Oregon State University
COLLEGE OF ENGINEERING | School of Electrical Engineering and Computer Science

LeaPformer: Enabling Linear Transformers for Autoregressive and Simultaneous Tasks via Learned Proportions

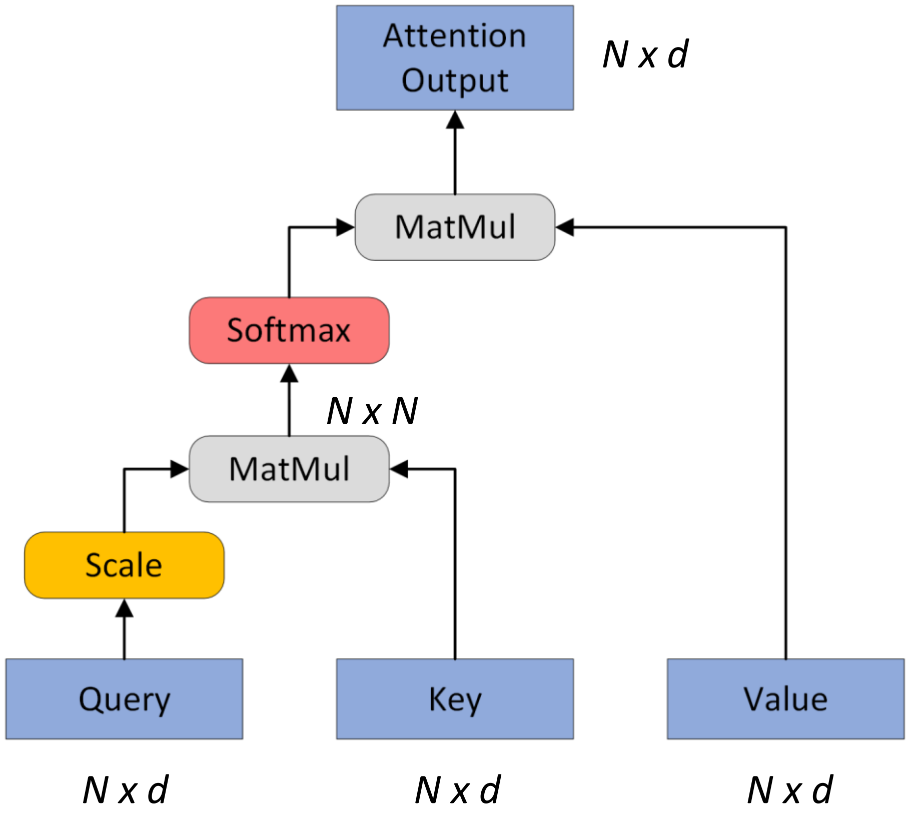
Victor Agostinelli, Sanghyun Hong, Lizhong Chen

ICML
International Conference On Machine Learning

Motivation

- Attention in transformers is bottlenecked in run-time and memory footprint when engaging with large sequences, but is still **de-facto for sequence modeling**!

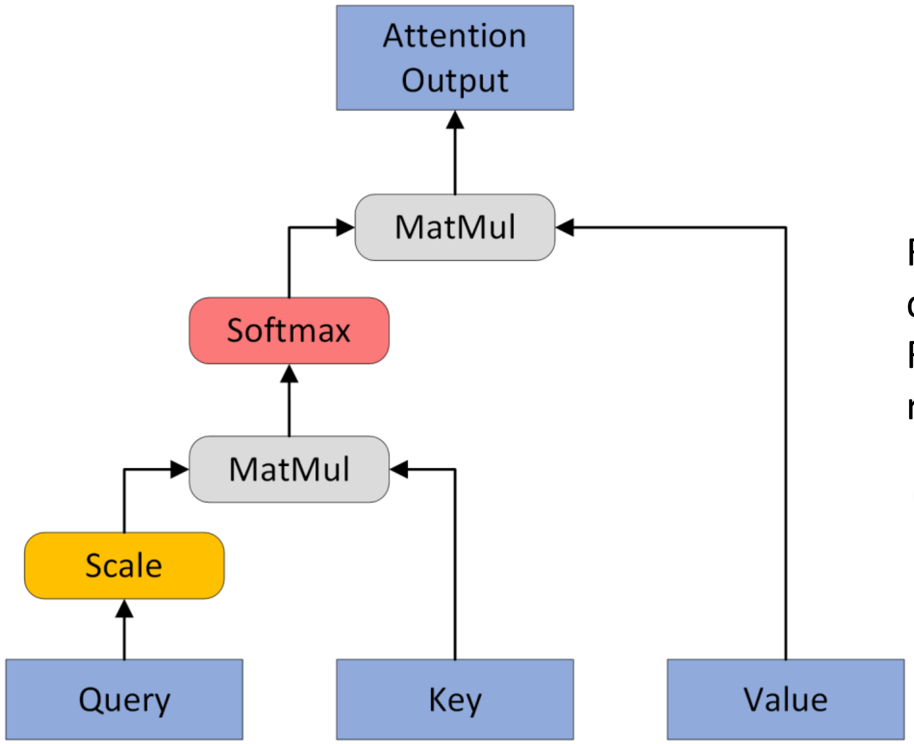
- Efficient attention variants have been studied almost since the introduction of transformers, but **remain niche** or result in **severe degradation** to model accuracy.



