MIX-LN: UNLEASHING THE POWER OF DEEP LAYERS BY COMBINING PRE-LN AND POST-LN

저자: Pengxiang Li, Lu Yin, Shiwei Liu

2025.05.19

발표자 : 양희원

목차

- 선정 이유
- 기존 기법
- 모델 제시
- 실험 및 결과
- 분석
- Personal Insight

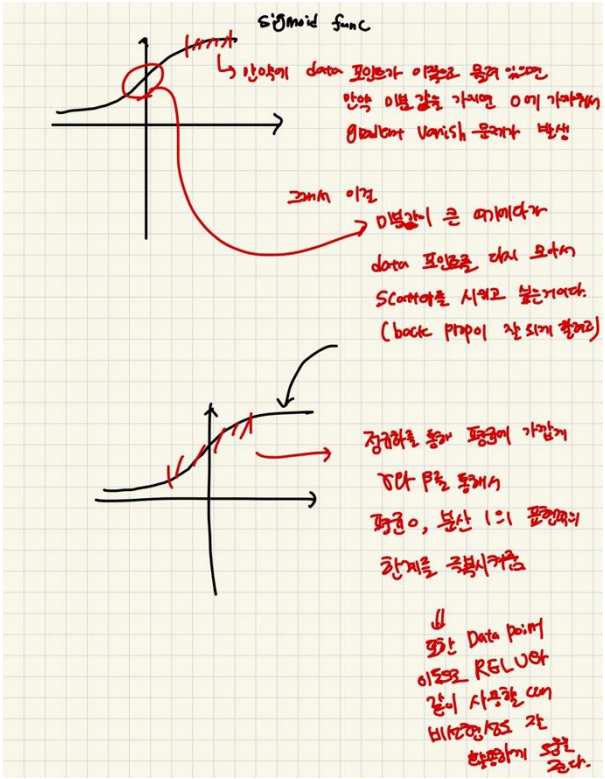
연구 배경

Transformer의 Layer Normalization

- LayerNorm은 Transformer 내부에서 gradient 안정화 및 학습 가속을 위해 필수적으로 사용됨

- 크게 두 가지 방식 존재
  - **Pre-LN** (모델 입력 직전에 정규화)
  - **Post-LN** (각 서브레이어 출력 직후 정규화)

Layer Normalziation



Layer Normalization $\mu = \frac{1}{d}\sum_{i=1}^{d} x_i$ $\sigma = \sqrt{\frac{1}{d}\sum_{i=1}^{d}(x_i - \mu)^2}$

$LN(x') = \gamma \times \frac{x'-\mu}{\sigma} + \beta$ (x'은 입력 vec, μ는 평균, γ, β는 shift 파라미터)

$$\frac{\partial \mu}{\partial x'} = \frac{1}{d} \times I$$

$$\frac{\partial \sigma}{\partial x'} = \frac{1}{2\sigma} \times \frac{\partial}{\partial x'}\left(\frac{1}{d}\sum_{i=1}^{d}(x_i' - \mu)^2\right)$$

$$= \frac{1}{2\sigma d}\sum_{i=1}^{d} 2(x_i' - \mu) \times \frac{\partial (x_i - \mu)}{\partial x'}$$

$$= \frac{1}{\sigma d}\sum_{i=1}^{d}(x_i - \mu) \times \left(e_i - \frac{1}{d}1\right) \text{ 이다.}$$

이걸 활용하여

$$\frac{\partial LN(x')}{\partial x'} = \frac{1}{\sigma} \times \frac{\partial (x'-\mu)}{\partial x'} - \frac{x'-\mu}{\sigma^2} \times \frac{\partial \sigma}{\partial x}$$

$$= \frac{1}{\sigma}\left(I - \frac{1}{d}11^T\right) - \frac{x'-\mu}{\sigma^2} \times \frac{1}{\sigma d}(x' - \mu 1)$$

$$= \frac{1}{\sigma}\left(I - \frac{1}{d}11^T - \frac{(x'-\mu 1)(x'-\mu 1)^T}{\sigma^2 d}\right) \text{ 이고}$$

