Multi Resolution Analysis (MRA) for Approximate Self-Attention

Zhanpeng Zeng

Sourav Pal

Jeffery Kline

Glenn Fung

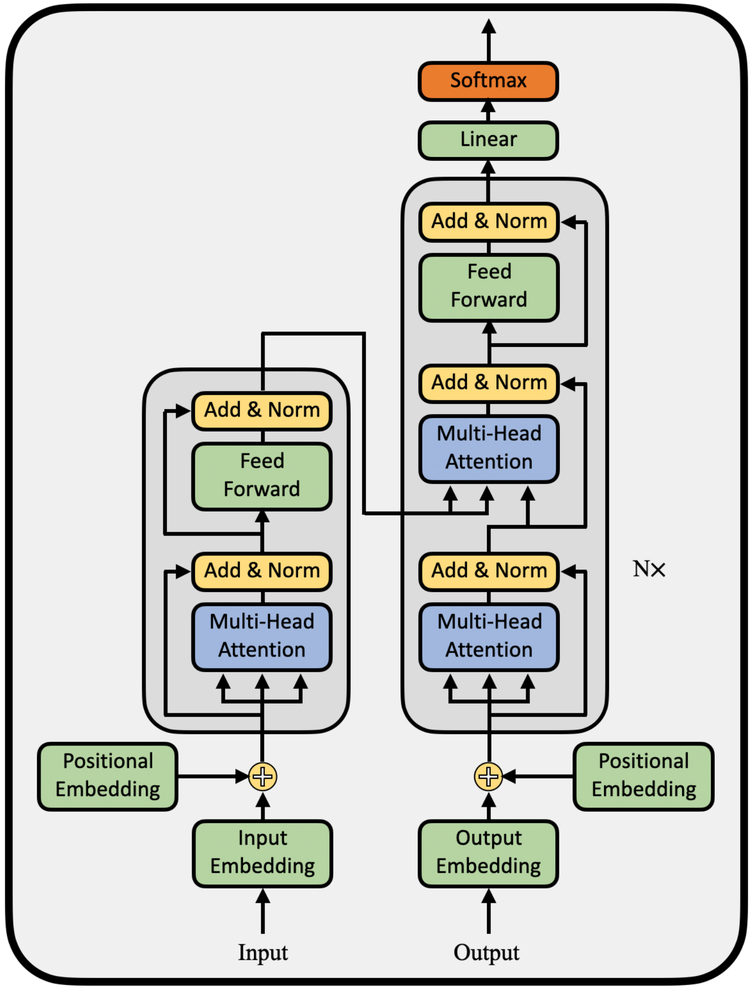
Vikas Singh

Motivation

Transformer

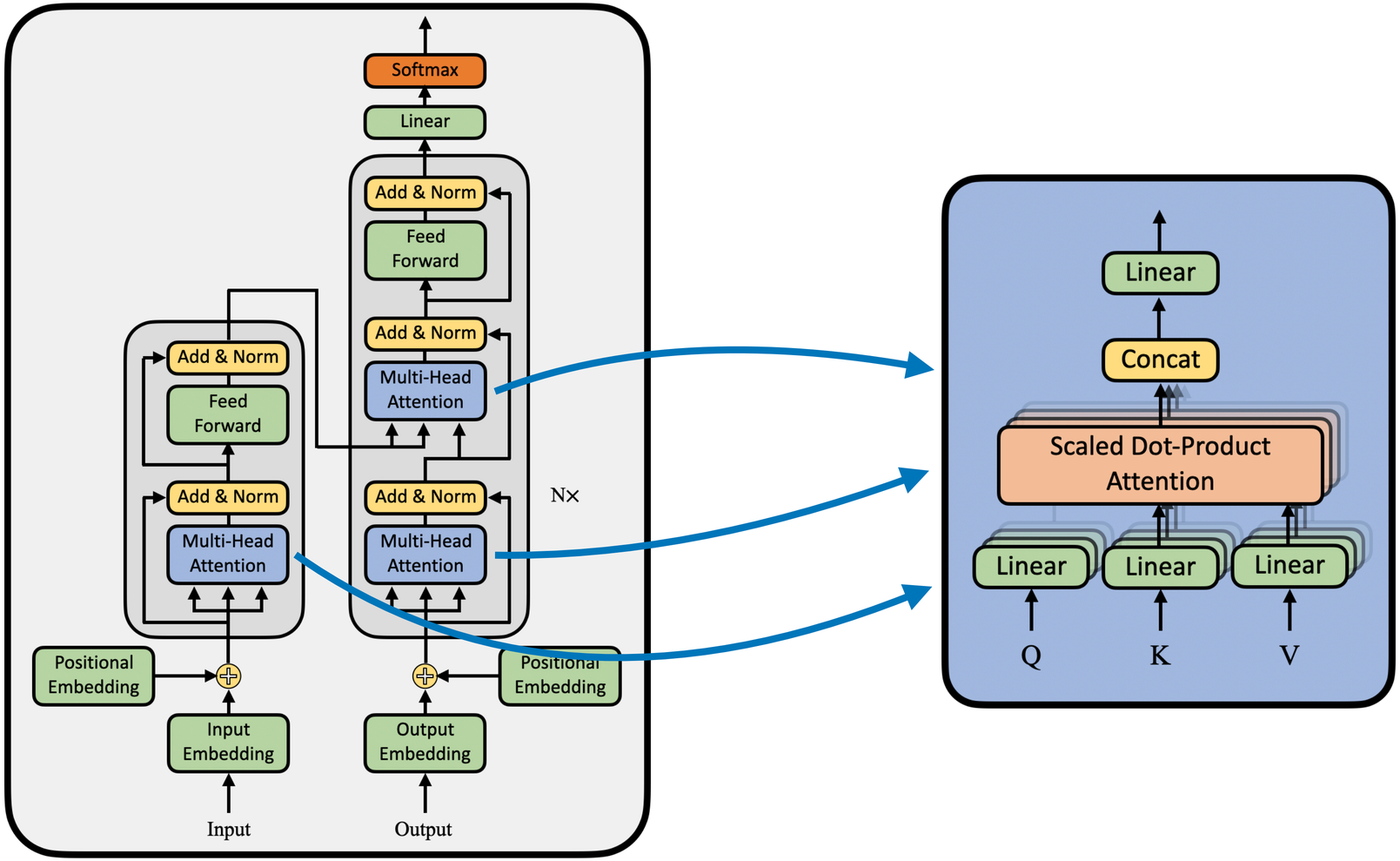
Motivation

Transformer



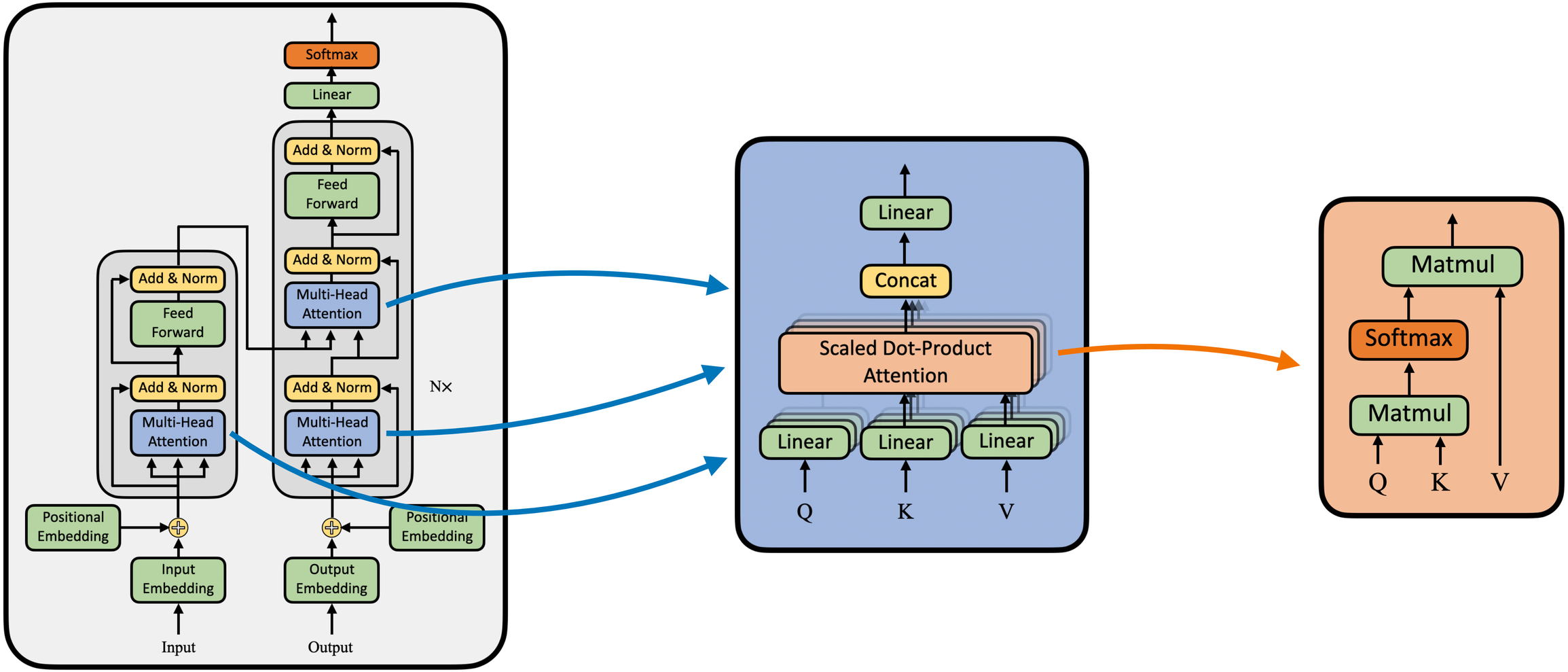
Motivation

Transformer



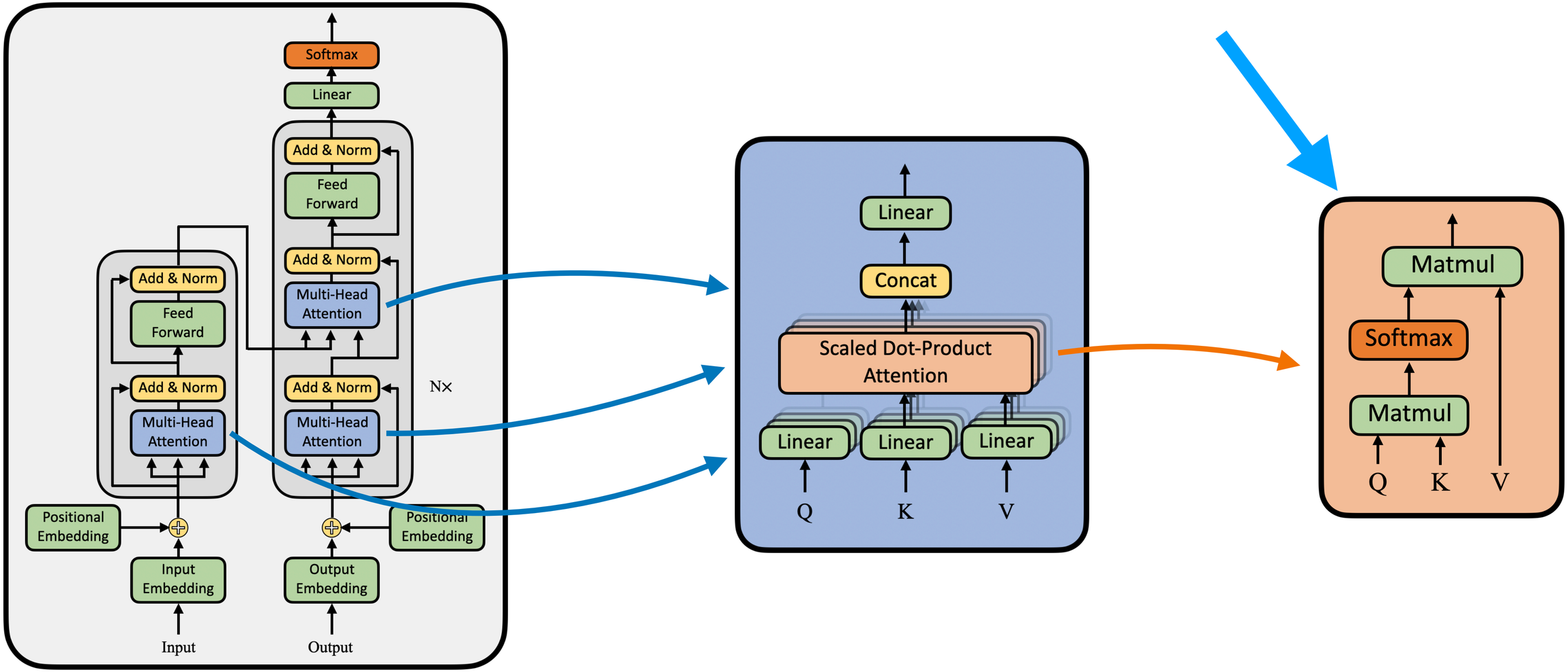
Motivation

Transformer



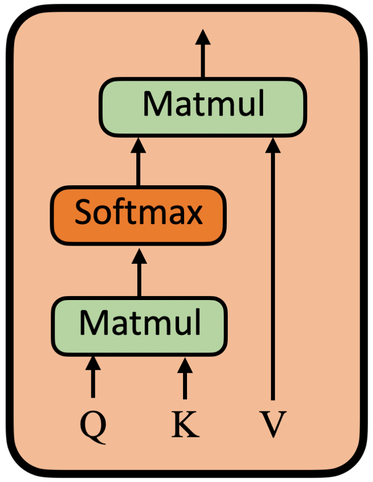
Motivation

Transformer



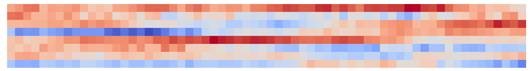
