ”, ’:’, ’D’, ’Mega’, ’n’, ’:’, ’You’, ”’”, ’re’, ’always’, ’welcome’, ’!’, ’But’, ’sorry’, ’boys’, ’,’, ’it’, ”’”, ’s’, ’going’, ’to’, ’be’, ”, ’a’, ’“’, ’girl’, ’s’, ’only’, ’”’, ’event’, ’(’, ’except’, ’for’, ’little’, ’Billy’, ’).’, ’Otherwise’, ’my’, ’house’, ’will’, ’never’, ’get’, ’cleaned’, ’today’, ”, ’:’, ’D’, ’¡/s¿’, ’¡pad¿’, ’¡pad¿’, ’¡pad¿’, ’¡pad¿’, ’¡pad¿’, ’¡pad¿’, ’¡pad¿’]

After encoding:

[’clean’, ’syn’, ’aparatul’, ’D’, ’Mega’, ’woman’, ’cerinte’, ’you’, ’know’, ’functioneaza’, ’Oui’, ’I’, ’don’, ’sanatate’, ’None’, ’matter’, ’clav’, ’I’, ’Can’, ’watch’, ’your’, ’kids’, ’indu’, ’Rose’, ’anne’, ’coloan’, ’Just’, ’Bring’, ’them’, ’over’, ’.’, ’Rose’, ’anne’, ’comenzi’, ’But’, ’Are’, ’YOU’, ’serious’, ’or’, ’just’, ’mad’, ’psiho’, ’us’, ’?’, ’Mega’, ’woman’, ’:’, ’No’, ’constru’, ’Britanie’, ”’”, ’mânt’,