여기서 가정에서 $\mu = 0$ 이라고 하면 $\|x\|_2 \simeq \sqrt{d}\,\sigma$ 이면

$$\frac{\partial LN(x')}{\partial x'} = \frac{\sqrt{d}}{\|x'\|_2} \times \left(I - \frac{x'x'^T}{\|x'\|_2^2}\right) \text{ 이다.}$$

Pre-LN

$$\text{Pre-LN}(x) = x + F(LN(x))$$ * Layer normalization을 Residual Connection 이전에 적용

$$\frac{\partial \text{Pre}}{\partial x} = I + \frac{\partial F(LN(x))}{\partial LN(x)} \times \frac{\partial LN(x)}{\partial x}$$ ⇒ x에 대해서 미분

$$= I + \frac{\partial F(LN(x))}{\partial LN(x)} \times \underbrace{\left(\frac{1}{\sigma_x} \times I\right)}_{①}$$

①

$\frac{\partial LN(x)}{\partial x}$ ⇒ Layer Normalization의 Jacobian Matrix이다.

(입력 벡터 x에 대하여 출력 벡터 LN(x)의 각 요소가 미치는 영향력)

$$\frac{\partial LN(x)}{\partial x} = \frac{\sqrt{d}}{\|x\|_2}\left(I - \frac{x x^T}{\|x\|_2}\right)$$

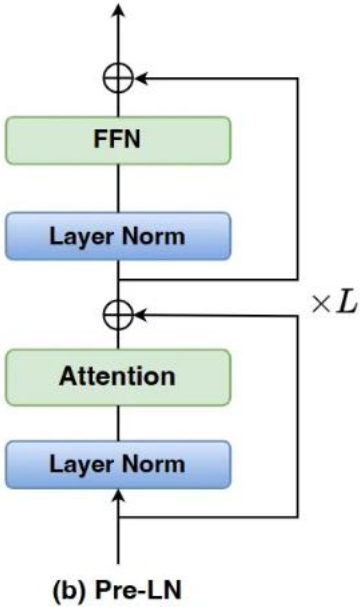
* $\sigma_x^2 = \frac{1}{d}\sum_{i=1}^{d}(x_i - \mu_{x_i})^2$ (정의) 여기서 $\mu_{x_i} = 0$ 이라고 가정

* $z = \frac{x - \mu_x}{\sigma_x}$ 라고 하자.

$$\frac{\sqrt{d}}{\sigma_x \sqrt{d}}\left(I - \frac{xx^T}{\sigma_x^2 d}\right) = \frac{1}{\sigma_x}\left(I - \frac{zz^T}{d}\right)$$

$$= \frac{1}{\sigma_x} I$$

⊕에서 x에 대한 편미분을 back prop할 때 안정적이어 초기 layer에 효과적이다. 그러나 후기 layer에서는 입력값이 그대로 출력값을 배출하기에 모델이 후기 layer에서 효과적으로 학습이 안된다!



(b) Pre-LN

Post-LN

$$\text{Post-LN}(x) = LN(x + F(x))$$

$$\frac{\partial \text{Post}}{\partial x} = \frac{\partial LN(x+F(x))}{\partial (x+f(x))} \left(I + \frac{\partial F(x)}{\partial x}\right)$$

↓

Layer Normalization의 Jacobian function이다.

↳ 1차 편미분 행렬.

$$\frac{\partial LN(x')}{\partial x'} \approx \frac{1}{\sigma_{x'}} I \quad (x' = x + f(x))$$

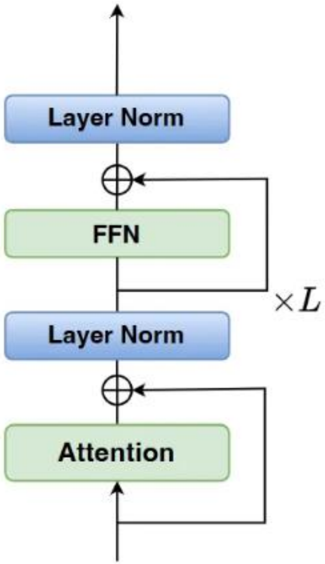
↳ 이게 layer에 여러 번 곱해지면서 <Spectral Norm> 초기 layer의 gradient가 소실되는 거임.

