Week 2 : Attention is All you Need

120220210 고재현

2023년 3월 22일

1 Introduction

Seq2Seq-Attention 모델은 Neural Machine Translation(NMT)에서 가장 유명한 모델 중 하나였다. Seq2Seq에 Attention 메커니즘을 추가함으로써 기존의 Seq2Seq 모델의 Long-Term Memory에 관한 문제점을 해결하고자 했다. 본 논문에서 Transformer 모델 [1, Transformer]을 제안함으로써 Seq2Seq 구조를 완전히 제거함으로써 문제를 해결하였다.

2 Transformer Architecture

트랜스포머의 수학적 표기는 [2]에 설명되어 있다. 논문에서는, 트랜스포머를 다중 self-attention 블록으로 구성된 새로운 인코더-디코더 아키텍처로 소개한다. 각 트랜스포머 레이어의 입력을 $\mathbf{X} \in \mathbb{R}^{n\times d}$라 하면 ($n$은 토큰 수, d는 각 토큰의 차원), 한 블록 레이어는 $f_\theta : \mathbb{R}^{n\times d} \to \mathbb{R}^{n\times d}$ 함수로 정의되고, $f_\theta(\mathbf{X}) =: \mathbf{Z}$로 정의되며, 다음과 같다.

$$\mathbf{A} = \frac{1}{\sqrt{d}}\mathbf{X}\mathbf{Q}(\mathbf{X}\mathbf{K})^\top, \tag{1}$$

$$\widetilde{\mathbf{X}} = \mathrm{SoftMax}(\mathbf{A})(\mathbf{X}\mathbf{V}), \tag{2}$$

$$\mathbf{M} = \mathrm{LayerNorm}_1(\widetilde{\mathbf{X}}\mathbf{O} + \mathbf{X}), \tag{3}$$

$$\mathbf{F} = \sigma(\mathbf{M}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2, \tag{4}$$

$$\mathbf{Z} = \mathrm{LayerNorm}_2(\mathbf{M} + \mathbf{F}), \tag{5}$$

여기서 식 1, 식 2, 식 3은 어텐션 계산을 나타내며, 식 4, 식 5은 위치별 feed-forward 네트워크 (FFN) 레이어이다. 여기에서 $\mathrm{Softmax}(\cdot)$은 row-wise softmax 함수를, $\mathrm{LayerNorm}(\cdot)$은 layer normalization 함수를 참조하며, σ는 활성화 함수를 나타낸다. $\mathbf{Q}, \mathbf{K}, \mathbf{V}, \mathbf{O} \in \mathbb{R}\mathbb{R}^{d\times d_f}$, $\mathbf{b}_1 \in \mathbb{R}^{d_f}$, $\mathbf{W}_2 \in \mathbb{R}^{d_f\times d}$, $\mathbf{b}_2 \in \mathbb{R}^d$는 레이어의 학습 가능한 매개변수이다. 또한, 자기 어텐션을 다중 어텐션으로 확장하기 위해 여러 어텐션 헤드를 고려하는 것이 일반적이다. 구체적으로, $\mathbf{Q}, \mathbf{K}, \mathbf{V}$는 H 개의 헤드로 분해되며, $d = \sum_{h=1}^{H} d_h$로 각각 $\mathbf{Q}^{(h)}, \mathbf{K}^{(h)}, \mathbf{V}^{(h)} \in \mathbb{R}^{d\times d_h}$로 표기된다. 어텐션 헤드에서 나온 행렬 $\widetilde{\mathbf{X}}^{(h)} \in \mathbb{R}^{n\times d_h}$는 연결(concatenated)하여 $\widetilde{\mathbf{X}}$를 얻는다. 이 경우, 식 1과 식 2는 각각 다음과 같이 수정될 수 있다.

$$\mathbf{A}^{(h)} = \frac{1}{\sqrt{d}}\mathbf{X}\mathbf{Q}^{(h)}(\mathbf{X}\mathbf{K}^{(h)})^\top,\ \widetilde{\mathbf{X}} = |_{h=1}^{H}(\mathrm{SoftMax}(\mathbf{A}^{(h)})\mathbf{X}\mathbf{V}^{(h)}). \tag{6}$$

다중 헤드 메커니즘은 모델이 다른 측면에서 표현을 암묵적으로 학습할 수 있도록 한다. 어텐션 메커니즘 외에도, 이 논문은 sine 및 cosine 함수를 서로 다른 주파수로 사용하여 각 토큰의 위치를 구분하기 위한 위치 임베딩으로 사용한다.

위치 임베딩은 각 토큰의 상대적인 위치 정보를 모델에 전달하기 위해 사용되며, 코사인 함수와 사인 함수의 주기적인 변화를 이용하여 표현된다. 위치 pos와 임베딩 차원 i에 대한 위치 임베딩 값 $PE_{pos,i}$은 다음과 같이 정의된다.

$$PE(pos, 2i) = sin(pos/(10000^{(2i/d)})) \tag{7}$$

$$PE(pos, 2i+1) = cos(pos/(10000^{(2i/d)})) \tag{8}$$

여기서 d는 임베딩 차원의 크기를 나타낸다. 위치 임베딩은 입력 토큰의 임베딩 벡터와 더해져 각 토큰의 상대적인 위치 정보를 반영한 새로운 임베딩 벡터를 생성한다.

참고 문헌

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[2] E. Min, R. Chen, Y. Bian, T. Xu, K. Zhao, W. Huang, P. Zhao, J. Huang, S. Ananiadou, and Y. Rong, "Transformer for graphs: An overview from architecture perspective," *arXiv preprint arXiv:2202.08455*, 2022.