Linear Attention

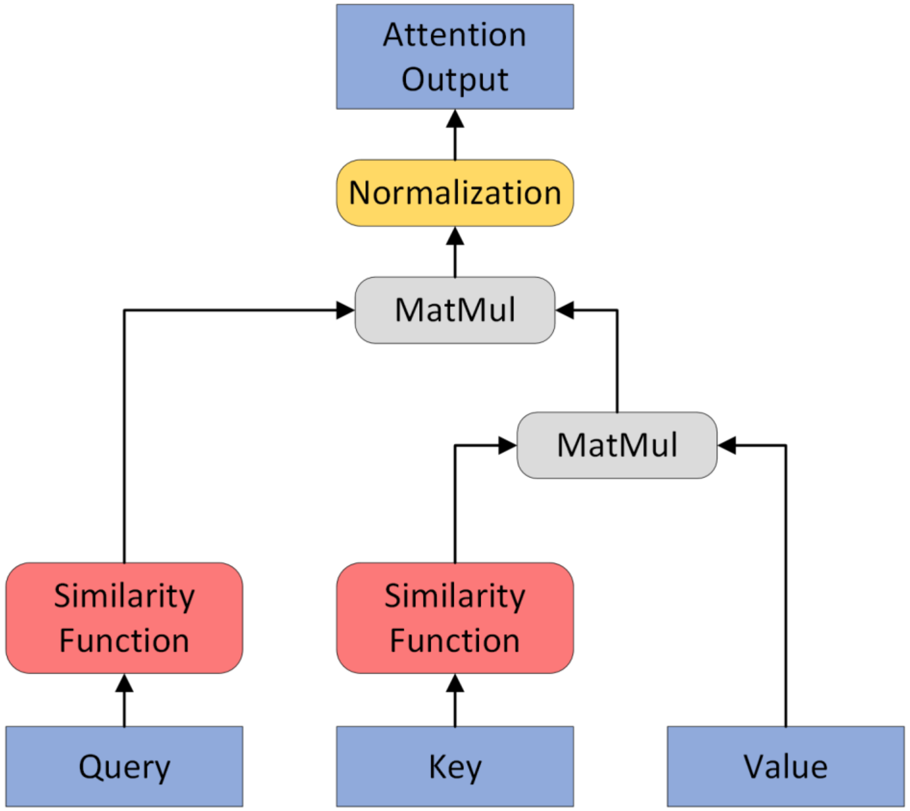
- Truly linear with no prerequisites? *O(n)* run-time and memory footprint!
  - Results in RNN-like recurrent behavior during inference!
  - Recurrent state is memory constraint during attention, much smaller than QK^T matrix! No need to cache key and value matrices!
  - Includes benefits like infinite LLM context, increased accessibility for edge devices, **extreme speedups**, etc.