(위 과정에서 $x \rightarrow x'$으로 치환하면 된다)

$\Rightarrow \prod_{l=1}^{L} \frac{1}{\sigma_x^l}$ L개의 multiple layer를 지나면 gradient vanish가 된다.

또한 layer를 여러 개가 지나가다 보면 σ_x^l이 1보다 커지는 걸 관측하였다.

이를 통해 Post-LN은 초기에 layer들이 gradient vanishment에 의해서 학습되지 않음을 알 수 있다.



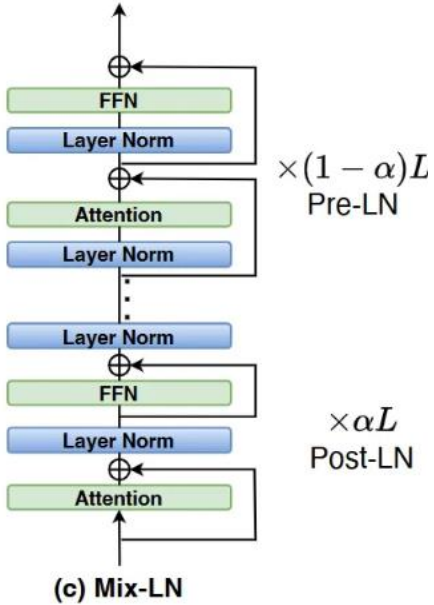
(a) Post-LN

Mix-LN

Mix-LN의 핵심 아이디어: Pre-LN과 Post-LN의 장점을 결합하여
초기 레이어에는 Post-LN을 적용하고
후반 레이어에는 Pre-LN을 적용

LLM의 총 레이어 수를 L이라고 할 때
처음 a L개의 레이어에는 Post-LN을 적용,
나머지 (1-a)L개의 레이어에는 Pre-LN을 적용

a는 0과 1 사이의 값을 가지는
hyperparameter로, 두 normalization 전략
간의 전환점을 결정



실험 method

Angular Distance: 특정 레이어의 입력과 그 다음 레이어의 입력 간의 각도 거리를 측정합니다.

$$d(x^{\ell}, x^{\ell+n}) = \frac{1}{\pi} \arccos\left(\frac{x_T^{\ell} \cdot x_T^{\ell+n}}{\left\| x_T^{\ell} \right\| \left\| x_T^{\ell+n} \right\|} \right)$$

$d(x_\ell, x_{\ell+n})$: 레이어 ℓ과 그 이후 n번째 레이어($\ell + n$) 사이의 Angular Distance를 나타냅니다.

x_ℓ^T: 레이어 ℓ에 입력되는 토큰 T를 의미합니다.

$x_{\ell+n}^T$: 레이어 $\ell + n$에 입력되는 토큰 T를 의미합니다.

$x_\ell^T \cdot x_{\ell+n}$: x_ℓ^T와 $x_{\ell+n}^T$의 내적(dot product)입니다.

$\|x_\ell^T\|$: 벡터 x_ℓ^T의 L2-norm(유클리드 노름)을 나타냅니다. 벡터의 크기를 측정하는 방법입니다.

$\arccos$: 역코사인 함수입니다. 내적값을 두 벡터 크기의 곱으로 나눈 값의 역코사인을 취하여 두 벡터 사이의 각도를 라디안 단위로 계산합니다.

$\frac{1}{\pi}$: 결과값을 [0, 1] 범위로 조정하기 위한 스케일링 요소입니다. arccos의 결과값은 [0, π] 사이의 값을 가지므로, π로 나누어 0과 1사이의 값으로 정규화합니다.

