Transformer-based Seq2Seq and Sparsity & Linearity

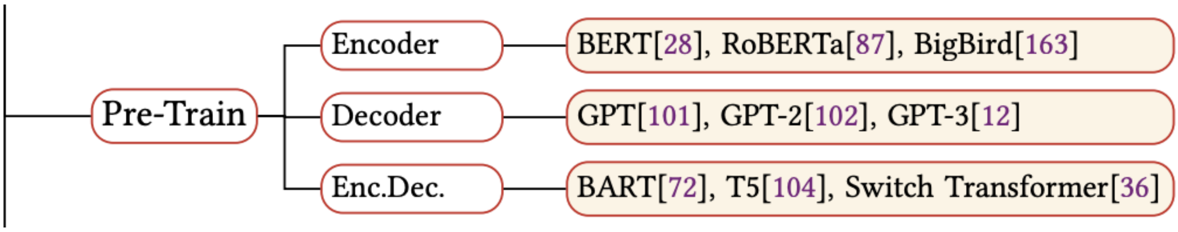
Author	@Dyan Lee
Created By	Ⓓ Dyan Lee
Date	@2021년 9월 10일
Organization	KAIST EdLab.
Presented @	
Published	Sep 2021
Tags	Kick-Off
Tags 1	Study
Transcript	
Type	Weekly Sync

Agenda

- Walkthrough a survey papers on Transformer, Efficient Transformer, Visual Transformer, and etc.
- Seek out what must be done first.
- Primary agenda will be Performer and more.

Summary

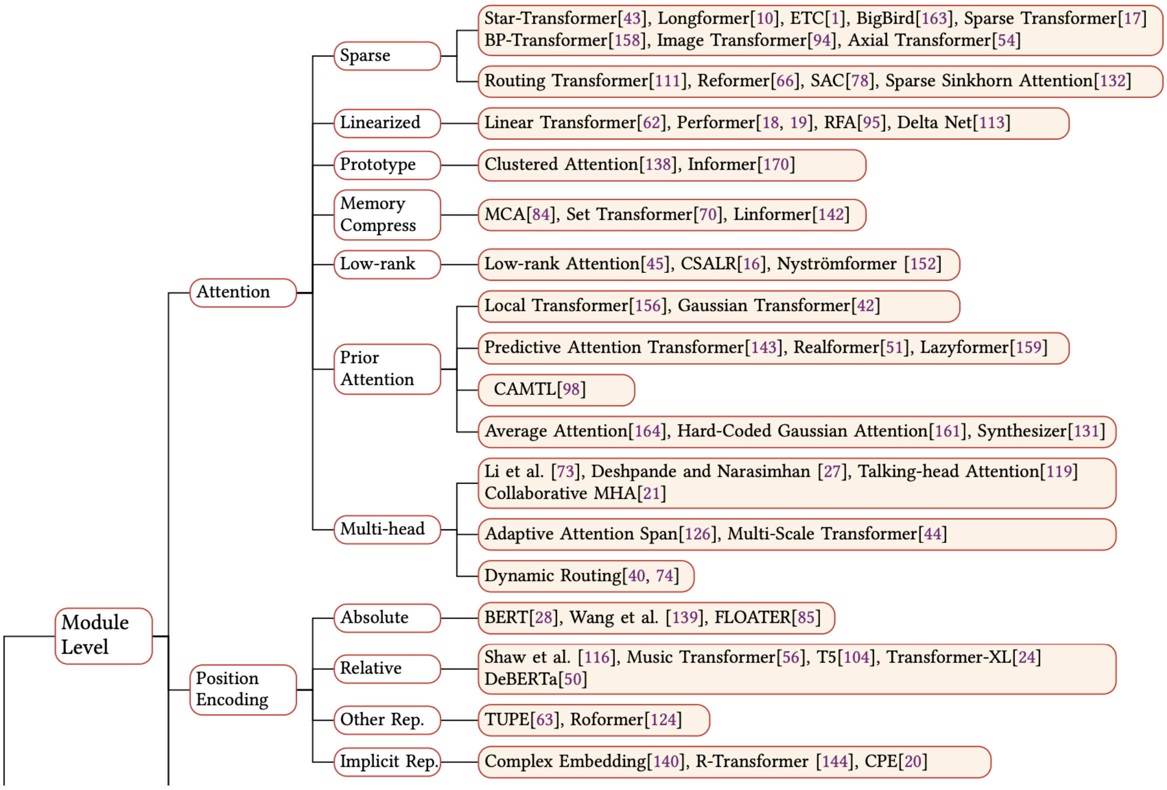
Seq2Seq Models


Pre-Train
Encoder
Decoder
Enc.Dec.
BERT[28], RoBERTa[87], BigBird[163]
GPT[101], GPT-2[102], GPT-3[12]
BART[72], T5[104], Switch Transformer[36]

