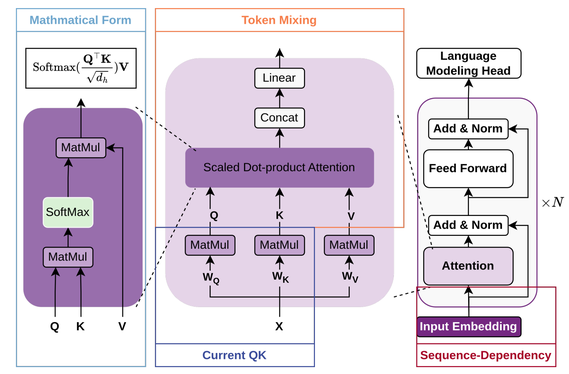
Deconstructing Attention: Investigating Design Principles for Effective Language Modeling

Huiyin Xue, Nafise Sadat Moosavi, Nikolaos Aletras[1]

1 School of Computer Science, University of Sheffield, United Kingdom

INTUITION

- **Attention Deconstruction** Are the following four key design principles truly essential, or could relaxing some of them suffice if applied selectively?
  1. `Token Mixing`: enabling multi-token interactions
  2. `Mathematical Form`: dot-product similarities plus softmax weighting
  3. `Sequence-Dependency`: adapting attention weights to each input
  4. `Current QK`: grounding attention in the current layer
- **Diagnostic Approach** Guided by Occam's Razor, we systematically relax these principles through controlled attention variants, evaluated in two settings: (1) *uniform* replacement across all layers, and (2) *hybrid* configurations that interleave standard and simplified modules.
- **Key Findings**
  - Token mixing and following the mathematical form are crucial for attention alternatives when applied uniformly, but not necessary for hybrid.
  - Sequence-dependency enhances the generalization ability.
  - Current QK is not as essential as expected.
  - Layer collaboration matters.
  - Standard layers in the hybrid architecture might serve as a normalization mechanism.
  - Strategically integrating a few standard attention layers within LMs can greatly improve, even overcome, limitations of less powerful attention mechanisms.



ATTENTION VARIANTS

STANDARD DOT-PRODUCT ATTENTION

$$\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i = \mathbf{X}\mathbf{W}_i^{Q,K,V}$$
$$\mathbf{H}_i = \mathrm{Att}(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i) = \mathrm{SoftMax}(\frac{\mathbf{Q}_i\mathbf{K}_i^\top}{\sqrt{d_h}})\mathbf{V}_i$$

RELAXING TOKEN MIXING

- MLP

$$\mathbf{O} = \mathrm{GatedMLP}(\mathbf{H}) = \mathrm{FC}_{\mathrm{Dn}}(\mathrm{SiLU}(\mathrm{FC}_{\mathrm{Gt}}(\mathbf{H})) \cdot \mathrm{FC}_{\mathrm{Up}}(\mathbf{H}))$$

RELAXING THE DERIVATION OF QK

- StaticEmbQK

$$\mathbf{Q}, \mathbf{K} = \mathbf{e}\mathbf{W}^{Q,K}; \quad \mathbf{V} = \mathbf{H}\mathbf{W}^{V}; \quad \mathbf{e} = \mathrm{Emb}(\mathbf{t})$$

RELAXING THE MATHEMATICAL FORM

- Approximate

$$\mathbf{A} \approx \mathrm{Taylor}\left(\mathbf{Q}\mathbf{K}^\top/\sqrt{d_h}\right)$$

- Non-approximate

$$\mathbf{A} = \mathrm{Softmax}\left((\mathbf{Q} \odot \mathbf{K})\mathbf{1}^\top/\sqrt{d_h}\right)$$
$$\mathbf{Q} = \mathrm{SiLU}\left(\mathbf{H}\mathbf{W}^{Q}\right); \quad \mathbf{K}, \mathbf{V} = \mathbf{H}\mathbf{W}^{K,V}$$

RELAXING SEQUENCE DEPENDENCY

- Random-fixed (RndEmbQK)

$$\mathbf{X} = \mathrm{TransformerBlock}^{(l)}(\epsilon), \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \sigma\mathbf{I})$$
$$\mathbf{Q}, \mathbf{K} = \mathbf{X}\mathbf{W}^{Q,K}; \quad \mathbf{V} = \mathbf{H}\mathbf{W}^{V}$$

- Text-fixed (FixedSeqQK)

$$\mathbf{X} = \mathrm{TransformerBlock}^{(l)}(\mathrm{Emb}(\mathbf{t}^s))$$
$$\mathbf{Q}, \mathbf{K} = \mathbf{X}\mathbf{W}^{Q,K}; \quad \mathbf{V} = \mathbf{H}\mathbf{W}^{V}$$