’serious’, ’jung’, ’intelege’, ’liked’, ’your’, ’kids’, ’cur’, ’cedat’, ’participanți’, ”’”, ’Dimensiuni’, ’nice’, ’enough’, ’randul’, ’incep’, ’if’, ’Britanie’, ’asking’, ’Copiii’, ’comenzi’, ’doing’, ’something’, ’dron’, ’they’, ’actually’, ’do’, ’it’, ’incep’, ’hugs’, ’D’, ’Rose’, ’anne’, ’:’, ’LOL’, ’masura’, ’okay’, ’cur’, ’Actually’, ’datorita’, ’I’, ’Think’, ’enburg’, ’When’, ’fructe’, ’telling’, ’children’, ’judeţul’, ’go’, ’ander’, ’Help’, ’aunt’, ’Me’, ’gie’, ’assisting’, ’cleaning’, ’dron’, ’they’, ”’”, ’ll’, ’think’, ’formularul’, ”’”, ’sustine’, ’fun’, ’dry’, ’Vince’, ’slav’, ’And’, ’When’, ’YOU’, ’Tell’, ’them’, ’Pflicht’, ’cleaning’, ’their’, ’own’, ’rooms’, ’stroke’, ’they’, ’saying’, ’YOU’, ”’”, ’ARE’, ’incep’, ’wasting’, ’monster’, ’?’, ’senzati’, ’:’, ’D’, ’Rose’, ’anne’, ’Asadar’, ’cerinte’, ’Exactly’, ’rift’, ’But’, ’Bill’, ’example’, ’example’, ’always’, ’cleaning’, ’cleaning’, ’ander’, ’windows’, ’at’, ’My’, ’parents’, ’house’, ’nac’, ’they’, ’don’, ’datorita’, ’no’, ’even’, ’need’, ’machiaj’, ’ask’, ’anymore’, ’instalat’, ’incep’, ’he’, ’simply’, ’knows’, ’fil’, ”’”, ’dus’, ’his’, ’job’, ’ionar’, ’Kyle’, ’:’, ’Does’, ’they’, ’pay’, ’him’, ’?’, ’incep’, ’:’, ’P’, ’Rose’, ’anne’, ’:’, ’Well’, ’denumirea’, ’sometimes’, ’nes’, ’but’, ’still’, ’I’, ’think’, ’incep’, ’he’, ’liking’, ’psiho’, ’doing’, ’fil’, ’.”’, ’I’, ’don’, ”’”, ’unsure’, ’unsure’, ’simti’, ’maybe’, ’cumva’, ”’”, ’cumva’, ’because’, ’no’, ’nobody’, ’checks’, ’incep’, ’if’, ’incep’, ’he’, ’did’, ’task’, ’well’, ’uire’, ’incep’, ’he’, ”’”, ’oase’, ’just’, ’Left’, ’alone’, ’volu’, ’do’, ’his’, ’thing’, ’masura’, ’so’, ’incep’, ’he’, ’feels’, ’like’, ’incep’, ’cumva’, ’man’, ’intr’, ’sanct’, ’house’, ’situatia’, ’incep’, ’cumva’, ’cumva’, ’?’, ’Mega’, ’nnie’, ’:’, ’Tell’, ’them’, ’I’, ’CAN’, ’pay’, ’with’, ’cookies’, ’odata’, ’if’, ’they’, ’Help’, ’me’, ’arra’, ’Rose’, ’anne’, ’slav’, ’drückt’, ’famous’, ’chocolate’, ’cookies’, ’?’, ’Mega’, ’n’, ’Asadar’, ’Why’, ’lucreaza’, ’?’, ’I’, ”’”, ’sustine’, ’feeling’, ’generous’, ’today’, ’incearca’, ’:’, ’D’, ’Rose’, ’anne’, ’Asadar’, ’Guys’, ’uleiul’, ’III’, ’Think’, ’herself’, ”’”, ’urmeaza’, ’cerinte’, ’gonna’, ’change’, ’herself’, ’mind’, ’dispoziti’, ’tanar’, ’Beer’, ’.’, ’Me’, ’g’, ’kindly’, ’Would’, ’YOU’, ’accept’, ’ONE’, ’additional’, ’Help’, ’ers’, ’?’, ’incep’, ’:’, ’D’, ’Mega’, ’n’, ’:’, ’you’, ’datorita’, ’welcome’, ’always’, ’welcome’, ’!’, ’BUT’, ’sorry’, ’boys’, ’datorita’, ’sarbatori’, ”’”, ’Hurricane’, ’gonna’, ’nunta’, ’be’, ’incep’, ’aveti’, ’“’, ’girl’, ’sectiune’, ’only’, ”’’, ’event’, ’(’, ’except’, ’excluding’, ’little’, ’Billy’, ’).’, ’Otherwise’, ’My’, ’house’, ’will’, ’Never’, ’get’, ’cleaned’, ’today’, ’varf’, ’hugs’, ’D’, ’ceilalti’, ’mașin’, ’cît’, ’cît’, ’reusit’, ’pastra’, ’pastra’, ’pastra’]

Memory of Block2

[’gardinen’, ’gardinen’, ’gardinen’, ’gardinen’, ’gardinen’, ’gardinen’, ’gardinen’, ’gardinen’, ’gardinen’, ’gardinen’, ’gardinen’, ’gardinen’]

As we can see the memory tokens in each slot have the same token for first block slot containing 12 tokens of **’Spitalul’**, the second block has 12 tokens of **’gardinen’**. It is worth noting that all memory tokens were first initialized with pad tokens at the beginning of training. For this reason if we compare with the pad tokens at the end of the second block we will see that the pad tokens at the end of encoding stage have different tokens. Pad tokens at the end of the second block were masked but this does not mean that pad tokens in the query did not depend on other block tokens plus they have different positional encoding, unlike the memory tokes that have the same positional encoding for each chunk. Until now we understand that both memory tokens and pad tokens have something in common: memory tokens of the slot depend on the related sequence tokens but not on other memory tokens of other slots, on the other side, pad tokens of the second block depend on the second block tokens but not on pad tokens. As an example of the attention map; figure 1, for the second block on the first layer, head five in the encoder[1]. Also we can note the encoded blocks have some noise some tokens were replaced with different tokens. For that encoder, loss was added to the model loss but that did not help to get better results.

As we can see in the attention map in figure 1 how all memory tokens of the second slot are interaction just with the related chunk, and how the chunk is interacting with all memory tokens in all slots.

