Abstractive Text Summarization using Encoder-Decoder Architecture

Naman Jaswani [1] **Rohit Panda** [2]

Abstract

In this paper, we analyse the results of Encoder-Decoder Architecture(9),(3) with Bahdanau attention (1) on abstractive text summarization task. We propose modifications to this baseline architecture using various attention mechanisms. We propose a metric based on cosine similarity, and show that it correlates well with BLEU scores(6).

1. Technical Details

1.1. Dataset

We use *Amazon Fine Food Reviews* dataset (5), which has columns for food reviews and their corresponding summary for around 500000 data points.

1.2. Baseline Model

An Encoder-Decoder model with Bahdanau attention(1).

1.2.1. Encoder

An Encoder reads a sequence of vectors $\mathbf{x} = (x_1, \cdots, x_T)$, & correspondingly generates hidden states $h_t = f(x_t, h_{t-1})$ for each time step t; where f is GRU cell(3).

1.2.2. Bahdanau Attention

Attention mechanism gives **context vector** representing importance of every Encoder state to a decoder state, at timestep t. In Bahdanau attention, score of each Decoder state h_t for the sequence of Encoder outputs h_s is given by: $score\left(h_t, \overline{h}_s\right) = v_a^\top \tanh\left(W_1 h_t + W_2 \overline{h}_s\right)$.

1.2.3. Decoder

Decoder is trained to predict the next word y_t given all previous $\{y_1, \cdots, y_{t-1}\}$ and the context vector c from attention.

1.2.4. Specifications

The results in Table 1 are produced using following specifications. We use pretrained *GloVe* embeddings(7) with an embedding size of 100. For **Encoder**, we use a *Bidirectional GRU* (8) and for **Decoder**, we use a *GRU* model(3). **Learning rate** is set to $3e^{-4}$, and the source and target maximum allowed lengths are fixed at 500 and 15 respectively.

2. Novel Contributions

Self Attention[1]**:** To capture the dependencies between various encoder states, we introduce *Self-Attention*(10) on top of Encoder outputs.

Attention on Decoder[2]**:** To generate contextually better sequences, we introduce *Bahdanau attention*(1) on already generated tokens, while predicting the next token.

Multihead Attention[1,2]**:** Taking inspiration from Transformers(10) which uses Multihead-attention, we introduce multiple Bahdanau attentions on Encoder states.

Variation to Multihead Attention[1]**:** We introduce a variation to multihead attention by combining one Bahdanau attention(1) model and one Luong attention(4) model. ($luong\ attention\ score\left(h_t, \overline{h}_s\right) = h_t^\top W \overline{h}_s$).

Metric on Cosine Similarity[2]**:** Summary is subjective and Bleu score(6) which does exact match, doesn't do justice. We introduce an equivalent which computes cosine similarity between embeddings of output and target tokens instead.

3. Results

Three of our Four variations perform similar to the Baseline in terms of Bleu score and Our metric without surpassing it, whereas Multihead Attention variation surpasses Baseline.

Model	BLEU score
Baseline	22.86
Multihead Attention	**25.41**

Model	Our Metric
Baseline	0.482
Multihead Attention	**0.516**

3.1. Sample Results

The following output sequences are generated for an input sentence (109 words) of a customer review on coffee.

Target: "nice stuff"

Baseline output: "dreamfields addition for the morning"

Self-Attention output: "my little break to brew"

Tools

- **Python** Version: 3.9.5
- **Torch** Version: 1.8.1 + cu111
- **Torchtext** Version: 0.9.1
- **Numpy** Version: 1.18.5
- **Pandas** Version: 1.2.4
- **BLEU Score** Version: 0.3
- **ROUGE Scores** Version: 1.0.0

References

[1] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2015.

[2] J. Bastings. The annotated encoder decoder.

[3] K. Cho, B. van Merrienboer, Ç. Gülçehre, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078, 2014.

[4] M. Luong, H. Pham, and C. D. Manning. Effective approaches to attention-based neural machine translation. *CoRR*, abs/1508.04025, 2015.

[5] J. McAuley and J. Leskovec. From amateurs to connoisseurs: Modeling the evolution of user expertise through online reviews. 2013.

[6] 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*, ACL '02, page 311–318, USA, 2002. Association for Computational Linguistics.

[7] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[8] M. Schuster and K. Paliwal. Bidirectional recurrent neural networks. *Signal Processing, IEEE Transactions on*, 45:2673 – 2681, 12 1997.

[9] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215, 2014.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.