Motivation

Transformer



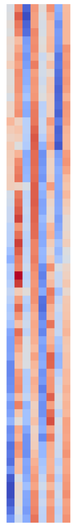
Motivation

Transformer

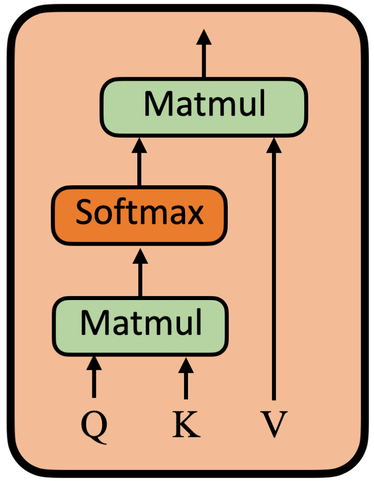


K^T

$Q, K, V \in \mathbb{R}^{n \times d}$

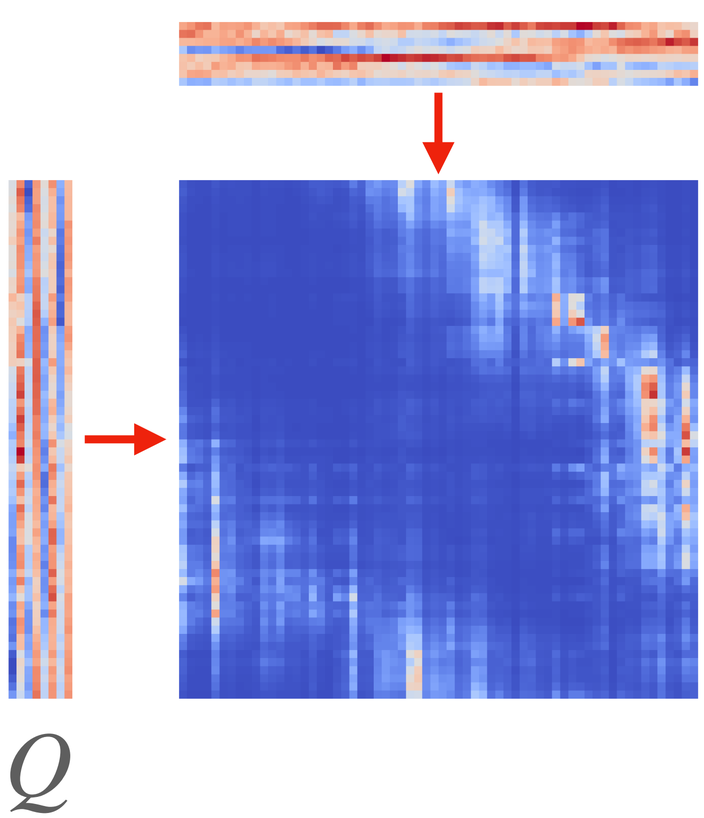


Q



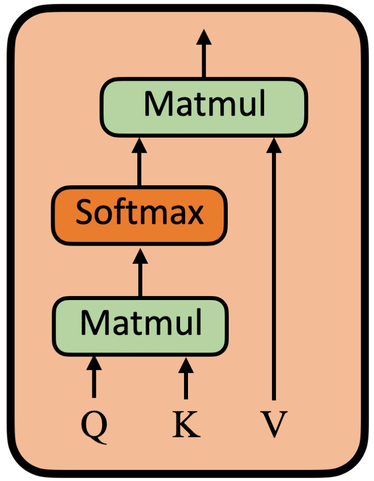
Motivation

Transformer



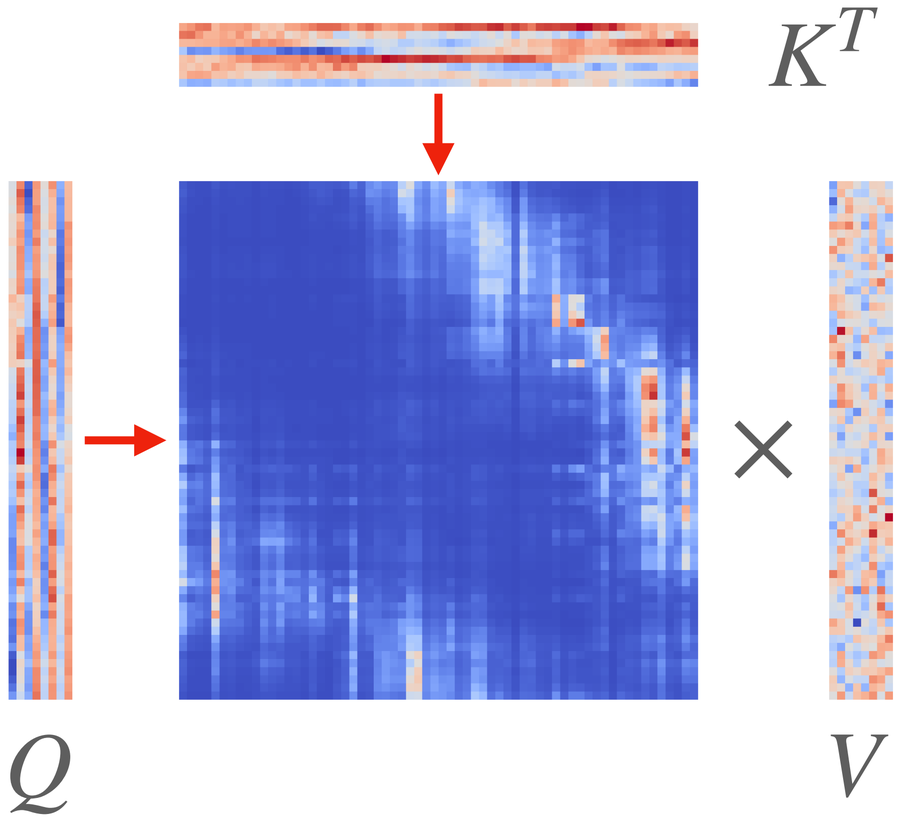
$$Q, K, V \in \mathbb{R}^{n \times d}$$

$$\exp(QK^T)$$



Motivation

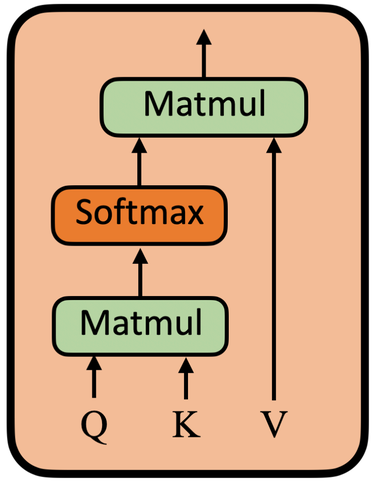
Transformer



$$Q, K, V \in \mathbb{R}^{n \times d}$$