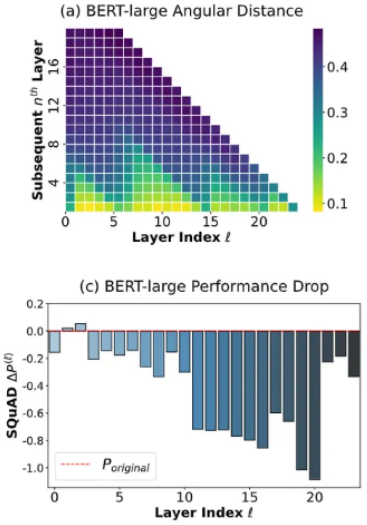
Angular Distance의 의미: Angular Distance는 두 레이어 간의 표현(representation) 유사성을 측정합니다. 값이 작을수록(0에 가까울수록) 두 레이어의 표현이 유사함을 의미합니다. 즉, 두 벡터가 가리키는 방향이 비슷합니다. 또힌 값이 클수록(1에 가까울수록) 두 레이어의 표현이 다름을 의미합니다. 즉, 두 벡터가 가리키는 방향이 상이합니다.

Performance Drop: LLM에서 특정 레이어 l 을 제거(pruning)하기 전과 후의 성능 차이

$$\Delta P(\ell) = P(\ell)_{\text{pruned}} - P_{\text{original}}$$

변화량이 적을 수록 해당 레이어를 제거해도 모델 출력에 미치는 영향이 줄어듬을 알려줍니다.

Post-Layer Normalization 결과(BERT)



기존 Post 모델을 활용하여 실험

BERT-Large는 Post-Layer Normalization (Post-LN)을 사용하는 모델입니다.

1번째 지표에서 노란색은 거리가 짧음(유사도가 높음)을, 보라색은 거리가 큼(유사도가 낮음)을 의미합니다.

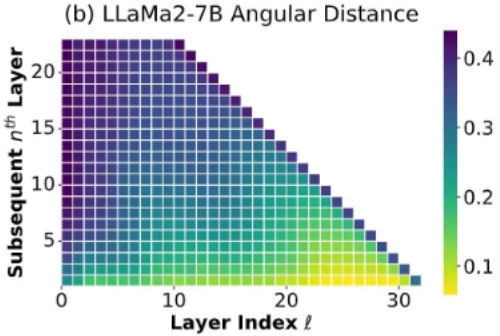
BERT-large의 앞부분 레이어들이 뒷부분 레이어보다 인접 레이어와 각도 거리가 짧은 경향이 있습니다. 특히, 3, 4, 9, 10, 11번째 레이어는 인접 레이어와 매우 유사합니다.

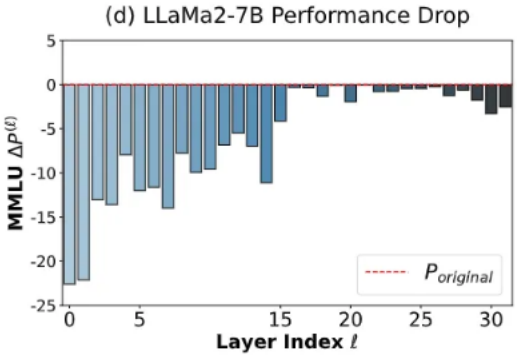
x축은 제거된 레이어의 인덱스, y축은 SQuAD 데이터셋에서의 성능 저하를 나타냅니다.

초기 레이어를 제거했을 때의 성능 저하가 더 깊은 레이어를 제거했을 때보다 훨씬 작습니다. 흥미롭게도 2, 3번째 레이어를 제거하면 성능이 약간 향상됩니다.

결론: BERT-Large 모델에서는 앞부분 레이어가 뒷부분 레이어보다 효과가 떨어지고 이는 앞부분 레이어가 뒷부분 레이어보다 효과가 떨어집니다.

