1. Before Transformer

- FNN: 독립 데이터만 학습 가능
- RNN: 시퀀스 데이터 처리, hidden state 전달
  - 문제: 기울기 소실/폭발 → 장기 의존성 어려움
- LSTM: Cell state + 게이트(Forget, Input, Output) → 장기 의존성 개선
- GRU: Reset/Update 게이트 → 단순화된 구조
- Seq2Seq: Encoder-Decoder 구조, Context Vector로 압축 → 병목 현상 문제
- Seq2Seq + Attention: 전체 hidden state 반영, 단어별 중요도 학습

2. Attention

- 아이디어: 특정 시점 단어 예측 시, 관련 입력 단어에 집중
- 구성 요소:
  - Query: 현재 처리 벡터
  - Key: 유사도 측정 기준
  - Value: 가중치 적용 벡터
- Dot-Product Attention 과정:
  1. Query–Key 유사도 계산 (score)
  2. softmax로 확률화
  3. Value에 가중치 적용 → weighted sum
  4. Decoder hidden state와 결합

3. Transformer

- 입력: Embedding + Positional Encoding
- Encoder
  - Self-Attention (Q=K=V)
  - Scaled Dot-Product Attention
  - Multi-Head Attention → 다양한 패턴 학습
- Decoder
  - Masked Multi-Head Attention (미래 정보 차단)
  - Position-wise Feed-Forward (ReLU + 선형 변환)
  - Residual Connection + Layer Normalization

- 특징: RNN 없이 병렬 연산 가능

4. 대표 모델들

모델	구조	학습 방식	방향성
GPT	Decoder	Auto-Regressive	단방향
BERT	Encoder	Auto-Encoding	양방향
BART	Encoder + Decoder	Auto-Encoding + Auto-Regressive	양방향 + 단방향
ELECTRA	Encoder + Discriminator	Replaced Token Detection	양방향
T5	Encoder + Decoder	Text-to-Text	양방향

5. 학습 방식

- GPT:
  - Pre-training (다음 단어 예측)
  - Fine-tuning (태스크별 분류기 추가)
  - Zero-shot/Few-shot 학습 가능
- BERT:
  - Masked Language Model (MLM)
  - Next Sentence Prediction (NSP)
- BART:
  - 입력 데이터에 noise 추가 → denoising 학습
  - 생성·번역·이해 모두 가능