Linear Attention Preliminaries



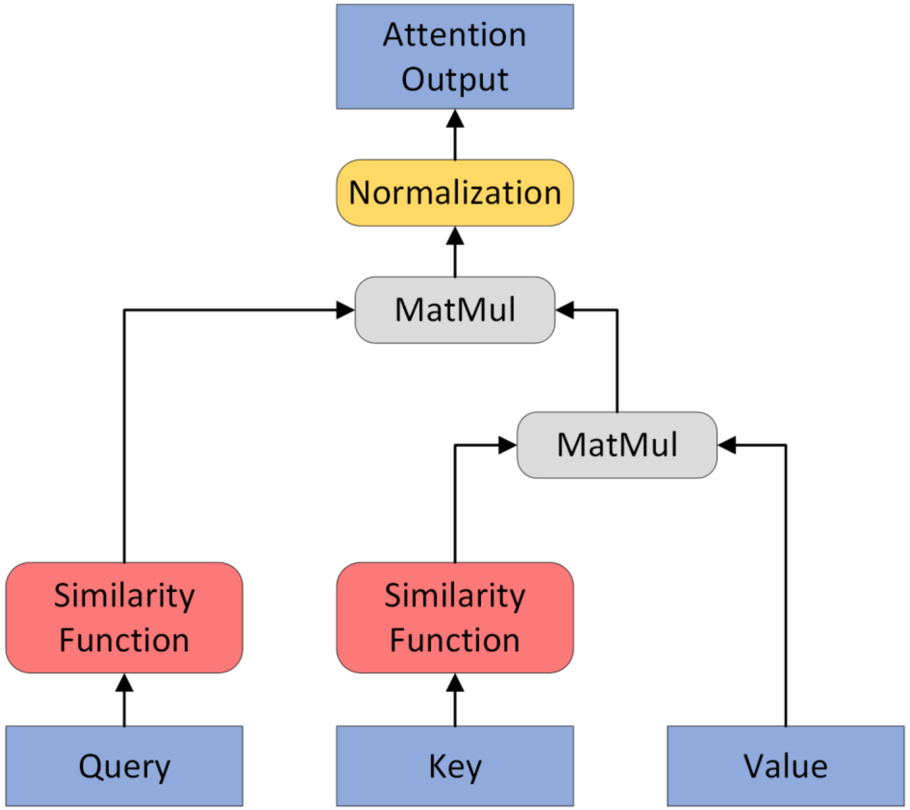
Softmax Attention: $O(n^2)$ Run-time and Mem.

Replace Softmax with a decomposable Similarity Function *S(…)* and reorganize computation.