Pre-Train

- BART[72], T5[104], Switch Transformer[36]
- gpt

Attention


Module Level
Attention
Position Encoding
Sparse
Linearized
Prototype
Memory Compress
Low-rank
Prior Attention
Multi-head
Absolute
Relative
Other Rep.
Implicit Rep.
Star-Transformer[43], Longformer[10], ETC[1], BigBird[163], Sparse Transformer[17] BP-Transformer[158], Image Transformer[94], Axial Transformer[54]
Routing Transformer[111], Reformer[66], SAC[78], Sparse Sinkhorn Attention[132]
Linear Transformer[62], Performer[18, 19], RFA[95], Delta Net[113]
Clustered Attention[138], Informer[170]
MCA[84], Set Transformer[70], Linformer[142]
Low-rank Attention[45], CSALR[16], Nyströmformer [152]
Local Transformer[156], Gaussian Transformer[42]
Predictive Attention Transformer[143], Realformer[51], Lazyformer[159]
CAMTL[98]
Average Attention[164], Hard-Coded Gaussian Attention[161], Synthesizer[131]
Li et al. [73], Deshpande and Narasimhan [27], Talking-head Attention[119] Collaborative MHA[21]
Adaptive Attention Span[126], Multi-Scale Transformer[44]
Dynamic Routing[40, 74]
BERT[28], Wang et al. [139], FLOATER[85]
Shaw et al. [116], Music Transformer[56], T5[104], Transformer-XL[24] DeBERTa[50]
TUPE[63], Roformer[124]
Complex Embedding[140], R-Transformer [144], CPE[20]

Sparsity

- **Star-Transformer**[43], **Longformer**[10], **ETC**[1], **BigBird**[163], **Sparse Transformer**[17]
  BP-Transformer[158], **Image Transforme**r[94], Axial Transformer[54]
- Routing Transformer[111], **Reformer**[66], SAC[78], **Sparse Sinkhorn Attention**[132]