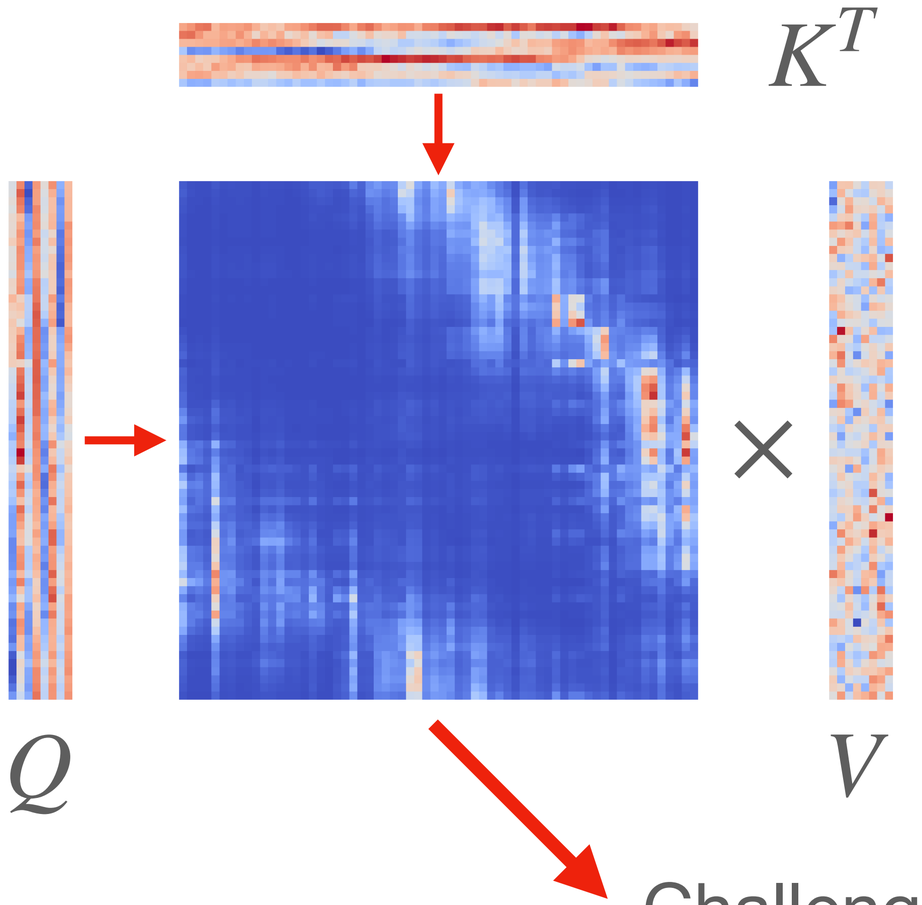
$$\exp(QK^T)$$

$$\exp(QK^T)V$$



Motivation

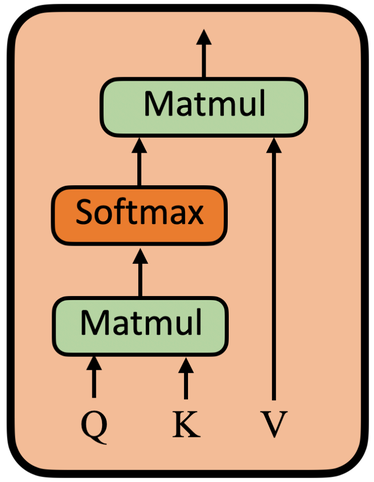
Transformer



$$Q, K, V \in \mathbb{R}^{n \times d}$$

$$\exp(QK^T)$$

$$\exp(QK^T)V$$



Challenge: $O(n^2)$ complexity

Related Works

REFORMER: THE EFFICIENT TRANSFORMER

Nikita Kitaev*
U.C. Berkeley & Google Research
kitaev@cs.berkeley.edu

Łukasz Kaiser*
Google Research
{lukaszkaiser,levsk

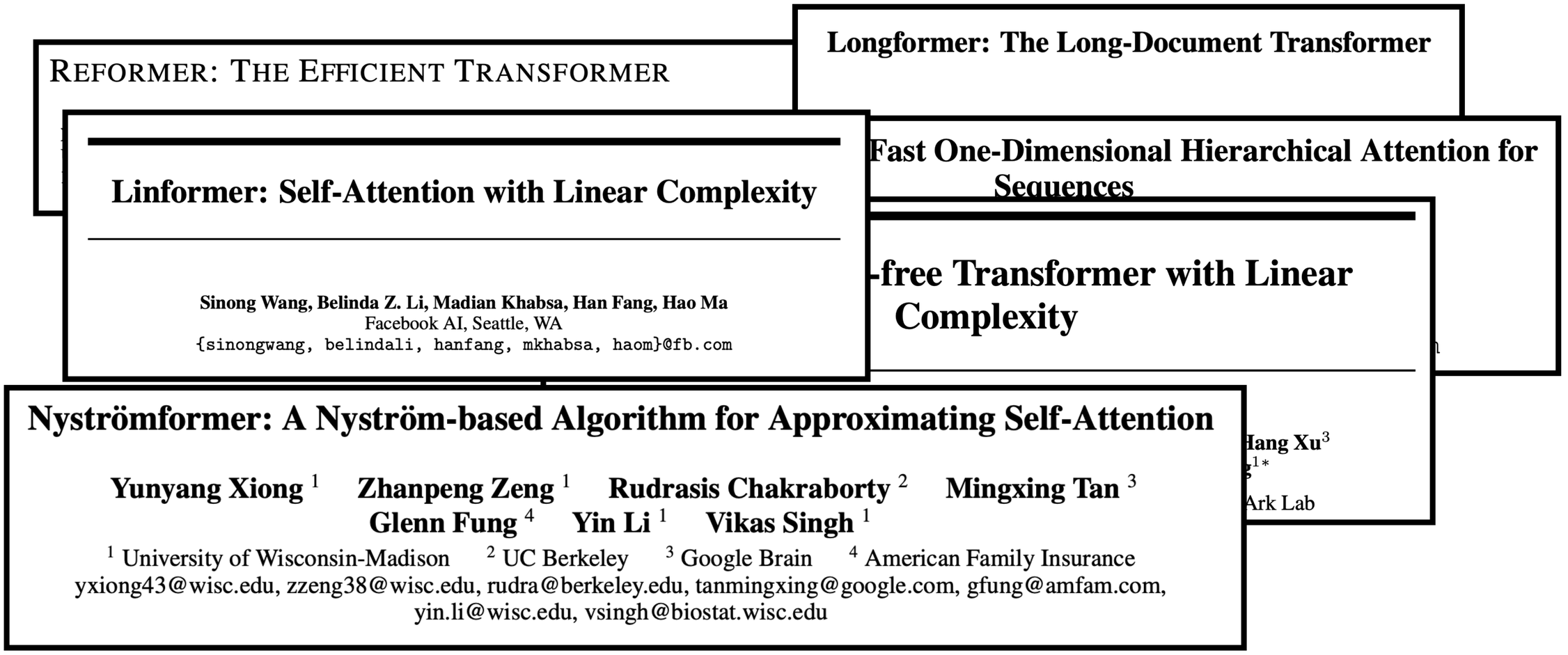
Longformer: The Long-Document Transformer

