Multi-Dimension Attention for Multi-Turn Dialog Generation (Student Abstract)

Billal Belainine, Fatiha Sadat, Mounir Boukadoum

University of Quebec in Montreal
201 President Kennedy, Montréal, QC H2X 3Y7
Quebec, Canada
belainine.billal@courrier.uqam.ca, (sadat.fatiha, boukadoum.mounir)@uqam.ca

Abstract

This paper presents a generative neural model for open and multi-turn dialog response generation that relies on a multi-dimension attention process to account for the semantic inter-dependence between the generated words and the conversational history, so as to identify all the words and utterances that influence each generated response. The performance of the model is evaluated on the wide scope DailyDialog corpus and a comparison is made with two other generative neural architectures, using machine learning metrics. The results show that the proposed model improves the state of the art for generation accuracy, and its multi-dimension attention allows for a more detailed tracking of the influential words and utterances in the dialog history for response explainability by the dialog history.

Introduction

Generative neural architectures with an encoder-decoder architecture have emerged as an efficient approach to long dialog modeling in open conversation systems. The hierarchical encoder-decoder (HRED) architecture (Serban et al. 2016) has played a significant role in this, and several variants have been proposed for improved responses. However, the forgetting problem of the underlying recurrent neural networks (RNNs) has been a major limiting factor for optimal performance. Xing et al. (2018) propose to address the latter problem with the Hierarchical Recurrent Attention Network (HRAN), where an attention mechanism analyzes the hidden states of the encoder to establish short-term links with the input and apply them at the output. Zhang et al. (2019) propose further improvement with ReCoSa (Relevant context with self-attention), where an RNN with self-attention mechanism is used at the encoder and the decoder before measuring the relevance between the response and the local contexts.

Implementation

Figure 1 provides the block diagram of our encoder-decoder model with multi-dimension attention. It is similar in essence to the transformer (Vaswani et al. 2017), but it uses a double position coding scheme at the input to vertically code the word positions within utterances and horizontally code the utterance positions in the dialog. Another difference with the Transformer is the use of a direct representation of the output data instead of a key-value approach, and of the cosine similarity in the attention determination instead of a regular dot-product. These changes result in a lower memory footprint, faster learning and improved sequence transduction performance.

Word and Utterance Position Coding

Two position vectors are added to the input embeddings in order to specify the word and utterance order in the input dialog. One of the vectors codes the word positions within the utterances and the other codes the utterance positions within the dialog. Given an embedding size d_{model}, the vertical and horizontal coding of the word positions in the dialog history generates two vectors:

$$PE_{vert}(pos_w) = \begin{cases} PE_{vert}(pos_w)_{2i} = sin(pos_w/10000^{2i/d_{model}}) \\ PE_{vert}(pos_w)_{2i+1} = cos(pos_w/10000^{2i/d_{model}}) \end{cases}$$

$$PE_{horiz}(pos_u) = \begin{cases} PE_{horiz}(pos_u)_{2j} = cos(pos_u/10000^{(2j)/d_{model}}) \\ PE_{horiz}(pos_u)_{2j+1} = sin(pos_u/10000^{(2j)/d_{model}}) \end{cases} \quad (1)$$

Multi-dimension Attention

At the encoder, our model projects each input embedding with its position code into two representation vectors q and $\acute{q}$, and two matrices Q and $\acute{Q}$ are obtained by stacking the inputs. Then, the cosine similarity of the two matrices is computed before applying the softmax function to the result and weight the elements of $\acute{Q}$:

$$\begin{aligned} \alpha &= Attention_{2D}(Q, \acute{Q}) \\ &= Softmax\left(cos\left(Q\acute{Q}^T\right)\right)\frac{\acute{Q}}{\|\acute{Q}\|} \\ &= Softmax\left(\frac{Q\acute{Q}^T}{\|Q\|\|\acute{Q}^T\|}\right)\frac{\acute{Q}}{\|\acute{Q}\|} \end{aligned} \quad (2)$$

Given the cosine image, there is no need to scale the softmax argument as done in the Transformer. The self-attention described by equation 2 is performed over h heads by setting the size of the q and $\acute{q}$ vectors to d_{model}/h. Then, the results are concatenated before a final projection to produce the encoder's output. The whole process is repeated for all the utterances in the input batch and a stack of self-attention matrices is obtained at the end, with the embedded position codes in each one specifying the words and utterances re-