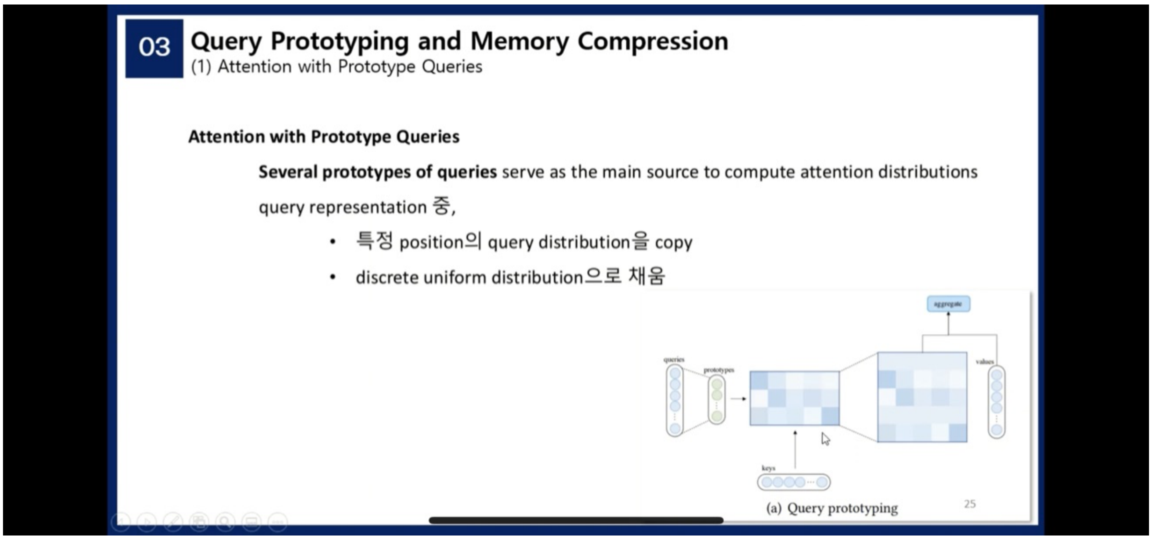
Linearised

- **Linear Transformer[62], Performer[18, 19], RFA[95], Delta Net[113]**

1. **Linear Transformers**
    - Feature Maps에 따른 분류.
    - Feature space 상에서의 orthogonality를 이용할 수 있는 feature map을 제안했다.
2. Performers
    -

Prototype

- Clustered Attention[138], **Informer**[170]



03 Query Prototyping and Memory Compression

(1) Attention with Prototype Queries

Attention with Prototype Queries

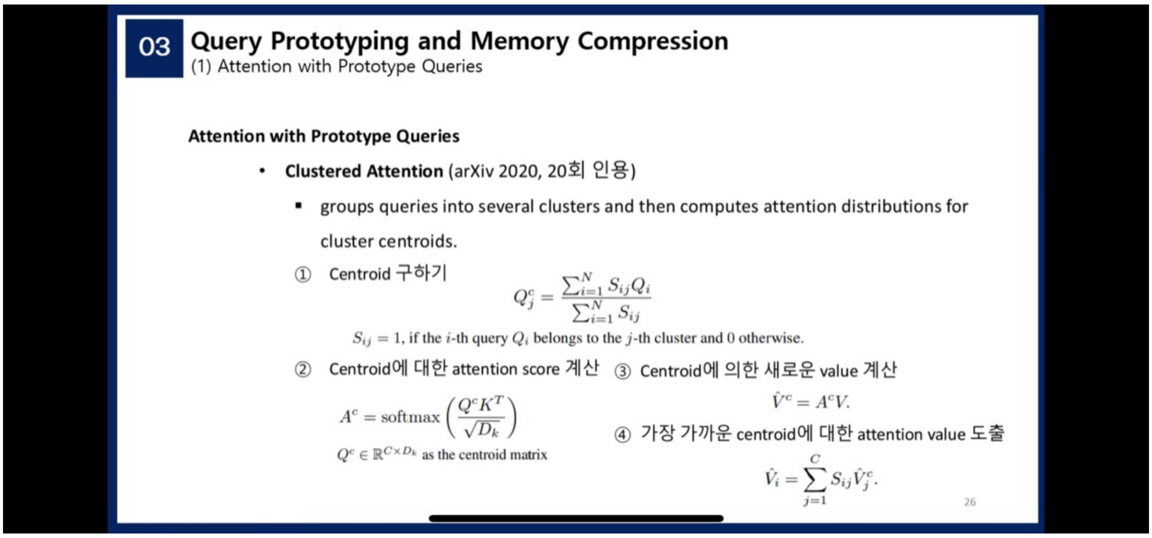
Several prototypes of queries serve as the main source to compute attention distributions

query representation 중,

- 특정 position의 query distribution을 copy
- discrete uniform distribution으로 채움

(a) Query prototyping

25



03 Query Prototyping and Memory Compression

(1) Attention with Prototype Queries

Attention with Prototype Queries

- **Clustered Attention** (arXiv 2020, 20회 인용)
  - groups queries into several clusters and then computes attention distributions for cluster centroids.

① Centroid 구하기

$$Q_j^c = \frac{\sum_{i=1}^{N} S_{ij} Q_i}{\sum_{i=1}^{N} S_{ij}}$$

$S_{ij} = 1$, if the i-th query Q_i belongs to the j-th cluster and 0 otherwise.