Pre-Layer Normalization 결과(LLaMa2)





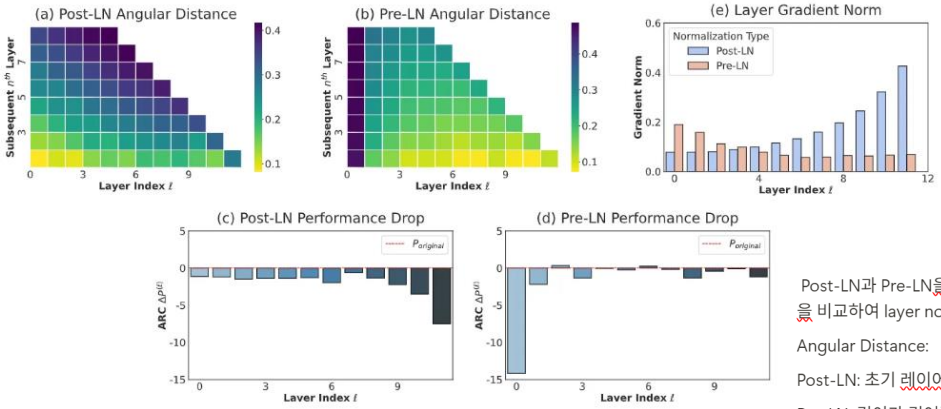
LLaMa2-7B는 Pre-Layer Normalization을 사용하는 모델입니다.

레이어가 깊어질수록(오른쪽으로 갈수록) 인접 레이어와의 각도 거리가 점차 감소합니다(보라색에서 노란색으로 변함). 20번째에서 30번째 레이어는 인접 레이어와 각도 거리가 매우 짧습니다.

더 깊은 레이어를 제거해도 정확도 손실이 거의 없지만, 초기 레이어를 제거하면 정확도가 크게 떨어집니다.

기존 Pre모델을 활용하여 실험

같은 모델을 통해 Pre, Post 비교(자체적으로 학습시킨 소규모 LLM)



Pre와 Post 모델을 동일한 거 활용하여 실험

Post-LN과 Pre-LN을 사용한 LLaMa-130M 모델의 Angular Distance, Performance Drop, Gradient Norm을 비교하여 layer normalization 방식에 따른 차이점입니다.

Angular Distance:

Post-LN: 초기 레이어에서 유사도가 높고, 깊이가 깊어질수록 레이어 간 구별이 뚜렷해집니다.

Pre-LN: 깊이가 깊어질수록 레이어 간 유사도가 점진적으로 감소하여 깊은 레이어에서 유사도가 매우 높게 나타납니다.

Gradient Norm:

Post-LN: 깊은 레이어에서 기울기가 크지만 초기 레이어에서 기울기 소실이 심각합니다.

Pre-LN: 초기 레이어에서 기울기 흐름이 좋지만 후반 레이어에서는 감소합니다.

Gradient Norm은 각 레이어의 기울기 크기를 나타내며, 이는 학습 과정에서 레이어가 얼마나 효과적으로 기여하는지를 보여줍니다. Post-LN은 깊은 레이어에서 큰 기울기를 가지지만 초기 레이어에서는 기울기 소실이 발생하며, Pre-LN은 초기 레이어에서는 기울기가 잘 유지되지만 깊은 레이어에서는 기울기가 감소합니다. 이는 각 normalization 방식이 신경망의 학습에 미치는 영향을 분석하는 데 중요한 지표로 활용됩니다.

실험 결과(Mix-LN의 우수성)

Table 1: Perplexity (↓) comparison of various normalization methods across various LLaMA sizes.

