<1>

데이터셋 선정 배경부터 말씀드리겠습니다. 연구팀은 트랜스포머의 성능을 입증하기 위해 의도적으로 성격이 다른 두 가지 데이터셋을 선택했습니다.

첫째, 영어**-**독일어**(WMT 2014)** 데이터셋입니다. 독일어는 복잡한 문법과 자유로운 합성어 구조로 인해 기계번역 분야에서 전통적인 **'난제(Hard Task)'**로 꼽힙니다. 연구팀은 이 데이터셋을 통해 트랜스포머의 Self-Attention이 기존 RNN 모델들이 어려워했던 **'장기 의존성(Long-term Dependency)'** 문제를 얼마나 효과적으로 해결하는지를 증명하고자 했습니다.

둘째, 영어**-**프랑스어 데이터셋은 입니다. 해당 데이터셋은 3,600만 문장이라는 방대한 규모를 자랑합니다. 이는 트랜스포머가 대규모 데이터 학습에서도 뛰어난 확장성을 가짐을 보여줍니다.

학습 최적화(Optimization) 부분에서도 흥미로운 디테일이 있습니다. **Adam Optimizer**를 사용하되, 그레디언트 크기를 추적하는 beta2 계수를 일반적인 0.999가 아닌 **0.98**로 조정했습니다. 이는 Attention 메커니즘 특유의 급격한 그래디언트 변화에 대응하여, 학습 초기의 불안정성을 잡고 수렴 속도를 높이기 위한 엔지니어링적 선택이었습니다.

또한 내부 구조적으로는 인코더와 디코더에 **Residual Connection**과 **Dropout**을 적용하여 깊은 신경망 학습 시 발생할 수 있는 기울기 소실 및 오버피팅팅 문제들을 예방했습니다.

<2>
트랜스포머 팀은 최적의 성능을 찾기 위해 **Base** 모델을 기준으로 다양한 실험을 진행했습니다. 레이어의 깊이, 임베딩 차원, FFN의 확장 차원, 그리고 헤드의 개수 등을 세밀하게 조정해 보았는데요. 이 과정을 통해 혼란도인 **Perplexity(PPL)**는 낮추고 기계 번역의 품질 지표인 **BLEU** 점수는 극대화하는 최적의 하이퍼파라미터 조합을 찾아낼 수 있었습니다.

<3>
표를 보시면 모델들과 모델들의 영어-독일어, 영어-프랑스어 번역 지수(BLEU), 그리고 Training Cost가 나와있습니다. Tranformer는 다른 모델들에 비해서 매우 적은 Training Cost를 가지고도
다른 모델보다 높은 BLEU 점수를 보여주고 있습니다. 특히 기존 구글의 RNN 기반 번역기인 GNMT보다 더 높은 BLEU 점수와 낮은 Training Cost를 보여줍니다.

논문의 마지막 부분을 보시면 저자들은 트랜스포머가 이미지, 비디오, 오디오 등으로 확장될 미래에 대해서 매우 기대가 된다고 밝혔습니다. 그리고 실제로 저자들은 각자 저마다의 비전을 실현시키기 위해서 구글을 퇴사하기도 하였습니다. 그러면 이제 트랜스포머에서 파생된 모델들을 살펴보겠습니다.

<4>
각각의 사진들은 BERT, GPT, VIT 모델들입니다.

(1)BERT는 트랜스포머의 **인코더(Encoder)**만을 떼어내어 만든 모델입니다. 다만, 트랜스포머의 인코더가 번역을 위해 문장의 맥락에 집중했다면, BERT는 언어의 **'문맥과 관계를 이해'**하는 데 초점을 맞췄습니다. 이를 위해 BERT는 기존 인코더 아키텍처에 두 가지 핵심 기능을 추가했습니다.