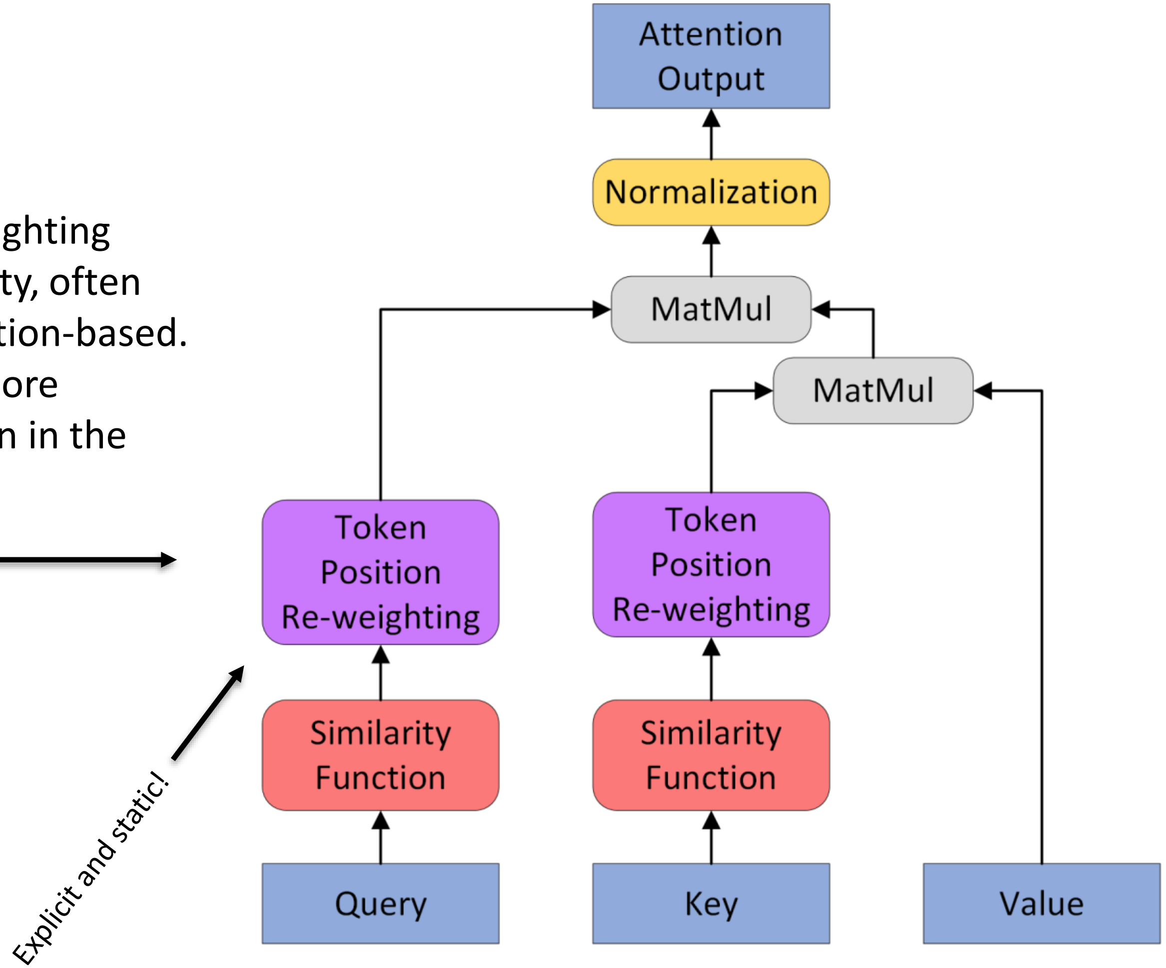


Linear Attention: $O(n)$ Run-time and Mem.

Re-weighting Functions

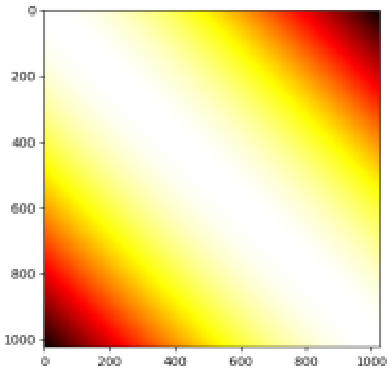


Add re-weighting functionality, often token position-based. Encodes more information in the sequence.