② Centroid에 대한 attention score 계산

$$A^c = \text{softmax}\left(\frac{Q^c K^T}{\sqrt{D_k}}\right)$$

$Q^c \in \mathbb{R}^{C \times D_k}$ as the centroid matrix

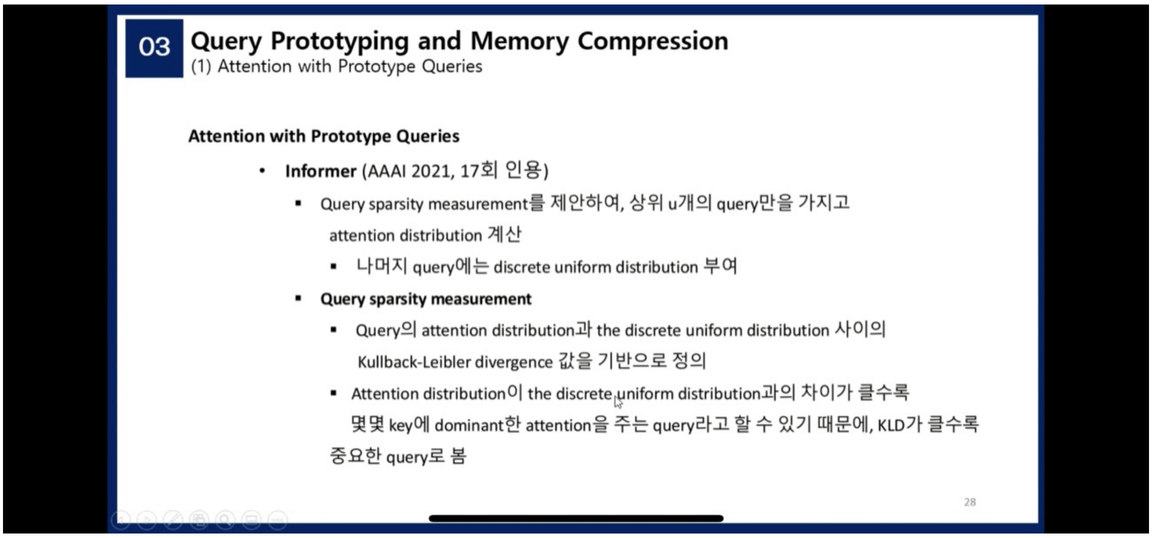
③ Centroid에 의한 새로운 value 계산

$$\hat{V}^c = A^c V.$$

④ 가장 가까운 centroid에 대한 attention value 도출

$$\hat{V}_i = \sum_{j=1}^{C} S_{ij} \hat{V}_j^c.$$

26



03 Query Prototyping and Memory Compression

(1) Attention with Prototype Queries

Attention with Prototype Queries

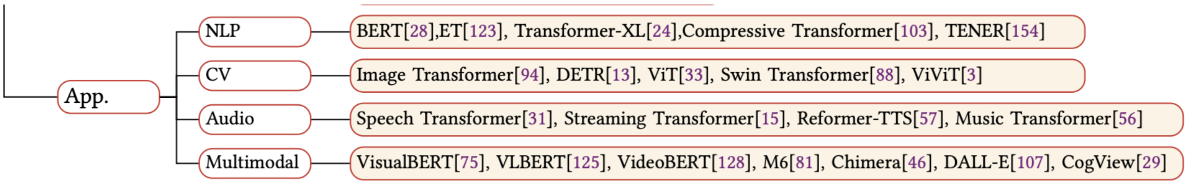
- **Informer** (AAAI 2021, 17회 인용)
  - Query sparsity measurement를 제안하여, 상위 u개의 query만을 가지고 attention distribution 계산
    - 나머지 query에는 discrete uniform distribution 부여
  - **Query sparsity measurement**
    - Query의 attention distribution과 the discrete uniform distribution 사이의 Kullback-Leibler divergence 값을 기반으로 정의
    - Attention distribution이 the discrete uniform distribution과의 차이가 클수록 몇몇 key에 dominant한 attention을 주는 query라고 할 수 있기 때문에, KLD가 클수록 중요한 query로 봄

28

Memory Compress

- MCA[84], Set Transformer[70], **Linformer**[142]

Multi-model



App.

NLP — BERT[28],ET[123], Transformer-XL[24],Compressive Transformer[103], TENER[154]

CV — Image Transformer[94], DETR[13], ViT[33], Swin Transformer[88], ViViT[3]

Audio — Speech Transformer[31], Streaming Transformer[15], Reformer-TTS[57], Music Transformer[56]

Multimodal — VisualBERT[75], VLBERT[125], VideoBERT[128], M6[81], Chimera[46], DALL-E[107], CogView[29]