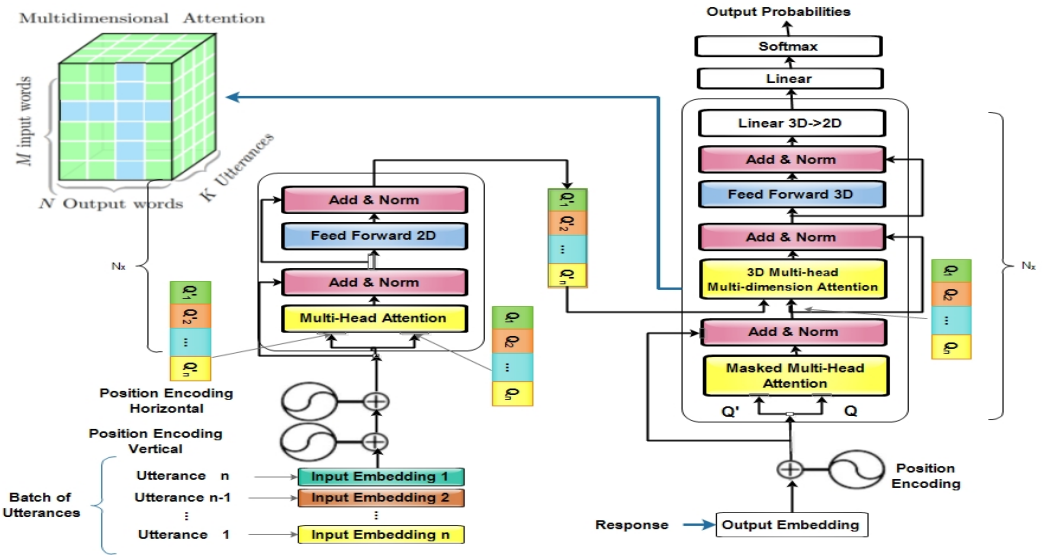

Figure 1: Computation graph of the proposed Multi-dimension Transformer model for conversation learning

ferred to. The decoder start in the same way as the encoder, with the already generated outputs as input and using a mask in the self-attention calculations to ignore not yet generated words. Then, an encoder-decoder attention is computed between the stack of outputs from the encoder and the result, therefore creating a 3D attention for predicting the decoder's output. Thus, given the stack of matrices $\acute{Q}_i$ from the encoder's output, with $i = 1, \cdots, n$, and the self-attention matrix Q of the decoder, we have:

$$\begin{aligned}\alpha &= Attention_{3D}(Q_{i<=n}, \acute{Q}) \\ &= Softmax_{out}\left(\left[cos\left(Q_1\acute{Q}^T\right), \cdots, cos\left(Q_n\acute{Q}^T\right)\right]\right)\frac{\acute{Q}}{\|\acute{Q}\|} \\ &= Softmax_{out}\left(\left[\frac{Q_1\acute{Q}^T}{\|Q_1\|\|\acute{Q}^T\|}, \cdots, \frac{Q_n\acute{Q}^T}{\|Q_n\|\|\acute{Q}^T\|}\right]\right)\frac{\acute{Q}}{\|\acute{Q}\|}\end{aligned} \tag{3}$$

We call this mechanism "multi-dimension attention" (MDA), since it reflects the dependencies between the decoder's output and all the words in the dialog history at once. We evaluated our Transformer with MDA model as well as three other models used for reference: (i) the Hierarchical Recurrent Encoder-Decoder (HRED) (Serban et al. 2016); (ii) the Relevant Contexts with Self-Attention (ReCoSa) (Zhang et al. 2019); (iii) the Basic Transformer (Vaswani et al. 2017) using concatenating multiple utterances together. The automatic machine evaluation included perplexity and the three embedding-based similarity measures proposed by

Corpus	**DailyDialog**				
Model	Average	Greedy	Extrema	PPL	BLEU
Transformer	0.514	0.372	0.311	83.41	0.42
HRED	0.636	0.459	0.391	85.08	0.56
ReCoSa	0.626	0.421	0.321	84.98	1.72
MDA	**0.671**	**0.524**	**0.398**	**79.62**	**2.87**

Table 1: Model comparison using the embedding metrics, perplexity and BLEU score

(Zhang et al. 2019): average embedding, extrema inclusion and greedy integration.evaluation. The perplexity metric measured the model's ability to account for the syntactic structure of the dialog (e.g., the turn-taking) and the syntactic structure of each utterance (e.g., the punctuation marks) (Serban et al., 2016a). The embedding metrics offer three ways to assess the similarity between the words in the model response and the ground truth.

Evaluation Results

Table 1 shows the obtained performances by the different models. It appears then that using a single encoder as out model does, with no hierarchical processing or additional modules can lead to more accurate prediction results. Our model also surpasses the other two for perplexity, with ReCoSa ranking second and HRED ranking last. This confirms the lower perplexity of attention-based models over RNN-based models

References

Serban, I. V.; Sordoni, A.; Bengio, Y.; Courville, A. C.; and Pineau, J. 2016. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models. In *AAAI*, 3776–3784.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Xing, C.; Wu, Y.; Wu, W.; Huang, Y.; and Zhou, M. 2018. Hierarchical Recurrent Attention Network for Response Generation. In *Proceedings of the AAAI*, volume 32.

Zhang, H.; Lan, Y.; Pang, L.; Guo, J.; and Cheng, X. 2019. ReCoSa: Detecting the Relevant Contexts with Self-Attention for Multi-turn Dialogue Generation. In Korhonen, A.; Traum, D. R.; and Màrquez, L., eds., *Proceedings of ACL 2019, Florence, Italy, July 28- August 2, 2019*, 3721–3730.