H-Transformer-1D: Fast One-Dimensional Hierarchical Attention for Sequences

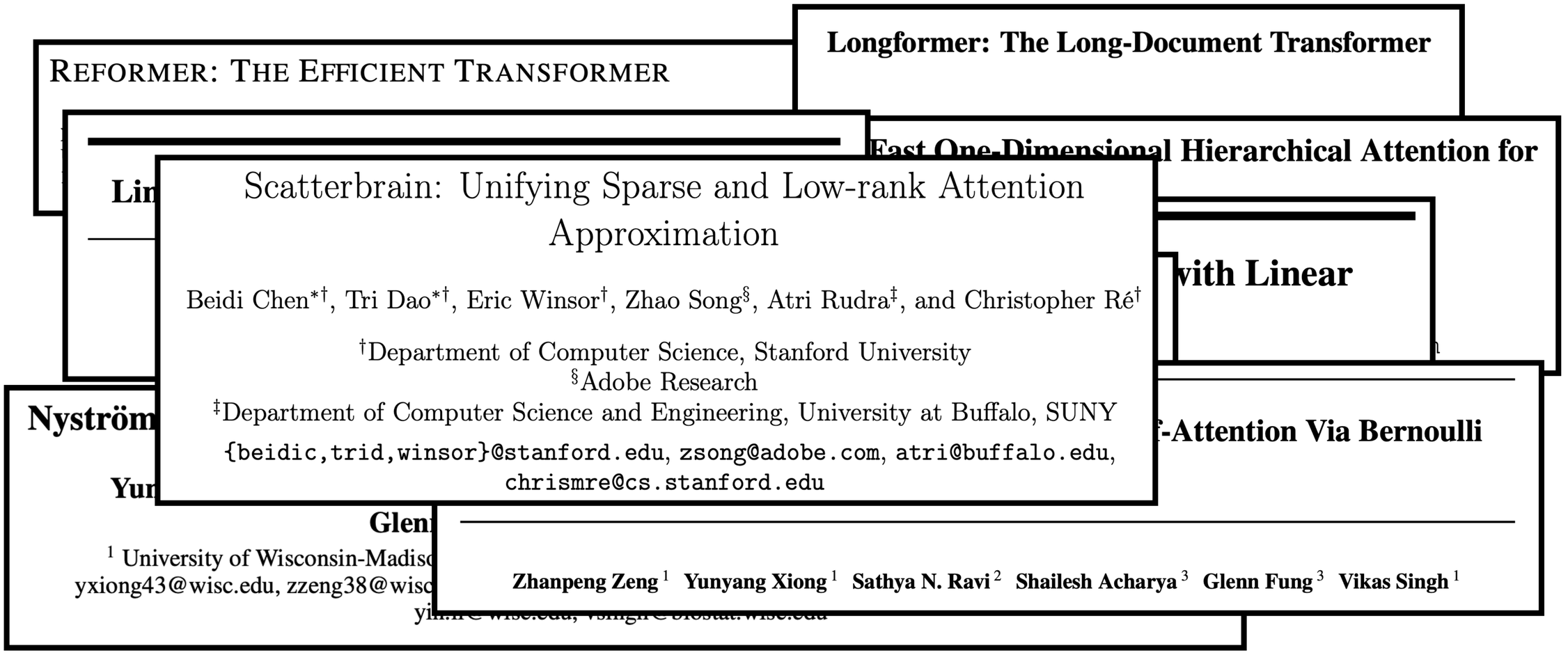
Zhenhai Zhu
Google Research
zhenhai@google.com

Radu Soricut
Google Research
rsoricut@google.com

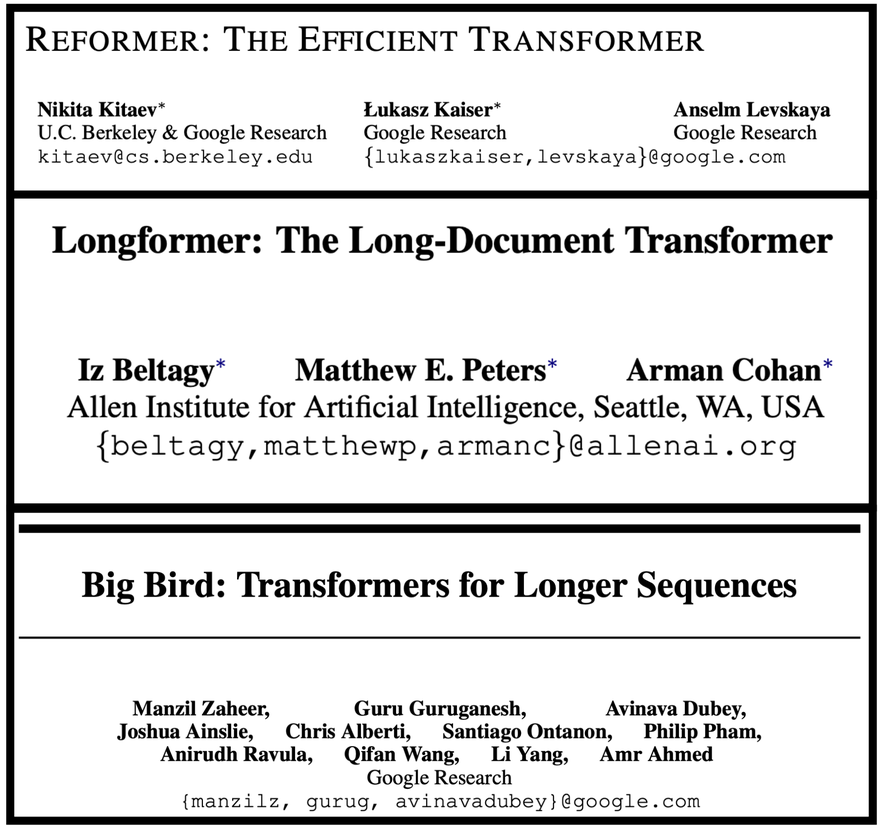
Related Works



Related Works



Related Works



REFORMER: THE EFFICIENT TRANSFORMER

Nikita Kitaev* U.C. Berkeley & Google Research kitaev@cs.berkeley.edu

Łukasz Kaiser* Google Research {lukaszkaiser,levskaya}@google.com

Anselm Levskaya Google Research

Longformer: The Long-Document Transformer

Iz Beltagy* **Matthew E. Peters*** **Arman Cohan***
Allen Institute for Artificial Intelligence, Seattle, WA, USA
{beltagy,matthewp,armanc}@allenai.org

Big Bird: Transformers for Longer Sequences

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, Amr Ahmed
Google Research
{manzilz, gurug, avinavadubey}@google.com

Related Works

REFORMER: THE EFFICIENT TRANSFORMER

Nikita Kitaev* U.C. Berkeley & Google Research kitaev@cs.berkeley.edu
Łukasz Kaiser* Google Research {lukaszkaiser,levskaya}@google.com
Anselm Levskaya Google Research

Longformer: The Long-Document Transformer

Iz Beltagy* **Matthew E. Peters*** **Arman Cohan***
Allen Institute for Artificial Intelligence, Seattle, WA, USA
{beltagy,matthewp,armanc}@allenai.org

Big Bird: Transformers for Longer Sequences

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, Amr Ahmed
Google Research
{manzilz, gurug, avinavadubey}@google.com

RETHINKING ATTENTION WITH PERFORMERS

Krzysztof Choromanski*[1], Valerii Likhosherstov*[2], David Dohan*[1], Xingyou Song*[1] Andreea Gane*[1], Tamas Sarlos*[1], Peter Hawkins*[1], Jared Davis*[3], Afroz Mohiuddin[1] Lukasz Kaiser[1], David Belanger[1], Lucy Colwell[1,2], Adrian Weller[2,4]
[1]Google [2]University of Cambridge [3]DeepMind [4]Alan Turing Institute

Linformer: Self-Attention with Linear Complexity

Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, Hao Ma
Facebook AI, Seattle, WA
{sinongwang, belindali, hanfang, mkhabsa, haom}@fb.com

