2025 여름 세미나

Signal Processing-Based Analysis of Transformer Block

Sogang University
Vision & Display Systems Lab, Dept. of Electronic Engineering

Presented By
김현빈

Contents

- Paper Review
  - Deep transformers without shortcuts: Modifying self-attention for faithful signal propagation (ICLR 2023)
  - Simplifying transformer blocks (ICLR 2024)

Paper Review

Deep transformers without shortcuts: Modifying self-attention for faithful signal propagation (ICLR 2023)

Introduction

- Introduction
  - 최근 neural network는 이론적 근거와 독립적으로 발전해오고 있었음.
    - 특히, 대부분의 현대 architecture는 특정 skip connection과 layer normalization을 배치하여 만들지만, 이 요소들의 역할이나 원칙이 완전히 이해되고 있지 않음
  - 이런 배경 속에서, skip connection 없이 깊은 신경망을 학습시킬 때, 신호 전달 원리를 적용하는 것이 관심을 받고 있음
    - Residual architecture의 효과를 설명하는 신호 전달 가설(signal propagation hypothesis)을 검증하여, DNN의 훈련 가능성(trainability)에 대한 이해를 명확히 할 수 있음
    - Residual 패러다임을 넘어 훈련 가능성을 달성할 수 있는 일반 원칙과 기술을 도출하면, 더 개선되거나 효율적인 아키텍처를 설계할 수 있는 가능성을 제공할 수 있기 때문
  - CNN에서 관련 내용들에 대한 연구는 이미 많이 진행되었으나, transformer에서는 아직 해결되지 않은 문제들이 많음.

Background

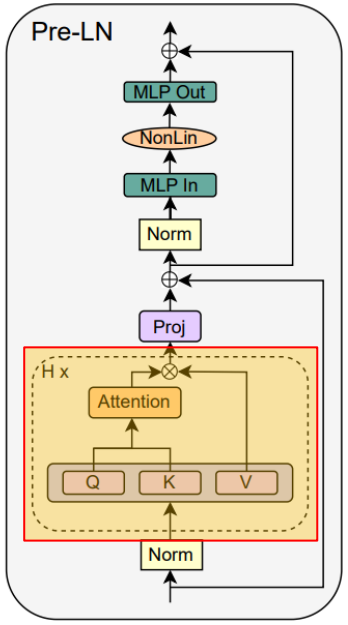
- Transformer blocks

$$\hat{\mathbf{X}}_l = \alpha \mathbf{X}_l + \beta\, \mathrm{MLP}(\mathrm{RMSNorm}(\mathbf{X}_l)),$$
$$\mathbf{X}_l = \alpha \mathbf{X}_{l-1} + \beta\, \mathrm{MHA}(\mathrm{RMSNorm}(\hat{\mathbf{X}}_{l-1}))$$
$$\mathrm{MHA}(\mathbf{X}) \triangleq \mathrm{Concat}\big(\mathrm{Attn}_1(\mathbf{X}), \ldots, \mathrm{Attn}_h(\mathbf{X})\big)\mathbf{W}^O$$

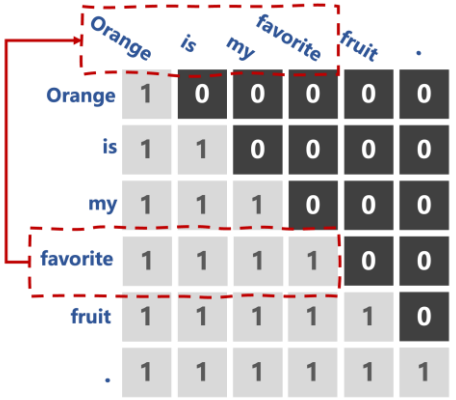
  - Self-attention

$$\mathrm{Attn}(\mathbf{X}) = \mathbf{A}(\mathbf{X})\mathbf{V}(\mathbf{X}),$$
$$\mathbf{A}(\mathbf{X}) = \mathrm{softmax}\left(\mathbf{M} \circ \frac{1}{\sqrt{d^k}}\mathbf{Q}(\mathbf{X})\mathbf{K}(\mathbf{X})^\top - \Gamma(1-\mathbf{M})\right)$$