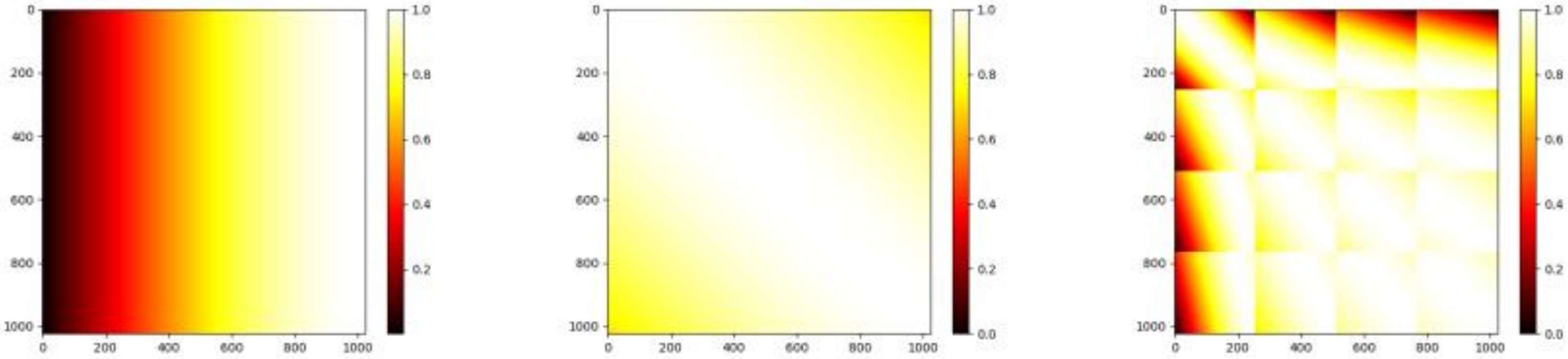


Limits of Explicit Positional Re-weighting

- Why are explicit positional re-weighting functions (SOTA is cosFormer) limited? Sequence length is needed, so autoregressive tasks become very difficult! Additionally, most simultaneous (i.e. streaming) tasks are **impossible**.
  - Autoregressive Language Modeling
  - Machine Translation (usually autoregressive) and Simultaneous Translation (T2T, S2T, etc.)



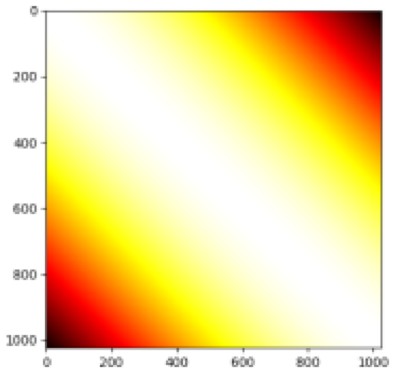
Static Attention
Concentration Pattern for
Bi-directional Attention



Problematic Attention Concentration
Pattern Alternatives for
Causal Attention

Limits of Static Attention Concentrations

- Static patterns are not very generalizable across tasks! For example:
  - Cross-attention concentrations between English-to-German vs. English-to-Chinese
  - Bidirectional language modeling vs. hierarchical math problem solving



Static Attention Concentration Pattern

Proposed Approach: LeaPformers

- Replace explicit positional re-weighting functions in sequences with sequence proportions. **No theoretical dependence** on explicit token positions!

- Replace static attention concentrations, with a **learnable component** based on proportions!

$$P_q = [P_{q,1},\ P_{q,2},\ \dots\ P_{q,N_1}], \quad 0 \leq\ P_{q,i} \leq 1$$

$$P_k = [P_{k,1},\ P_{k,2},\ \dots\ P_{k,N_2}], \quad 0 \leq P_{k,j} \leq 1$$

$$S(Q_{h,i}, K^T_{h,j}) = S_q(Q_{h,i})S_k(K^T_{h,j})\sigma(P_{q,i}, P_{k,j})$$

$$\downarrow$$

$$P_q(Q_{h,i}) = P_{q,i} = LeaP_Q(Q_{h,i})$$

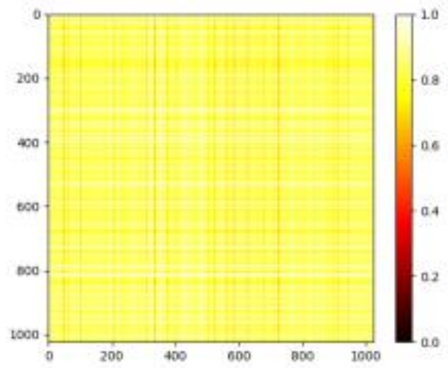
$$P_k(K_{h,j}) = P_{k,j} = LeaP_K(K_{h,j})$$