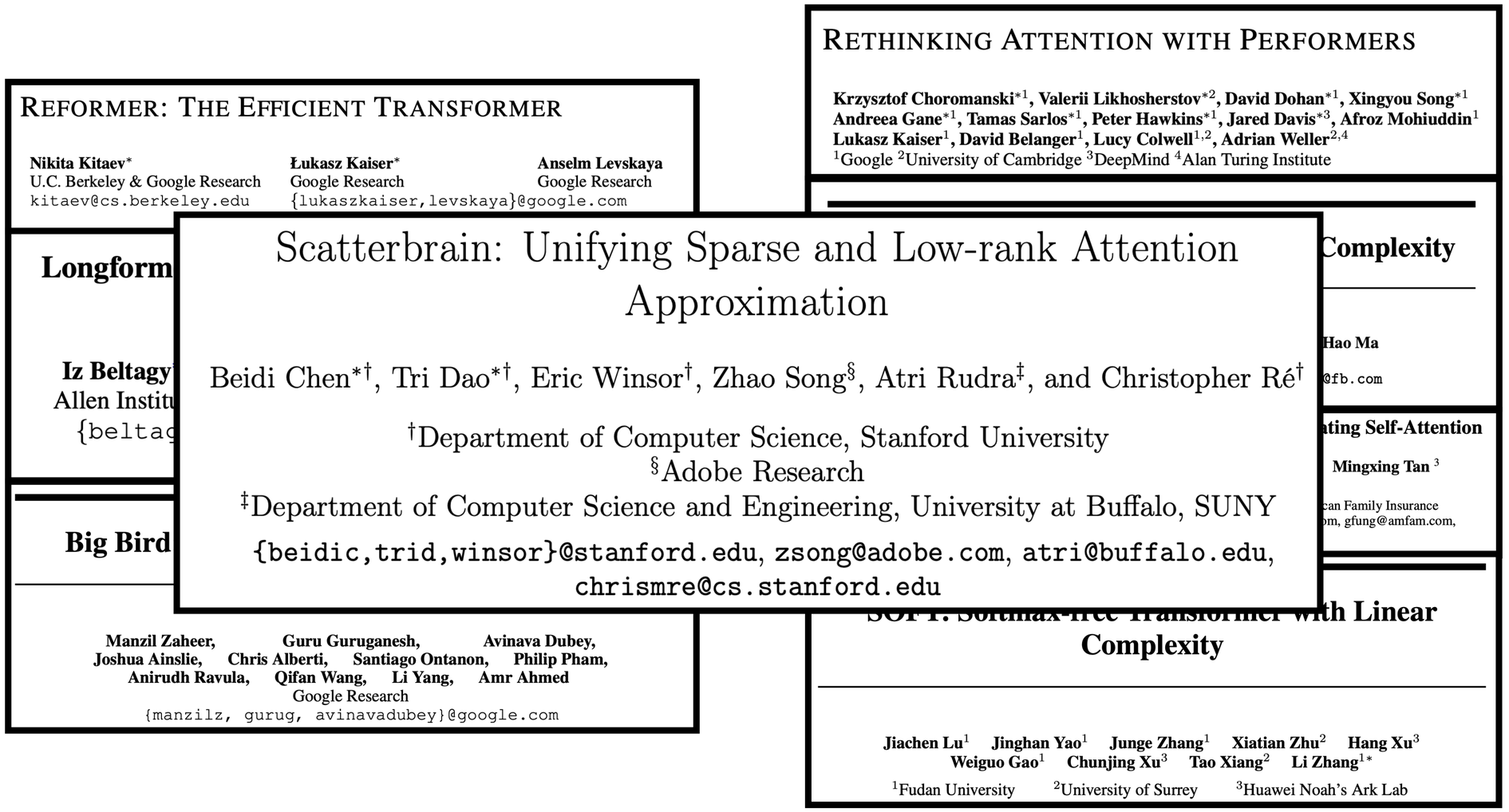
Nyströmformer: A Nyström-based Algorithm for Approximating Self-Attention

Yunyang Xiong [1] Zhanpeng Zeng [1] Rudrasis Chakraborty [2] Mingxing Tan [3] Glenn Fung [4] Yin Li [1] Vikas Singh [1]
[1] University of Wisconsin-Madison [2] UC Berkeley [3] Google Brain [4] American Family Insurance
yxiong43@wisc.edu, zzeng38@wisc.edu, rudra@berkeley.edu, tanmingxing@google.com, gfung@amfam.com, yin.li@wisc.edu, vsingh@biostat.wisc.edu

SOFT: Softmax-free Transformer with Linear Complexity

Jiachen Lu[1] Jinghan Yao[1] Junge Zhang[1] Xiatian Zhu[2] Hang Xu[3] Weiguo Gao[1] Chunjing Xu[3] Tao Xiang[2] Li Zhang[1*]
[1]Fudan University [2]University of Surrey [3]Huawei Noah's Ark Lab

Related Works



REFORMER: THE EFFICIENT TRANSFORMER

Nikita Kitaev* U.C. Berkeley & Google Research kitaev@cs.berkeley.edu

Łukasz Kaiser* Google Research {lukaszkaiser,levskaya}@google.com

Anselm Levskaya Google Research

RETHINKING ATTENTION WITH PERFORMERS

Krzysztof Choromanski*[1], Valerii Likhosherstov*[2], David Dohan*[1], Xingyou Song*[1] Andreea Gane*[1], Tamas Sarlos*[1], Peter Hawkins*[1], Jared Davis*[3], Afroz Mohiuddin[1] Lukasz Kaiser[1], David Belanger[1], Lucy Colwell[1,2], Adrian Weller[2,4]
[1]Google [2]University of Cambridge [3]DeepMind [4]Alan Turing Institute

Scatterbrain: Unifying Sparse and Low-rank Attention Approximation

Beidi Chen*†, Tri Dao*†, Eric Winsor†, Zhao Song§, Atri Rudra‡, and Christopher Ré†

†Department of Computer Science, Stanford University
§Adobe Research
‡Department of Computer Science and Engineering, University at Buffalo, SUNY
{beidic,trid,winsor}@stanford.edu, zsong@adobe.com, atri@buffalo.edu, chrismre@cs.stanford.edu

Longform

Iz Beltagy
Allen Institu
{beltag

Complexity

Hao Ma
@fb.com

ting Self-Attention

Mingxing Tan [3]
can Family Insurance
om, gfung@amfam.com,

Big Bird

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, Amr Ahmed
Google Research
{manzilz, gurug, avinavadubey}@google.com

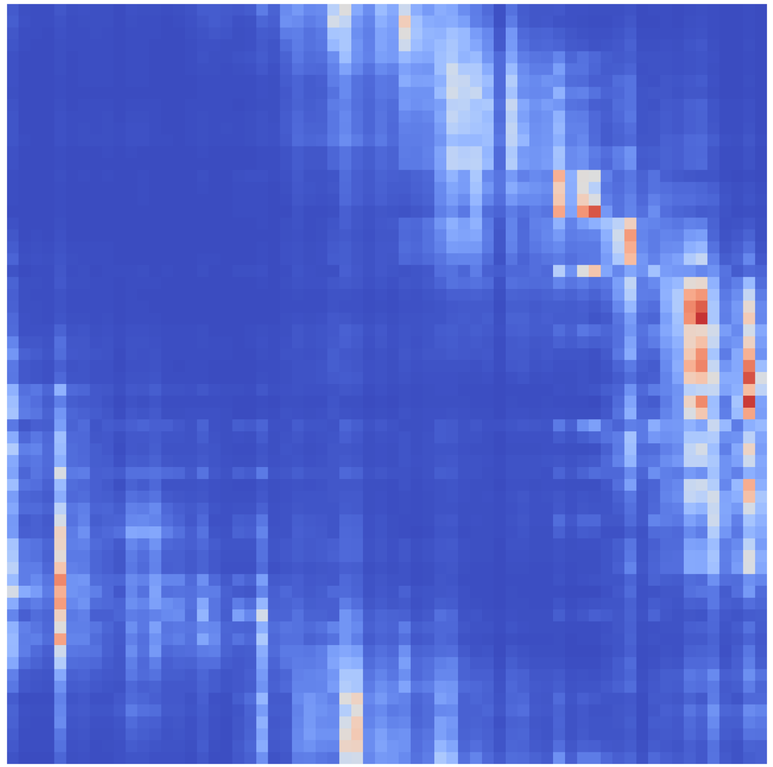
SOFT: Softmax-free Transformer with Linear Complexity

Jiachen Lu[1] Jinghan Yao[1] Junge Zhang[1] Xiatian Zhu[2] Hang Xu[3] Weiguo Gao[1] Chunjing Xu[3] Tao Xiang[2] Li Zhang[1*]
[1]Fudan University [2]University of Surrey [3]Huawei Noah's Ark Lab