    - Learnable parameter로 구성된 linear projection layer로 입력 x를 key, query, value 생성
    - Query, key간의 연산($K \cdot Q^T$)을 통해 얻은 가중치를 value에 적용
  - MLP layers
    - Residual branch(Normalized output of Attention block)에 적용
  - 본 논문에서는, MLP 블록이 없는 attention block에 초점
  - Causal masked attention
    - Focus to perform next-token prediction
    - 보통 여러 토큰을 병렬적으로 처리하기에 attention weight에서 이후 토큰을 mask



Attention block



Causal mask

Introduction

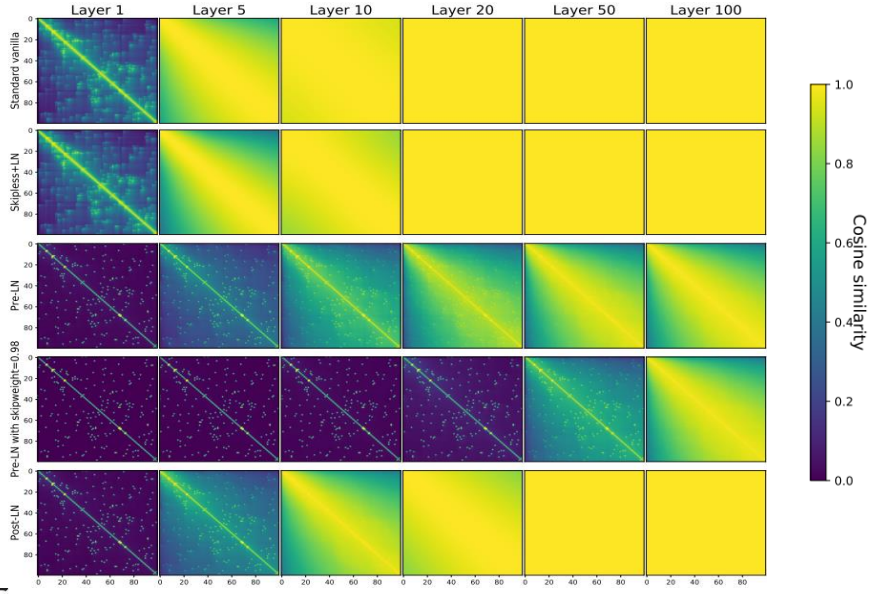
- Preliminaries
  - Location-wise kernel matrix Σ_l
    - $\Sigma_l = XX^T$, $(\Sigma_l[i,j] = X_l[i] \cdot X_l[j],\ inner\ product)$
    - Attention 레이어에서 각 위치가 다른 위치와 얼마나 연관성이 있는지를 표현
    - 신호가 어떻게 전달되는지를 수학적으로 분석할 수 있음.
  - Simplified formula for kernel matrices in deep attention-only transformers
    - $A(X) = softmax\left(\frac{1}{\sqrt{d_k}}(XW_Q)(XW_K)^T\right) = softmax\left(\frac{1}{\sqrt{d_k}}XW_QW_K^TX^T\right)\cdots XX^T$
    - $\mathrm{X}_{l+1} = \mathrm{A}_l\mathrm{X}_l\mathrm{W}_V$
    - $\Sigma_{l+1} = X_{l+1}X_{l+1}^T = (A_lX_lW_V)(A_lX_lW_V)^T \approx A_lX_lX_l^TA_l^T\ (Let\ W_V\ is\ otrhogonal\ initialized)$