THEORETICAL EFFICIENCY

- FLOP/it for each attention across all variants

Attention	Prefill	Decode
Standard	$4BL^2d + 6BLd^2$	$6Bd^2 + 4BLd$
MLP	$6BLd^2$	$6Bd^2$
Approx.	$14BLd^2$	$10Bd^2$
Non-apx.	$6BLd^2$	$6Bd^2$
RndEmbQK	$2L^2d + 2BL^2d + 6BLd^2$	$2Ld + 2BLd + 6Bd^2$
FixedSeqQK	$2L^2d + 2BL^2d + 6BLd^2$	$2Ld + 2BLd + 6Bd^2$
StaticEmbQK	$4BL^2d + 6BLd^2$	$6Bd^2 + 4BLd$

- Activation memory for each attention across all variants

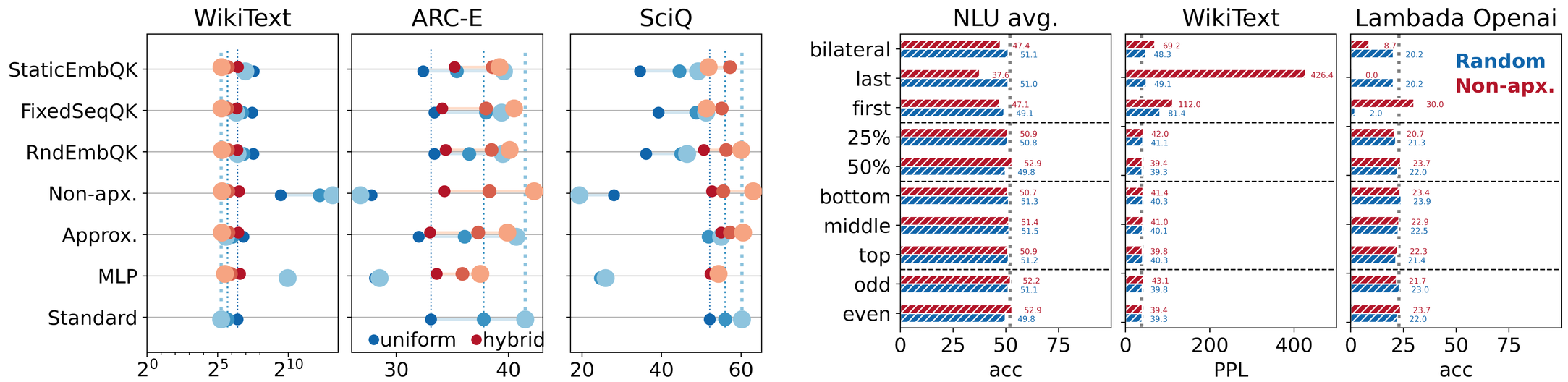
Attention	Activation memory
Standard	$8BLd + 2BL^2h/t$
MLP	$8BLd/t$
Approx.	$11BLd/t + 3Bd^2/ht$
Non-apx.	$8BLd + 4BLh/t$
RndEmbQK	$4BLd + 8Ld + 2L^2h/t$
FixedSeqQK	$4BLd + 8Ld + 2L^2h/t$
StaticEmbQK	$8BLd + 2BL^2h/t$

- Cache size (in bytes) per layer across all attention variants required during inference

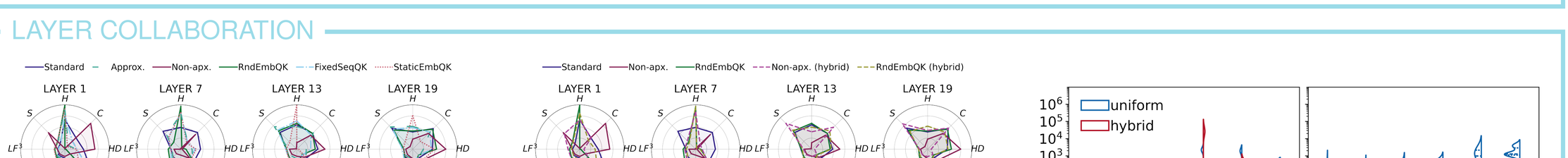
Attention	Cache Size for Inference
Standard	$4BLd$
MLP	0
Approx.	$6Bd + 4Bd^2/h$
Non-apx.	$2Bd + 4Bh$
RndEmbQK	$2(B+1)Ld$
FixedSeqQK	$2(B+1)Ld$
StaticEmbQK	$4BLd$

EMPIRICAL RESULTS

MODEL PERFORMANCE



LAYER COLLABORATION



e-mail: {hxue12, n.s.moosavi, n.aletras}@sheffield.ac.uk

Presented at the International Joint Conference on Natural Language Processing Asia-Pacific Chapter of the Association for Computational Linguistics, 2025 in Mumbai, India on 20-24 Dec 2023

IJCNLP-AACL 2025