page 1. VIT Contribution

- 완전히 Convolution과 안녕하고 SoTA를 달성했다

- Vision Transformer에서 중요한 것은 세 가지이다.

① 이미지 패치(patch)를 단어와 같이 다루었다.

② 아키텍처는 Transformer의 엔코더 부분이다.

③ 거대한 데이터 세트인 JFT-300M으로 사전학습했다.

- SoTA보다 뛰어난 성능을 약 15분의 1의 계산 비용만으로 얻을 수 있었다.

- 사전학습 데이터 세트와 모델을 더욱 크게 하여 성능을 향상시킬 수 있는 여지가 있다.

page 2 What is transformer

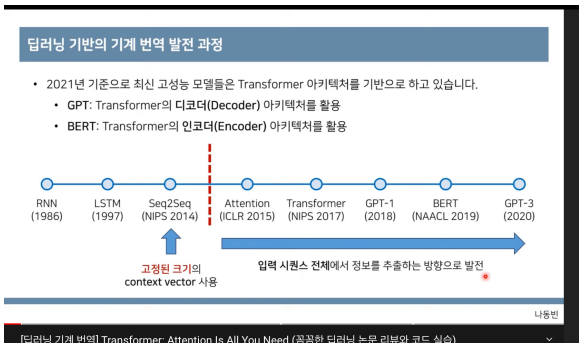
Transformer 의의

○ **Transformer 기여**

- 기존의 Sequence Transduction(변환) 모델은 인코더(Encoder)와 디코더(Decoder)를 포함하는 구조를 바탕으로, 순환 신경망(Recurrent)나 Convolutional Layer를 사용함
- **좋은 성능을 보인 모델의 특징** : Attention 메커니즘을 활용해서, 인코더와 디코더를 연결한 모델
- **Attention 메커니즘만을 사용하는 "Transformer"라는 새로운 구조를 제안**
  1. 기계번역(Machine Translation) Task에서 매우 좋은 성능
  2. 학습 시, 우수한 병렬화(Parallelizable) 및 훨씬 더 적은 시간 소요
  3. 구문분석(Constituency Parsing) 분야에서도 우수한 성능 → 일반화(Genelization)도 잘됨

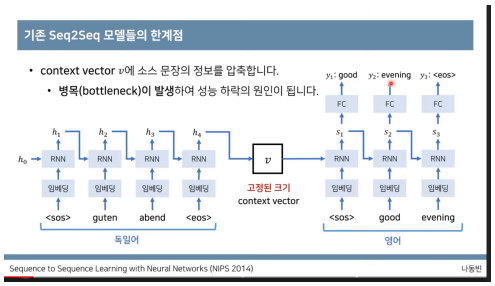
※ 구문분석이란?

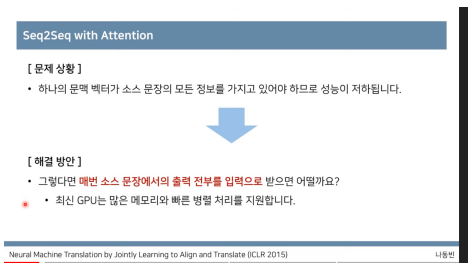
page 3 기계번역(MT)에서 transformer 발전 과정



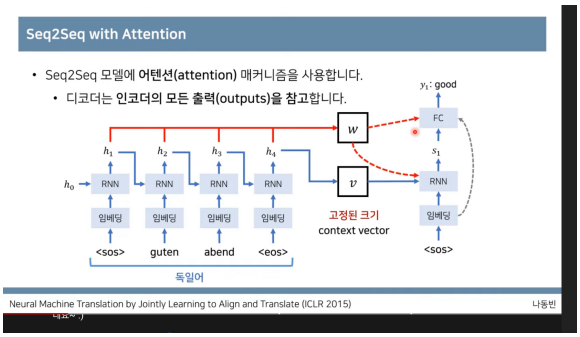
page 4 Related Work (Seq2Seq)

Seq2Seq 아키텍처 그림 + Seq2Seq limitation + 해결방안





page 5 Related work (Seq2Seq with Attention)



page 6 Related work (Transformer : Attention is All you need)

위 그림에서 RNN 제거

page 7 Attention Architecture

아키텍처 그림

page 8 ~11 or 12 How the Attention work

Query Key Value 설명

page 13 ~ Vision transformer 논문 정리

https://engineer-mole.tistory.com/133

↳ 되게 깔끔하게 정리 잘 되어 있어서 이 흐름대로 만들어짐