Multi-Resolution Approximation

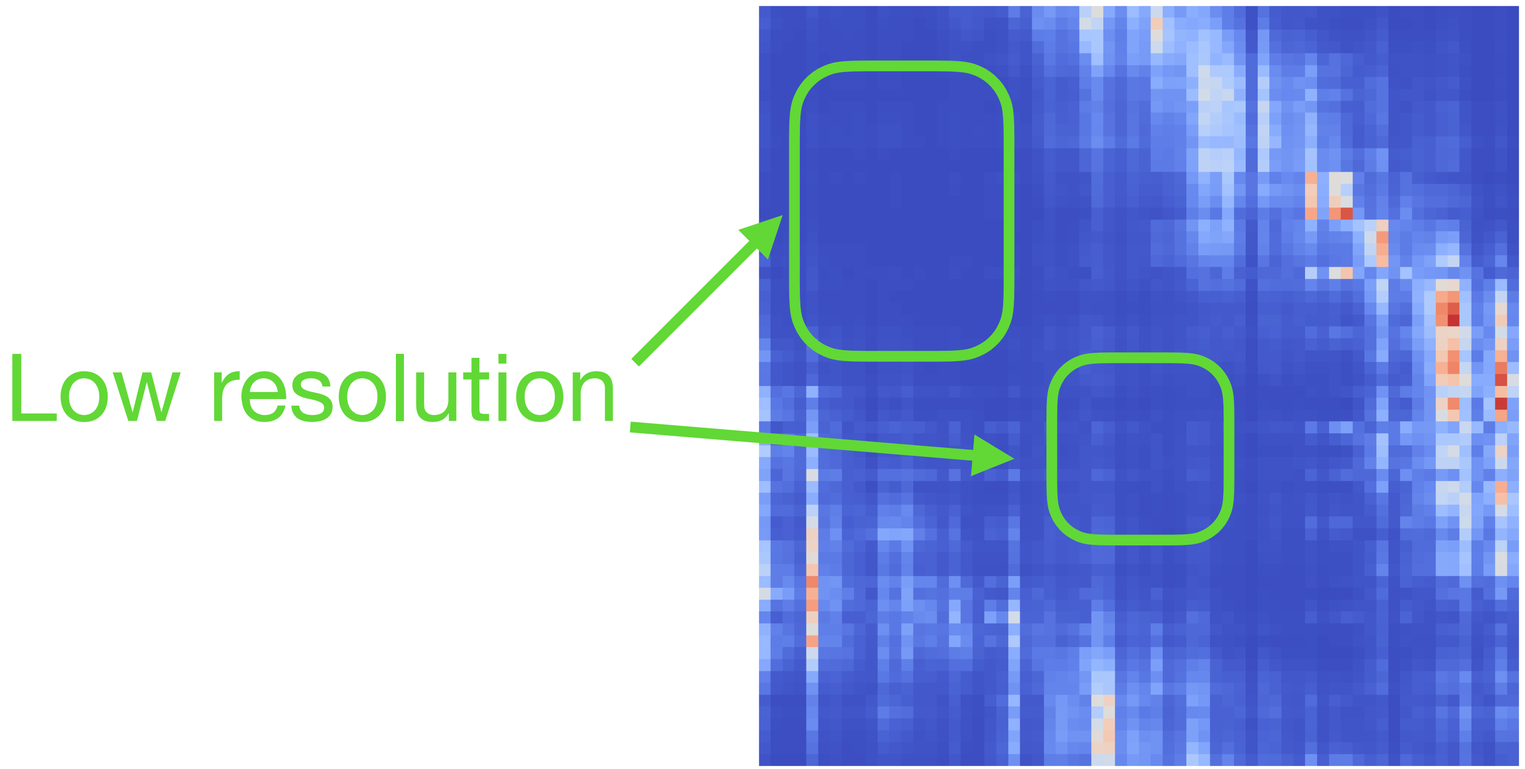
Multi-Resolution Approximation

Why?



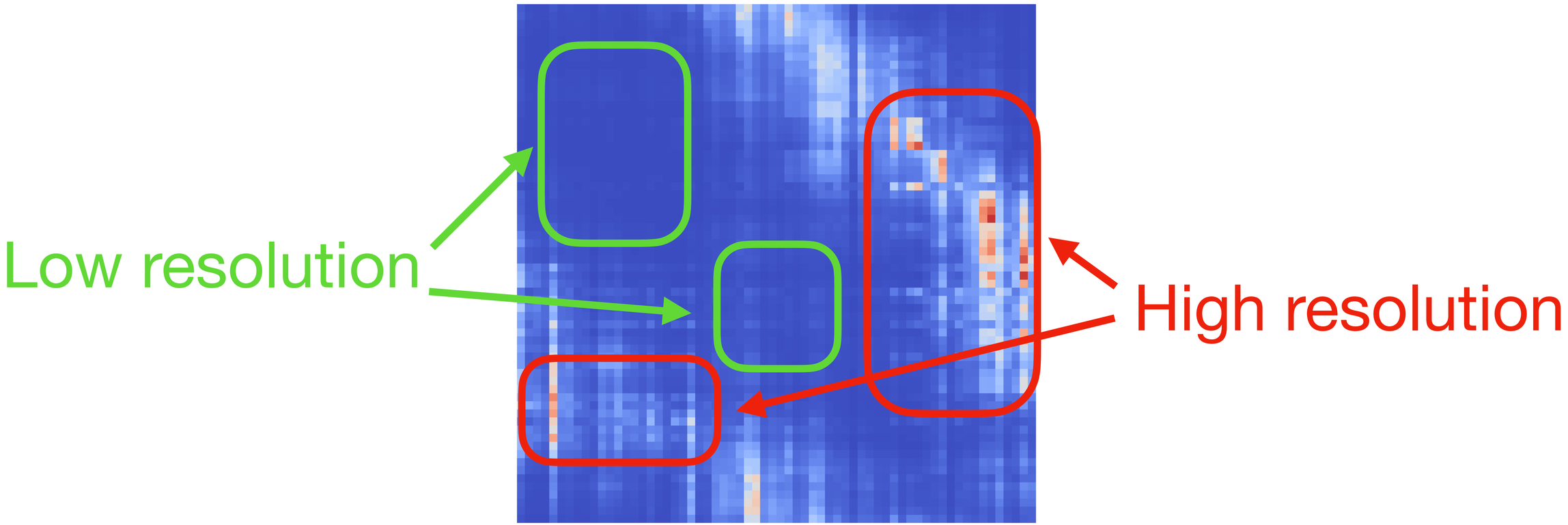
Multi-Resolution Approximation

Why?



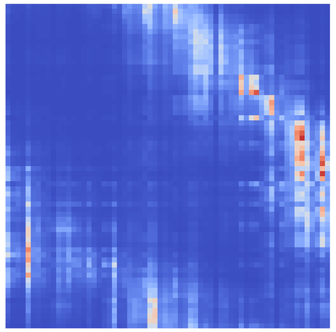
Multi-Resolution Approximation

Why?