[1] To see all the attention maps for all chunks of example 115 refer here.

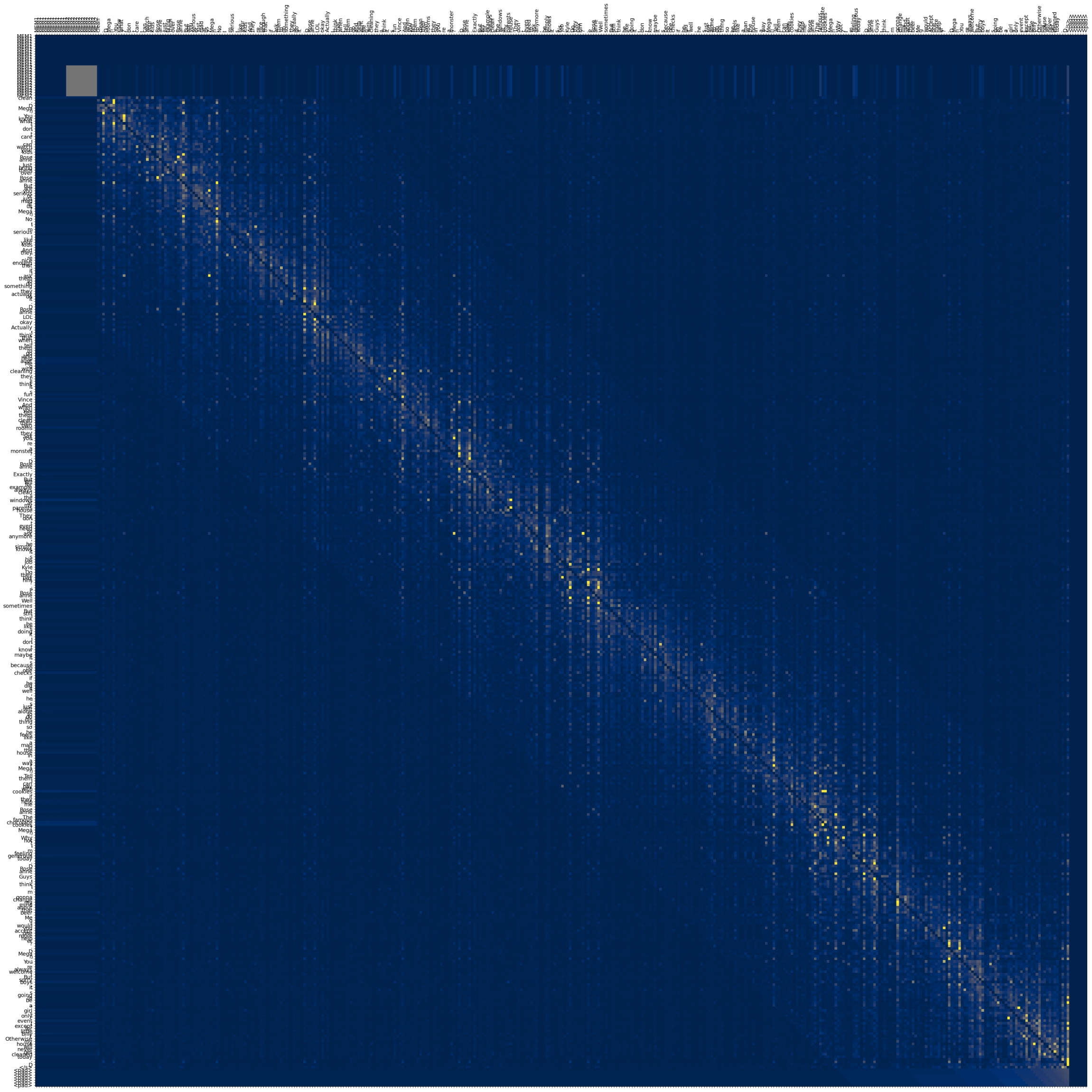

Fig. 1. Attention map of the second block of the example 115 from SAMSum data set
.