Towards Understanding Position Embeddings

Rishi Bommasani
Department of Computer Science
Cornell University
Ithaca, NY, 14853, USA
rb724@cornell.edu

Claire Cardie
Department of Computer Science
Cornell University
Ithaca, NY, 14853, USA
cardie@cs.cornell.edu

Abstract

This work describes initial work on probing position representations for understanding how they factor into current neural models. We primarily consider absolute position embeddings learned in the context of pre-trained self-attentive models. We observe that position embeddings may be an attractive object for study given that position is constrained therefore pose the question of whether position embeddings may be an interesting candidate for pretraining in transfer learning regimes (given that the position feature space may display greater structure than the word meaning space where pretrained embeddings are already effective for transfer learning). All results and complete experimental conditions are made available.[1]

1 Introduction

Neural models for natural language processing (NLP) have been applied to a wide array of tasks with increasing success over recent years. By their nature, these models lessen the need for explicit feature engineering and instead require innovation to model discrete features of natural language in a manner conducive for neural computation. A fundamental primitive in this process is embedding methods, which learn mappings from discrete feature spaces to continuous fixed-dimensional vector spaces (Collobert et al., 2011). Embedding methods have been applied successfully for representing words (Pennington et al., 2014; Mikolov et al., 2013), sentences (Peters et al., 2018; Conneau et al., 2017), and positional features (Vaswani et al., 2017). However, these embedding methods have only been rigorously analyzed in the case of words and sentences (Mimno and Thompson, 2017; Conneau et al., 2018) whereas positional embeddings/encodings have been comparatively understudied.

While embedding methods for words and sentences are naturally motivated given the empirical performance of neural networks and other machine learning algorithms that leverage distributed representations, the reasoning for position embeddings is more of a recent artifact of minimizing sequential computation.
Position embeddings have seen two key treatments in recent work:

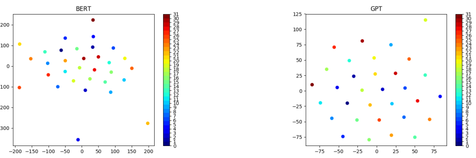
Absolute Position - Self-attentive encoders have recently seen widespread adoption as explicit encoders for a diverse class of tasks (Werlen et al., 2018) or as pre-trained contextualizers (Liu et al., 2019) for downstream problems. One fundamental reason is these models are conducive for modern hardware because they minimize/remove sequential processing that is a hallmark of a recurrent models. Position embeddings have been a key part of this shift by providing an alternative mechanism to model the sequential nature of language.

Relative Position - In unrelated work to self-attentional approaches, relative position with respect to linguistic objects have been considered for attention in relation extraction (Zhang et al., 2017). Recently, Dai et al. (2019) proposed a modification to the approach of Vaswani et al. (2017) by modelling position relatively. Relative position encoding is seen as a key innovation towards variable-length context modelling in this work.
A key motivation for considering position embeddings is that the underlying feature of position can be useful for both syntactic and semantic whereas conventionally word/sentence representations have been primarily applied to semantic tasks even if they can successfully model syntax (Goldberg, 2019). In this work, we pose three central questions towards understanding current

[1] https://github.com/rishibommasani/PositionEmbeddings

		Reference	
	Human	BERT	GPT
Human	100	21.19	18.84
BERT	25.42	100	48.05
GPT	22.78	48.46	100

Table 1: Average token alignment is given by the percentage of tokens in the reference that match the token at the same position in the candidate.



(a) BERT (b) GPT

Figure 1: t-SNE visualizations for BERT and GPT position embeddings

methods for position embedding and provide introductory results in tackling the first question:

1. How are current position embeddings related?
2. How should we encode position?
3. Are position embeddings transferrable?

2 Analysis

As a proxy for general understanding of embeddings for absolute position in self-attentive encoders, we consider the position embeddings in two recent models: GPT (Radford et al.), BERT (Devlin et al., 2018). Both models embed absolute position in $\mathbb{R}^{784}$ and consider 512 positions (GPT-2 considers 1024). We use embeddings from pretrained models[2] and note that the pretraining data and original initializations of these models may influence our analysis. For visual clarity, we report results on the first 32 positions in this paper.

Tokenizer Alignment In studying position embeddings for different models, a natural assumption would be to expect similarities in the representation for position i in models A and B. However, since different models use different tokenization schemes, it is necessary to understand the alignment in tokenization distributionally to understand the validity of this assumption. In particular, BERT uses WordPiece tokenization (Wu et al., 2016) for English whereas GPT uses BPE (Sennrich et al., 2016). To analyze this, we study tokenization of 1000000 Wikipedia sentences.[3] We also consider alignment with tokenizing by whitespace as this is often how humans (such as the authors) reason about absolute position implicitly.

Table 1 provides corpus level statistics on the token alignment and demonstrates that for all pairs of tokenizers, the exact tokens being represented at a specific position are quite different. This implies that we cannot make judgments about position embeddings without considering that they are frequently being used to describe fundamentally different tokens.

Embedding visualization Word embedding methods have been shown to capture human judgments about lexical relatedness and similarity. Unlike words, positions do not evoke the same types of human-interpretable properties and proposing vector operations to codify some relationship is less clear. As an example, while it may be clear what we hope for $\vec{queen} - \vec{woman} + \vec{man}$ to be, it is less clear what the closest position embedding should be for the sum and difference of various position embeddings. We therefore visualize word embeddings using t-SNE (Maaten and Hinton, 2008) for both BERT and GPT to help grapple with this. Figure 1 shows starkly different clustering behavior for boths sets of embeddings. The BERT embeddings appear to be more tightly clustered and we see that the neighbor sets for the same positions are quite dissimilar across the two plots.

3 Conclusion

We provide introductory results for better understanding position embeddings. We find that while position may be a ubiquitous feature, it is is clear that current position embeddings cannot be studied without controlling for the impacts of tokenization and that there are pronounced differences in the geometry of position embeddings even between embeddings that model absolute position for self-attentive architectures.

[2] https://github.com/huggingface/pytorch-pretrained-BERT

[3] https://blog.lateral.io/2015/06/the-unknown-perils-of-mining-wikipedia/

References

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(Aug):2493–2537.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $ &!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136.

Zihang Dai, Zhilin Yang, Yiming Yang, William W Cohen, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

David Mimno and Laure Thompson. 2017. The strange geometry of skip-gram with negative sampling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2873–2878.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1715–1725.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Lesly Miculicich Werlen, Nikolaos Pappas, Dhananjay Ram, and Andrei Popescu-Belis. 2018. Self-attentive residual decoder for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1366–1379.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45.