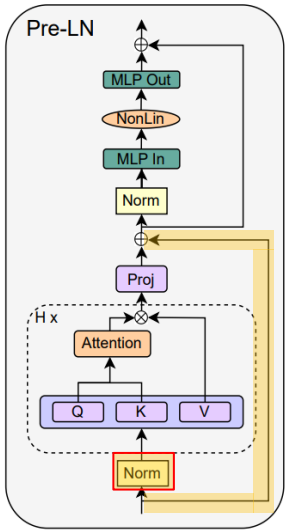
      $\therefore \Sigma_{l+1} = A_l\Sigma_lA_l^T$
    - **Kernel matrix의 변화를 통해, attention matrix가 입력 간의 관계를 어떻게 수정하고, 출력 간의 새로운 관계를 생성하는지 확인할 수 있음**

Introduction

- Rank collapse in deep skipless transformers
  - 기존 연구에 따르면, transformer를 깊게 쌓았을 때, 후반 레이어의 kernel matrix가 rank 1 matrix로 수렴한다는 문제가 있고, network의 훈련 가능성을 저해함.
    - 이 문제는 attention layer에서 발생하는 근본적인 문제이지만, 직접적인 해결 방안이 존재하지 않아 두가지 방법을 통해 우회적으로 해결함
      - Residual connection
      - Layer normalization

✓**Pre-LN:** $\boldsymbol{X_{l+1} = X_l + Residual(LN(X_l))}$, Post-LN: $X_{l+1} = LN(X_l + Residual(X_l))$





Method

- Constructing trainable deep transformers without shortcuts
  - Simplified formula for kernel matrices in deep attention-only transformers
    - 저자들은 attention matrix A에 대한 3가지 requirements를 정함

1. $\Sigma_{l+1} = A_l \Sigma_l A_l^T$ must be well-behaved at each block and avoiding rank collapse and exploding/vanishing diagonal values

- Rank collapse
  - ✓Σ_l의 rank가 1로 수렴하여, 입력 간의 관계 정보가 상실되어 학습이 불가
- Signal vanishing/explosion
  - ✓Σ_l의 대각 성분이 너무 작아지거나 커져, 입력 정보가 소실되거나 왜곡

2. A_l must be elementwise non-negatives

- $Softmax$의 계산 결과이기 때문에, 모든 원소가 음수가 아니여야함

3. ~~A_t should be lower triangular, for compatibility with causal masked attention.~~

Method

- Constructing trainable deep transformers without shortcuts
  - Identity attention: Value-skipinit
    - Attention matrix를 identity matrix ($\alpha = 1, \beta = 0$) 로 초기화하고, 점진적으로 α, β를 변화하며 학습해가자!

$$Attn(X) = A(X)V(X) \rightarrow \left(\boldsymbol{\alpha I} + \beta A(X)\right) \cdot V(X)$$

    - Identity matrix가 포함됐기 때문에, 신호의 일부가 온전히 전파될 수 있음

    - 이 방법을 통해, skip connection 없이 transformer block을 구성하여 학습시킬 수 있었지만, 엄연하게 “skipless” attention이라고 볼 수는 없음

Method

- Constructing trainable deep transformers without shortcuts
  - 아래의 두가지 series의 attention을 제안함
  - Signal preserving attention methods
    - Uniform Signal Preserving Attentions (U-SPA)

$$\Sigma_l(\rho_l) = (1-\rho_l)I_T + \rho_l 11^T, (\rho_l \leq \rho_{l+1})$$

      - 대각원소가 1이고, 나머지 원소가 모두 $\rho_l < 1$.
      - I_T 를 통해 자기 자신과의 관계를 유지하면서, 11^T 로 인해 모든 위치가 같은 정도로 영향을 주도록 설계
      - $\rho_L < 1$이기 때문에, rank collapse를 예방할 수 있음

    - Exponential Signal Preserving Attentions (E-SPA)

$$\left(\Sigma_l(\gamma_l)\right)_{i,j} = \exp(-\gamma_l|i-j|)\,.$$

      - 대각원소가 1이고, 나머지 원소는 대각원소에서 멀어지면서 지수적으로 감소.
      - **위치 간 거리를 고려**하여 자연스러운 Attention 구조를 제공이 가능함