Training Tokens	LLaMA-71M 1.1B	LLaMA-130M 2.2B	LLaMA-250M 3.9B	LLaMA-1B 5B
Post-LN	35.18	26.95	1409.09	1411.54
DeepNorm	34.87	27.17	22.77	1410.94
Pre-LN	34.77	26.78	21.92	18.65
Mix-LN	**33.12**	**26.07**	**21.39**	**18.18**

Mix-LN은 모든 모델 크기에서 가장 낮은 perplexity를 달성하여 Pre-LN과 Post-LN 모두를 능가합니다. 특히 LLaMA-71M 및 LLaMA-250M에서 Pre-LN에 비해 상당한 성능 향상을 보입니다.

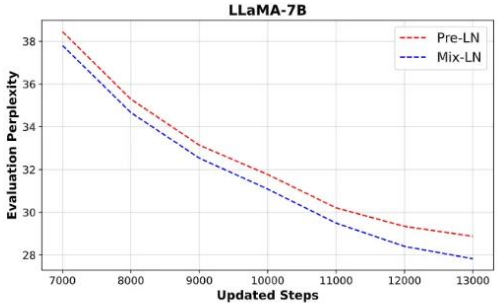


Figure 4: Training curve (eval perplexity) of `Mix-LN` and Pre-LN with LLaMa-7B.

그래프는 Mix-LN 방법이 Pre-LN 방법에 비해 LLaMA-7B 모델의 훈련 과정에서 더 낮은 Perplexity를 달성하여 더 우수한 성능을 나타냄을 보여줍니다. 이 논문에서 제시하는 Mix-LN의 효과를 뒷받침하며, 특히 대규모 언어 모델(LLM)의 깊은 레이어를 효율적으로 활용하는 데 Mix-LN이 기여할 수 있음을 보여줍니다. Mix-LN은 Pre-LN과 Post-LN의 장점을 결합하여 더 균형 잡힌 학습을 가능하게 하고, 이는 모델의 전반적인 성능 향상으로 이어집니다

Mix-LN의 우수성을 보여주는 다양한 지표

Table 3: RLHF comparison of final reward (↑) of Pre-LN and Mix-LN with LLaMA-1B.

Method	Model	Final Reward
Pre-LN	LLaMA-1B	0.75
Mix-LN	LLaMA-1B	**1.32**

RLHF(Reinforcement Learning from Human Feedback, 인간 피드백 기반 강화 학습) 보상 결과는 언어 모델이 인간의 선호도에 맞춰 얼마나 잘 작동하는지를 나타내는 지표입니다. Mix-LN은 LLM의 깊은 레이어의 잠재력을 최대한 활용하여 모델 크기를 늘리지 않고도 모델의 용량을 향상시킵니다.

Table 4: Accuracy (↑) comparison of Pre-LN and Mix-LN on ViT models.

Model	ViT-Tiny	ViT-Small
Pre-LN	67.30	75.99
Mix-LN	**67.34**	**76.40**

Mix-LN이 Vision Transformer 모델의 성능을 향상시키는 데 효과적인 normalization 기술이고 vision분야에서도 활용가능성이 높습니다.

α에 따른 학습 능력 및 각 레이어의 표현력

Table 5: Perplexity of LLaMA-1B with various Post-LN ratios α.

	Pre-LN	Mix-LN					Post-LN
Post-LN ratios α	0	16.7%	25.0%	33.0%	41.7%	50.0%	100%
Perplexity	18.65	18.34	**18.18**	18.41	18.55	18.86	1434

α는 hyperparameter로써 α=0 일때는 Pre-LN α=100이면 post-LN의 특성을 띄게 만들었다.

Mix-LN은 Pre-LN과 Post-LN의 장점을 결합하여 더 나은 성능을 달성하며, 특정 비율(α = 25.0%)에서 최적의 성능을 보입니다. Post-LN은 학습 불안정성으로 인해 성능이 좋지 않습니다.