$$P_q(Q_h) = [LeaP_Q(Q_{h,1}),\ \dots\ , LeaP_Q(Q_{h,N_1})]$$

$$P_k(K_h) = [LeaP_K(K_{h,1}),\ \dots\ , LeaP_K(K_{h,N_2})]$$

Impact of Proposed Approach

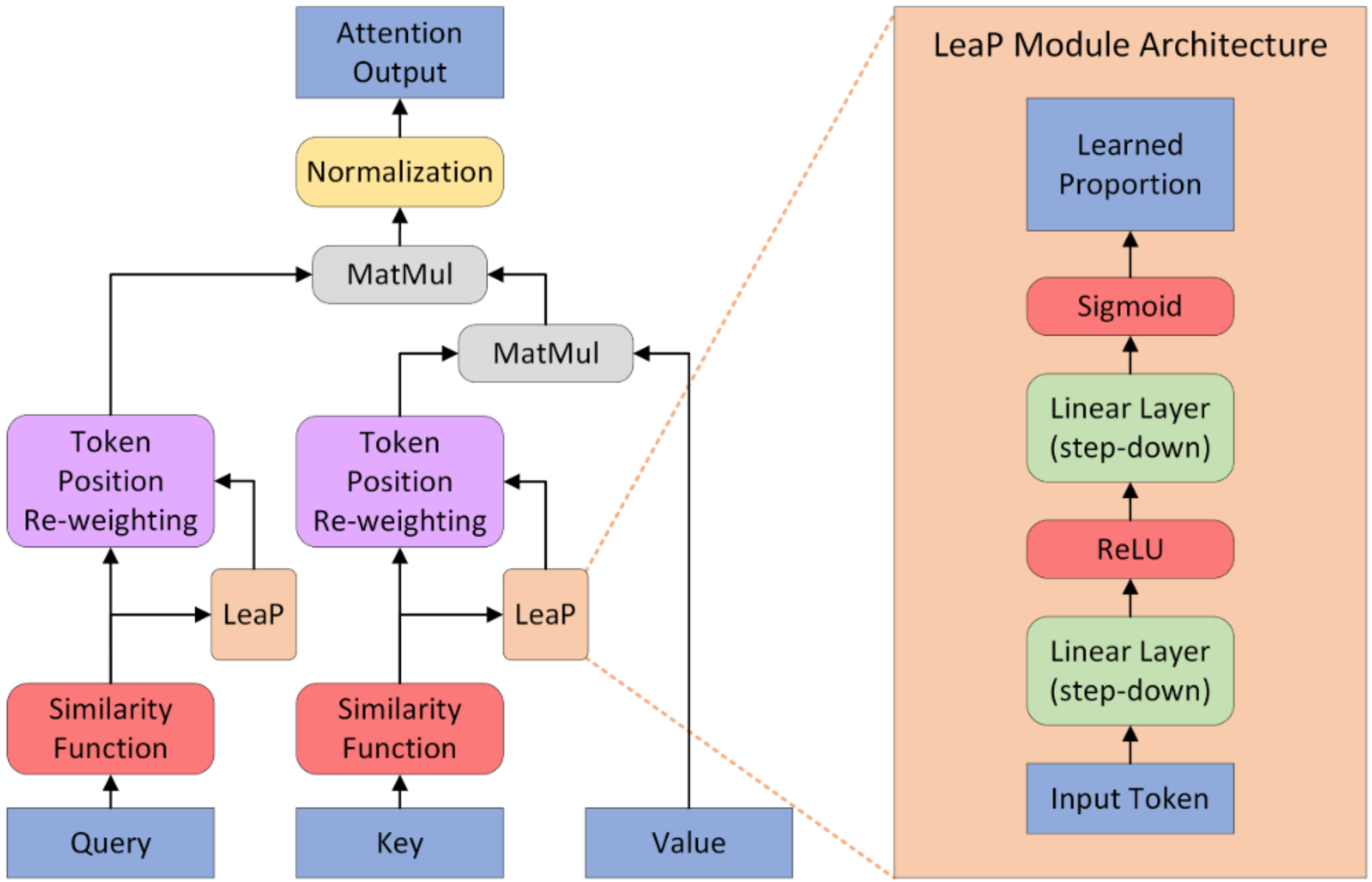
- **LeaPformers** solve the aforementioned issues!
  - Proportion representations mean **no theoretical blockade** to autoregressive and simultaneous applications.
  - Combined with the desire to eliminate static representations, a **learnable representation** allows for dynamism in attention concentrations!
  - Architectural changes are minimal to maintain linear attention throughput in addition to improving accuracy.



Dynamic Attention Concentration Pattern

LeaPformer Architecture and Results

- Small accuracy loss, but up to **7.8x faster** than softmax with **7.5x less memory!**
- Roughly equal in accuracy to BigBird, but **3.3x faster** and **3.1x less memory**!
- Improves on cosFormer! **More accurate** on all tasks, only slightly slower. Adapts seamlessly to generation!



ICML
International Conference
On Machine Learning

Oregon State University
College of Engineering

Want to find out more?

Find us at our poster session or send us an e-mail!
(agostinv, chenliz)@oregonstate.edu