Method

- Reverse Engineering Self-Attention Layers at Initialization
  - 저렇게 되도록 어떻게 Attention matrix A_l을 설계할건데?
    - Attention matrix의 변형

$$A(X) = softmax\left(\frac{1}{\sqrt{d_k}}Q(X)K(X)^T\right) \rightarrow softmax\left(\frac{1}{\sqrt{d_k}}Q(X)K(X)^T + B\right)$$

      - 위와 같이 attention matrix를 변형한 뒤, B(bias term)을 조절하여 U-SPA, E-SPA의 형태로 조절 가능

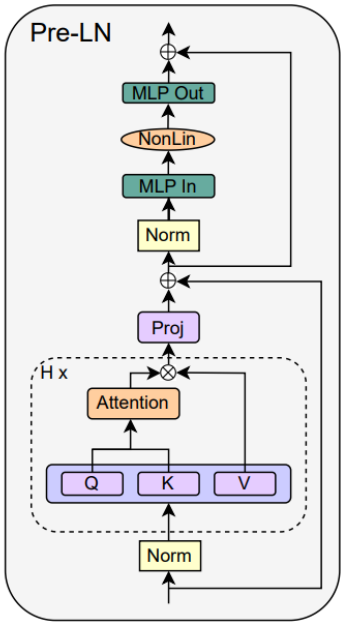
    - Query & Key initialization
      - Orthogonal Initialization
        - ✓W_Q, W_K를 Orthogonal Matrix (직교 행렬)로 초기화하여 QK^T가 랜덤한 방향으로 분포하면서 신호 소멸 방지
        - ✓Attention Matrix A(X)가 극단적인 값을 가지지 않고 적절한 형태를 유지
      - Zero- Initialization
        - ✓W_Q, W_K를 0으로 초기화하여, B만 남겨 우리가 원하는 형태로 attention 구조를 제어할 수 있음

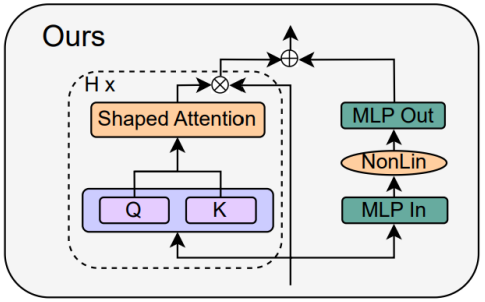
Paper Review

Simplifying transformer blocks (ICLR 2024)

Introduction

- Introduction
  - 앞선 논문은 신호 전파의 관점에서 attention block을 분석하여, skip connection 없이도 잘 작동하도록 attention block을 구성하는 방법에 대한 연구를 진행
  - 본 논문은 여기서 더 나아가서, attention block의 구성요소를 간소화하는 방법에 대한 연구를 진행
    - Removing skip connection
    - Removing value and projection layers
    - Removing MLP sub-block skip connection
    - Removing normalization layers





Method

- Removing skip connection
  - 기존 논문에서, Value-SkipInit 이라는 방법을 통해 attention matrix를 초기화하는 방법을 제안 (trainable scalars α, β initialized to 1 and 0 respectively)

$$Attn(X) = A(X)V(X) \rightarrow \left(\boldsymbol{\alpha I} + \beta A(X)\right) \cdot V(X)$$

  - 이후, 발전된 방법인 shaped attention 이 제안됐는데, 본 논문은 이를 사용.
    (α, β, γ are trainable, and C is a constant (not trained) centering matrix)

$$\boldsymbol{Attn(X)} = \boldsymbol{A(X)V(X)} \rightarrow (\boldsymbol{\alpha I} + \boldsymbol{\beta A(X)} - \boldsymbol{\gamma C}) \cdot \boldsymbol{V(X)}$$