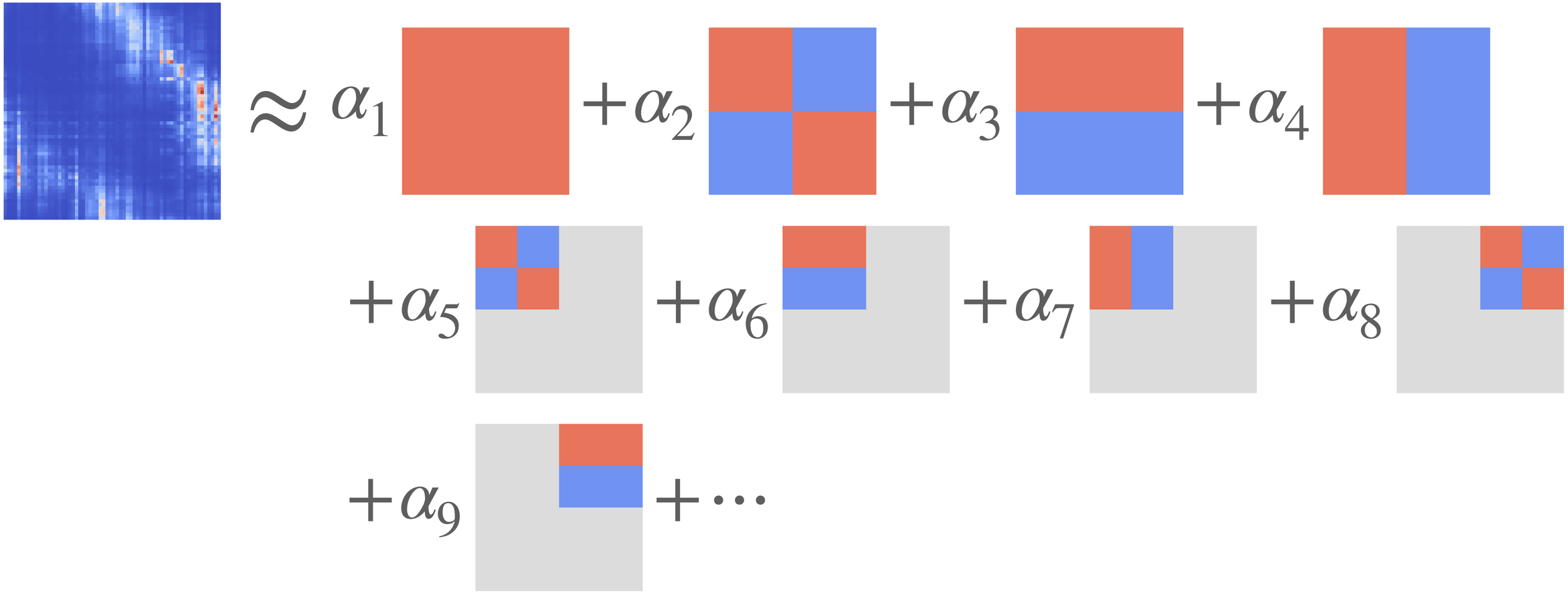
Multi-Resolution Approximation

Wavelets



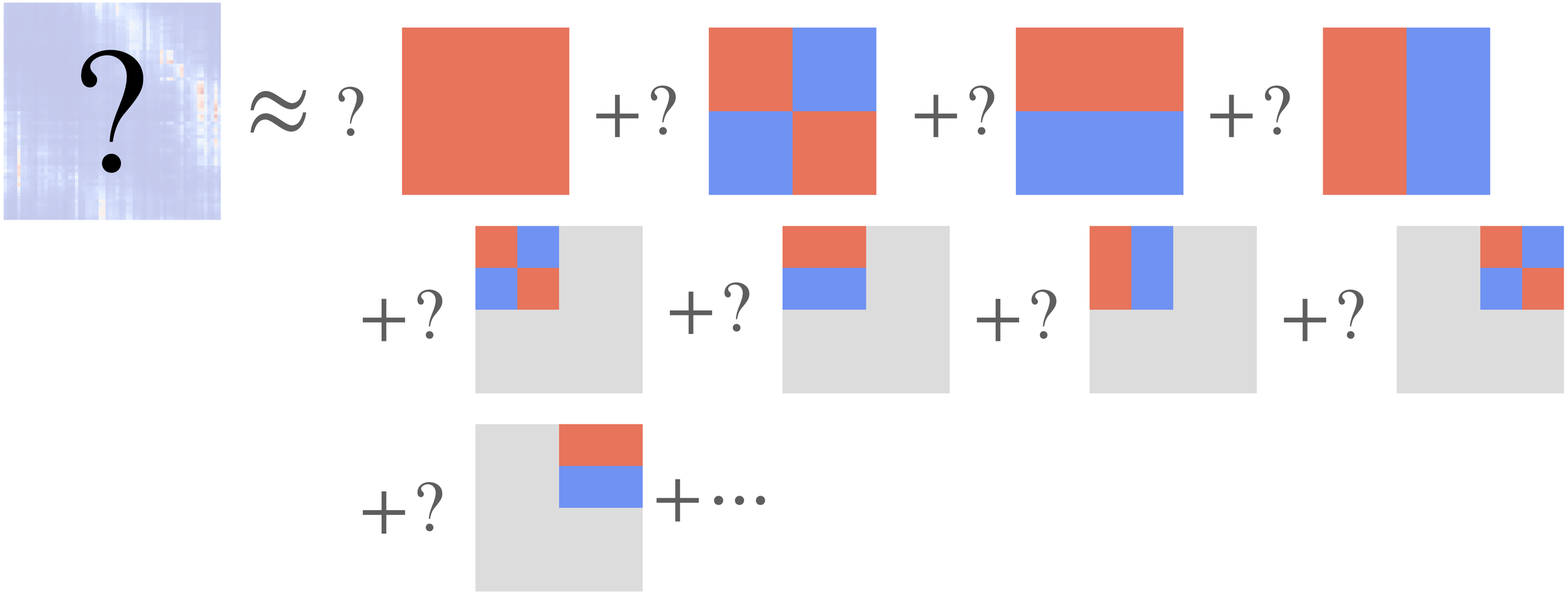
Multi-Resolution Approximation

Wavelets



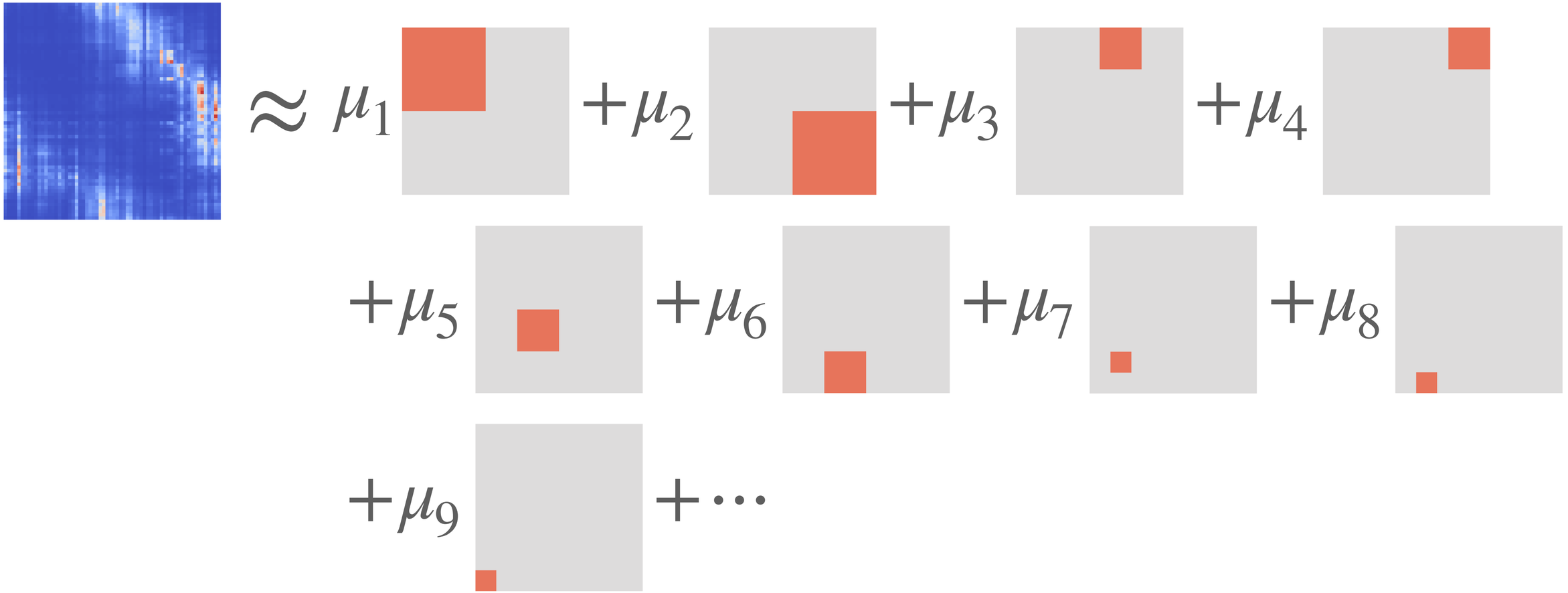
Multi-Resolution Approximation

Wavelets



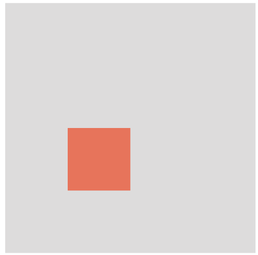
Multi-Resolution Approximation

Wavelets



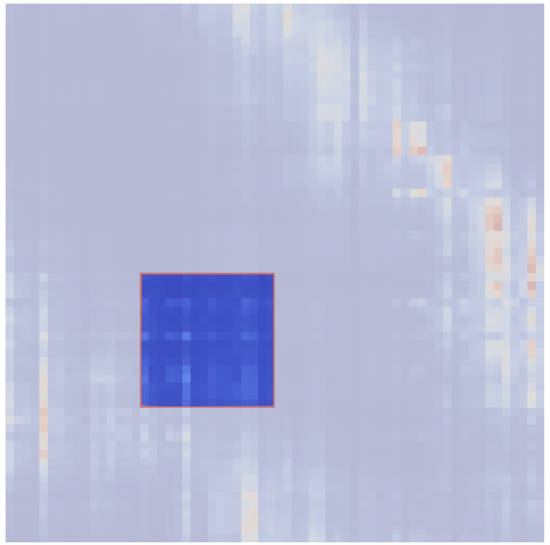
Multi-Resolution Approximation

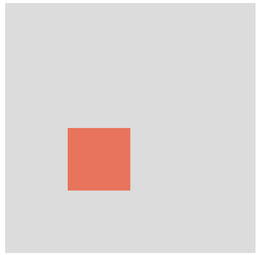
Coefficient Approximation

$\cdots + \mu_5$  $+ \cdots$

Multi-Resolution Approximation

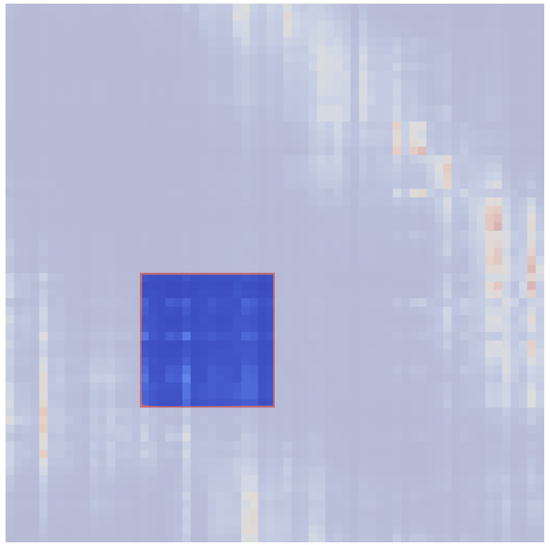
Coefficient Approximation

$\mu_5 = Avg($  $)$

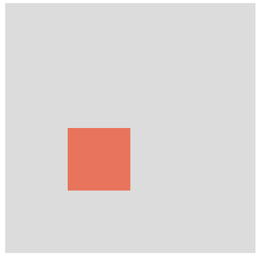
$\cdots + \mu_5$  $+ \cdots$

Multi-Resolution Approximation

Coefficient Approximation

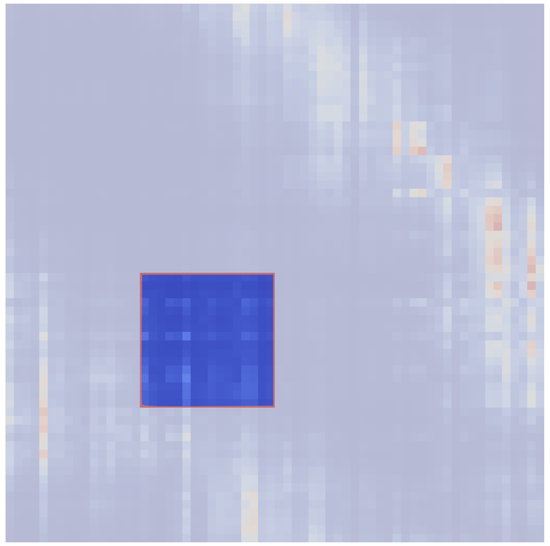
$$\mu_5 = Avg($$  $$)$$

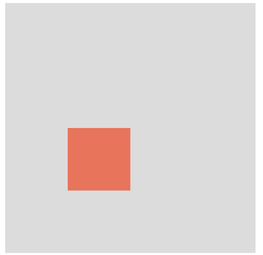
$$\frac{1}{N} \sum_{(B_5)_{i,j} \neq 0} \exp(QK^T)_{i,j}$$