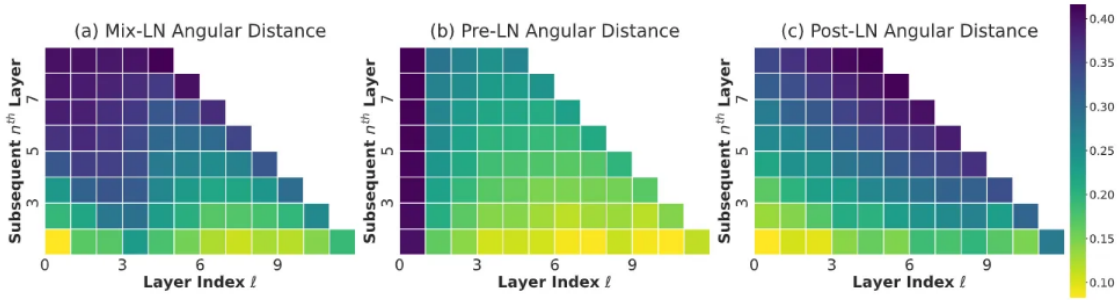
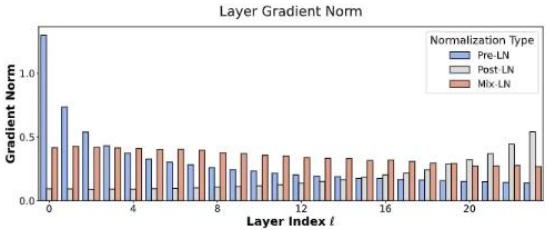


Figure 5: Angular distance from initial layer ℓ (x-axis) with block size n (y-axis) of LLaMA-130M.

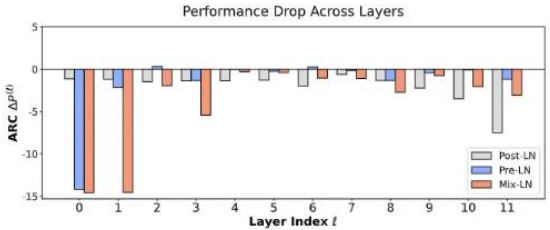
Mix-LN이 Pre-LN 및 Post-LN의 단점을 보완하고, 레이어 간 표현의 다양성을 높이고 Pre-LN이나 Post-LN에 비해 더 균등한 학습을 가능하게 합니다.

학습 목표

Mix-LN이 Pre-LN 및 Post-LN의 단점을 보완하고, 레이어 간 표현의 다양성을 높이고 Pre-LN이나 Post-LN에 비해 더 균등한 학습을 가능하게 합니다.



(a) Layer gradient norm of LLaMA-250M with various normalization techniques.



(b) Performance drop comparison of LLaMA-130M across layers for Pre-LN, Post-LN, and Mix-LN.

Mix-LN은 Pre-LN과 Post-LN의 단점을 보완하여 레이어 전체에 걸쳐 안정적인 Gradient Norm을 유지하고, 초기 레이어에서도 성능을 유지하면서 깊은 레이어의 기여도를 높이는 효과를 보입니다.

Table 6: Comparison against other normalization methods on LLaMA-250M.

Model	Pre-LN	Admin	Group-LN	Sandwich-LN	Mix-LN
LLaMA-250M	23.39	24.82	23.10	23.26	**22.33**

Mix-LN이 다른 Normalization 기법들보다 LLaMA-250M 모델의 성능을 향상시킨다는 것을 의미합니다. 또한 Mix-LN을 통해 학습을 안정화시키고, 모델의 표현력을 높이기 때문으로 해석할 수 있습니다.

Personal Insight

- LLM의 gradient vanishing 문제는 CNN만큼 심각하지 않음
  - LLM은 깊이보다는 width와 데이터 양으로 성능 확보
  - CNN은 100+ 레이어, LLM은 48개 이하 Transformer 블록 사용

- Transformer는 전역 정보 병렬 처리 → 얕은 구조로도 표현력 확보

- 논문 내 α 하이퍼파라미터 → 학습 가능한 파라미터로 설정 가능성 제안
  - 모델이 스스로 최적 전환점 학습 → adaptive design 확장

- 깊은 레이어는 문맥 통합 및 요약 생성에 핵심적 역할을 하지 않을까.. (e.g. text summarization)