    - $\boldsymbol{\alpha I}$: 처음에는 각 토큰이 자기 자신에게 더 집중(attend)하도록 유도
    - $\boldsymbol{\gamma C}$: Attention 값을 특정 기준으로 정렬하는 normalization 역할을 함

Method

- Removing skip connection
  - Weight Reparameterization
    - Skip connection을 제거하게 되면, 기존보다 학습속도가 느려진다는 문제점이 존재함.
    - 이를 위해, Self-Attention의 값 및 pro*jection weight* W_V, W_P를 다음과 같이 나눔.

$$\boldsymbol{W_V = \alpha_V W_V^{init} + \beta_V \Delta W_V}, \qquad \boldsymbol{W_P = \alpha_P W_P^{init} + \beta_P \Delta W_P}$$

      - W^{init} : 초기 projection layer (기본적으로 random orthogonal matrix로 설정)
      - ΔW : 학습 가능한 부분 (초기값 0), α:초기값 유지 정도, β: 학습 가능한 부분 변화량

  - Weight Reparameterization의 장점
    - W^{init}를 random orthogonal matrix로 설정하면 신호가 고르게 전파됨
    - ΔW를 초기에 0으로 설정하면, 학습 초기에 값이 급격하게 변하는 걸 방지할 수 있음
    - β를 작게 설정하면,$\left(= O\left(\frac{1}{\sqrt{L}}\right)\right)$ skip connection이 있는 것처럼 동작하여 학습 속도를 높일 수 있음

Method

- Removing value and projection layers
  - Random orthogonal matrix vs Identity Matrix initialization
    - 기존 orthogonal matrix initialization을 사용하게 되면, 신호가 고르게 분산되지만, 깊은 네트워크에서는 여전히 학습 속도가 느려질 수 있음
    - 반면, $W_V^{init} = W_P^{init} = I$로 설정하면 신호가 더 잘 보존되고 학습속도가 더 빨랐음

$$\boldsymbol{W_V = \alpha_V W_V^{init} + \beta_V \Delta W_V}, \qquad \boldsymbol{W_P = \alpha_P W_P^{init} + \beta_P \Delta W_P}$$

    - 따라서, $\boldsymbol{\beta_V = \beta_P = 0}$으로 설정하게 되면, $\boldsymbol{W_V = W_P = I}$로 유지되며, 신호가 계속 더 잘 보존된다는 것을 통해, value와 projection layer가 더 이상 필요하지 않다는 것을 알 수 있음

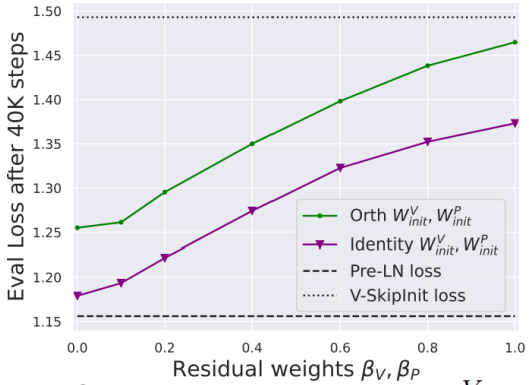


Figure 3: Restricting updates to $\mathbf{W}^V, \mathbf{W}^P$, through smaller β_V, β_P, recovers training speed in skipless transformers ($\alpha_{\text{SA}} = 0$).