첫째, 두 문장 사이의 관계를 파악하기 위해 문장을 구분해 주는 **Segment Embedding**을 Embedding 과정에 추가하였고, 두번째로 해당 문장을 함축적으로 표현하는 **`[CLS]` Tokenizing**을 진행합니다.

이러한 구조 덕분에 BERT는 단순한 단어 임베딩을 넘어, 문장의 의미와 관계까지 함축하는 **'범용 언어 특징 추출기(Feature Extractor)'**로서 LLM이 자연어를 처리할 수 있는 기반을 마련했습니다.

(2)두 번째로 살펴볼 모델은 GPT입니다. GPT는 앞서 본 BERT와 반대로, 트랜스포머의 디코더(Decoder) 부분만을 떼어내어 만든 아키텍처입니다.

가장 큰 구조적 특징은 인코더가 사라졌다는 점입니다. 이에 따라, 기존 디코더에서 인코더의 정보(Key, Value)를 받아오던 연결 고리인 '인코더-디코더 어텐션(또는 Cross-Attention)' 층도 함께 제거되었습니다.

학습 방식 또한 다릅니다. GPT는 **'다음 토큰 예측(Next Token Prediction)'**이라는 방식을 사용합니다. 즉, 정답(Ground Truth) 텍스트를 보여주고, 모델이 **지금까지 나온 단어들의 맥락(Context)**을 파악하여 그다음에 올 가장 적절한 단어를 맞추도록, 역전파를 통해 파라미터를 최적화합니다.

(3)
마지막으로 살펴볼 모델은 **비전 트랜스포머(ViT)**입니다.

기존의 CNN 기반 모델들은 '인접한 픽셀끼리 연관성이 높다'는 **지역적 특성(Locality)**에 집중했습니다. 하지만 이 때문에 반대로 이미지 전체를 아우르는 **전역적인 문맥(Global Context)**을 파악하는 데는 한계가 있다는 지적을 받아왔습니다.

ViT는 이 문제를 해결하기 위해, 트랜스포머의 인코더를 비전 분야에 도입했습니다. 핵심은 이미지를 텍스트의 단어처럼 **'패치(Patch)'**라는 작은 조각으로 자른 뒤, 이를 일렬로 나열하여 인코더에 입력했다는 점입니다.

이 과정을 통해 ViT는 CNN처럼 국소적인 정보에 갇히지 않고, 이미지 전체의 관계를 한 번에 학습할 수 있게 되었습니다.

이상으로 저희 5조는 트랜스포머 아키텍처의 구조적 특징과, 이를 바탕으로 발전한 BERT, GPT, ViT와 같은 파생 모델들에 대해 알아보았습니다.

2017년 구글 리서치 팀이 논문을 발표하고, 2024년 GTC 에서 저자들이 다시 한 자리에 모이기까지... AI는 생활의 많은 것들을 바꿔 놓았습니다.

그리고 그 AI 기술들의 핵심 엔진인 '트랜스포머'에 대해 깊이 있게 공부해 볼 수 있어 매우 뜻깊은 기회였습니다. 발표를 들어주셔서 감사합니다.

TMI)

- 트랜스포머의 가중치 초기화 기법: Xavier Initialization.
  레이어 간 신호의 분산(Variance)을 일정하게 유지하여, 학습 초기의 기울기 소실 및 폭발을 방지하는 초기화 방법. -> 그 당시 NLP 국룰 기법.

- Attention에 sqrt(d_k)로 나눈 값을 넣은 이유?: Softmax 특성상 입력값이 클수록 1에 가까워진다. 그리고 1에 가까워질수록 기울기가 0이되어 vanishing gradient가 생길 수 있기 때문에 의도적으로 헤더의 차원에 제곱근을 씌우고 나눠서 이를 방지한다.

- GeLU 함수:
  0이하부터는 스위치를 꺼버리는 ReLU와 달리 음수값도 약간의 조정을 한 다음 보내기 때문에 기울기가 완전히 0이 되는 상황을을 막아준다