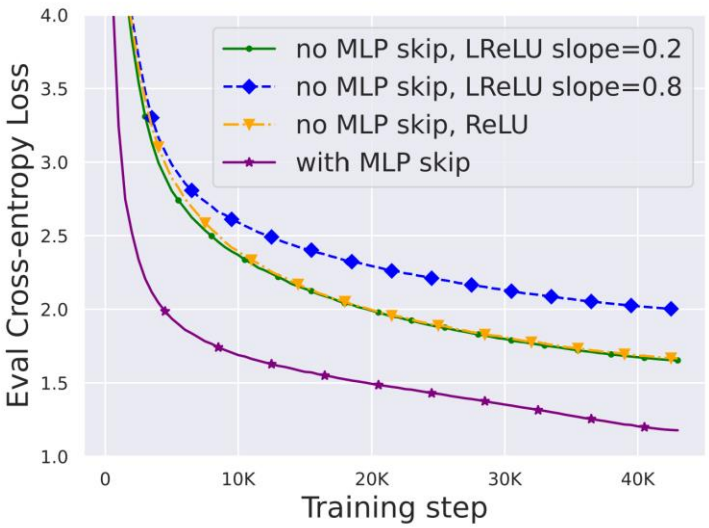
→ # params 13% 감소 및 계산량 감소
학습속도 저하 없이 기존 baseline과 동일한 성능을 유지

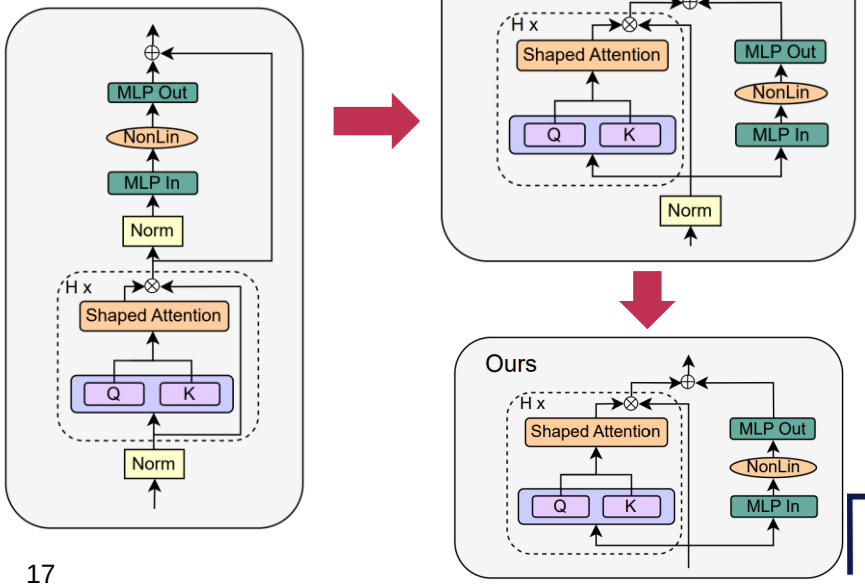
Method

- Removing MLP sub-block skip connection
  - MLP block에서는 skip connection을 제거하면, 오히려 학습속도가 저하됨.
  - 따라서, 아래와 같이 attention과 MLP 연산을 병렬적으로 처리하여, 2번의 skip connection을 한번으로 줄여 학습속도 및 latency를 감소시킴

$$\mathbf{X}_{\text{out}} = \alpha_{\text{comb}}\,\mathbf{X}_{\text{in}} + \beta_{\text{FF}}\,\text{MLP}(\text{Norm}(\mathbf{X}_{\text{in}})) + \beta_{\text{SA}}\,\text{MHA}(\text{Norm}(\mathbf{X}_{\text{in}})),$$

- Removing normalization layers
  - β_{FF}와 shaped attention을 사용함으로써, normalization의 역할을 대체 가능





Experiments

- Depth Scaling
  - Transformer의 깊이를 18->72로 확장하면서 성능 변화를 비교
  - 제안된 방법과 baseline(pre-LN) 모두, 깊어지면서 성능이 향상되며, 비슷한 학습곡선을 가지는 것을 확인
  - 반면, Value-SkipInit 모델은 모델의 깊이가 깊어질 수록 학습 속도가 저하되며 성능 격차가 벌어지는 것을 볼 수 있음

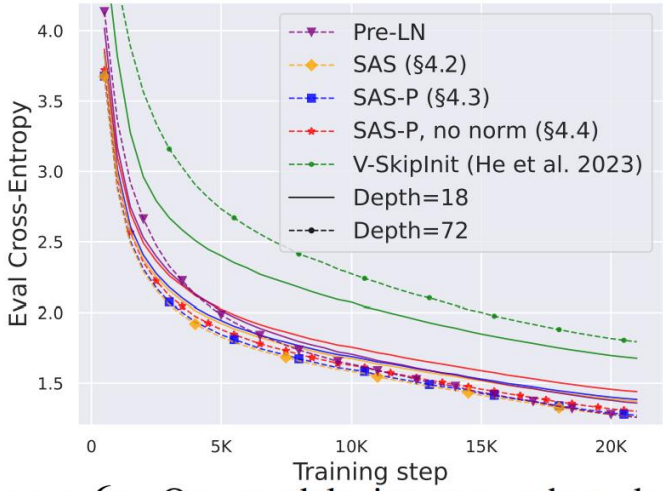
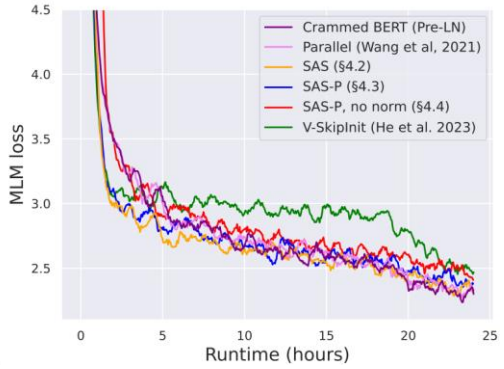


Figure 6: Our models improve when deeper (dashed, marked lines) vs. shallower (solid lines), unlike V-SkipInit (He et al., 2023).

Experiments

- BERT
  - 간소화된 Transformer 블록들이 **Autoregressive Decoder-Only 모델뿐만 아니라, 다양한 데이터셋과 아키텍처에서도 좋은 성능을 내는지** 실험
  - BERT 및 GLUE benchmark에서 성능을 평가하여 비교
    - Pretraining 속도 비교
      - 제안된 방법은, baseline BERT와 거의 동일한 학습속도를 유지하는 것을 확인
      - 스킵 연결을 제거하고 값(Value) 및 투영(Projection) 가중치를 수정하지 않은 Value-SkipInit 모델은 학습 속도가 크게 감소
    - Fine-tuning (GLUE benchmark)
      - 기존 BERT 모델 대비 16% 더 빠른 학습 속도를 가지면서 동일한 성능을 유지하는 것을 확인



Block	GLUE	Params	Speed
Pre-LN (Crammed)	$78.9_{\pm.7}$	120M	1
Parallel	$78.5_{\pm.6}$	120M	1.05
V-SkipInit	$78.0_{\pm.3}$	120M	0.95
SAS (Sec. 4.2)	$78.4_{\pm.8}$	101M	1.09
SAS-P (Sec. 4.3)	$78.3_{\pm.4}$	101M	1.16
SAS-P, no norm	-	101M	1.20

Conclusion

- Deep transformers without shortcuts: Modifying self-attention for faithful signal propagation (ICLR 2023)
  - Transformer에서 Skip Connection을 제거하면서 성능을 유지할 수 있는 방법을 제안
    - 그러나, 기본 Pre-LN Transformer보다 학습 속도가 느림
      → 학습 시간을 더 늘리면 동등한 성능을 달성할 수 있음

- Simplifying transformer blocks (ICLR 2024)
  - 기존 transformer를 대체할 수 있는 SAS 및 SAS-P 블록 제안
    - 기존 방법 대비 학습 속도 개선, 매개변수 감소, 긴 학습에서도 성능 유지
  - Value-SkipInit 방식은 깊은 네트워크에서의 확장성이 낮아 실용성이 떨어짐

감사합니다