$$\cdots + \mu_5$$  $$+ \cdots$$

Multi-Resolution Approximation

Coefficient Approximation

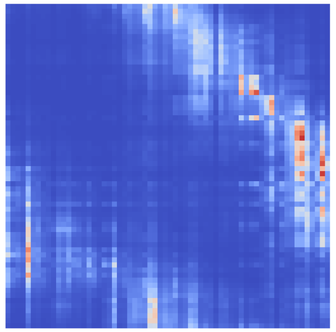
$\mu_5 = Avg($  $)$

$\cdots + \mu_5$  $+ \cdots$

$$\frac{1}{N} \sum_{(B_5)_{i,j} \neq 0} \exp(QK^T)_{i,j} \xrightarrow{\text{Efficient}} \exp(\frac{1}{N} \sum_{(B_5)_{i,j} \neq 0} (QK^T)_{i,j})$$

Multi-Resolution Approximation

Component Selection



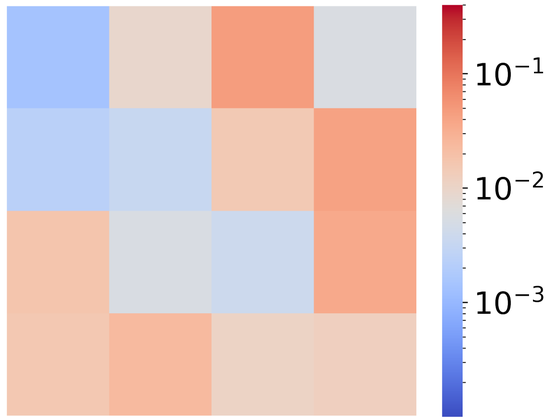
$$\approx \mu_1 \; ? \; + \mu_2 \; ? \; + \mu_3 \; ? \; + \mu_4 \; ?$$

$$+ \mu_5 \; ? \; + \mu_6 \; ? \; + \mu_7 \; ? \; + \mu_8 \; ?$$

$$+ \mu_9 \; ? \; + \cdots$$

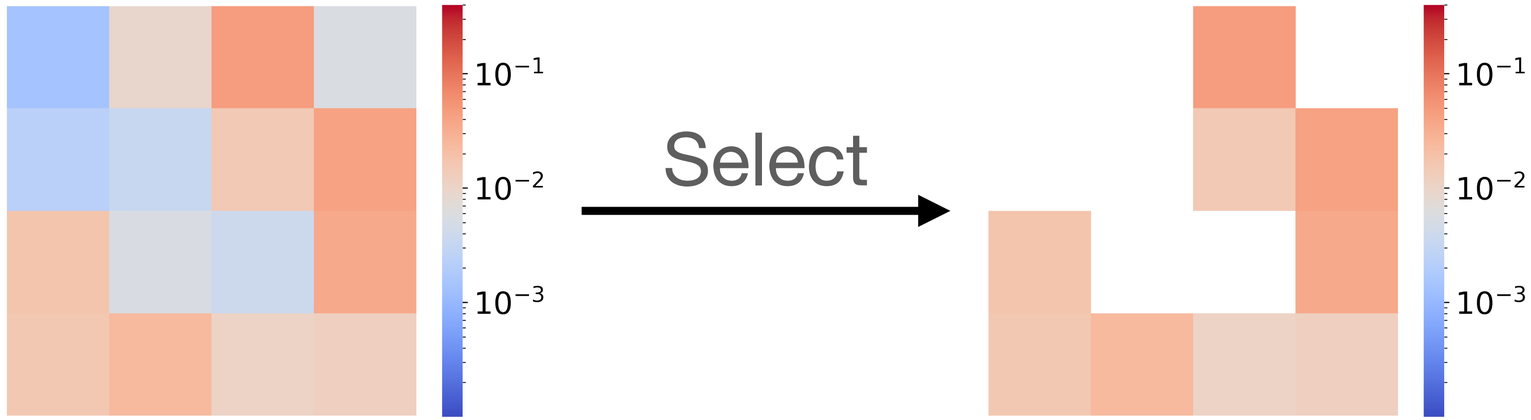
Multi-Resolution Approximation

Illustration



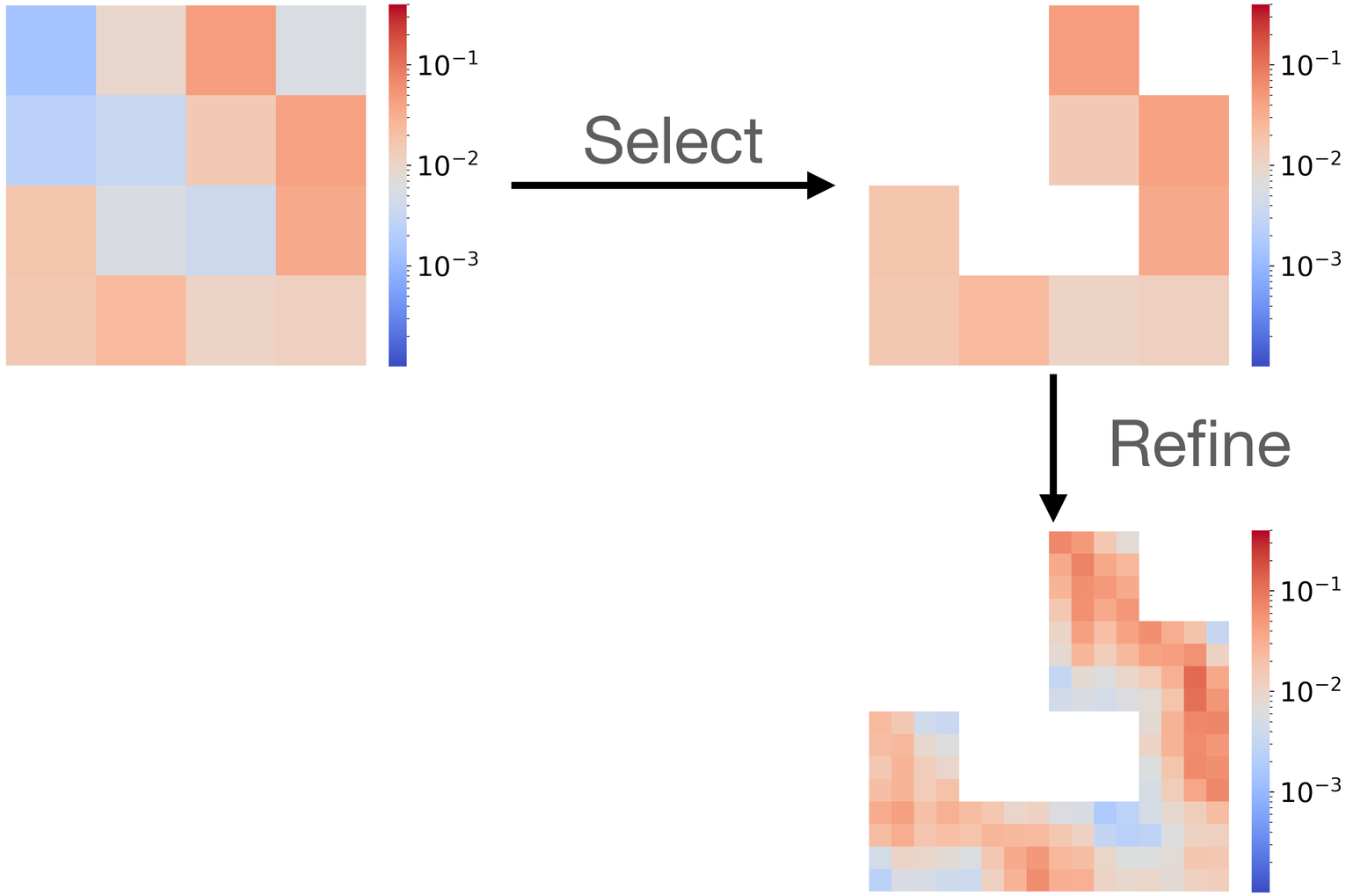
Multi-Resolution Approximation

Illustration



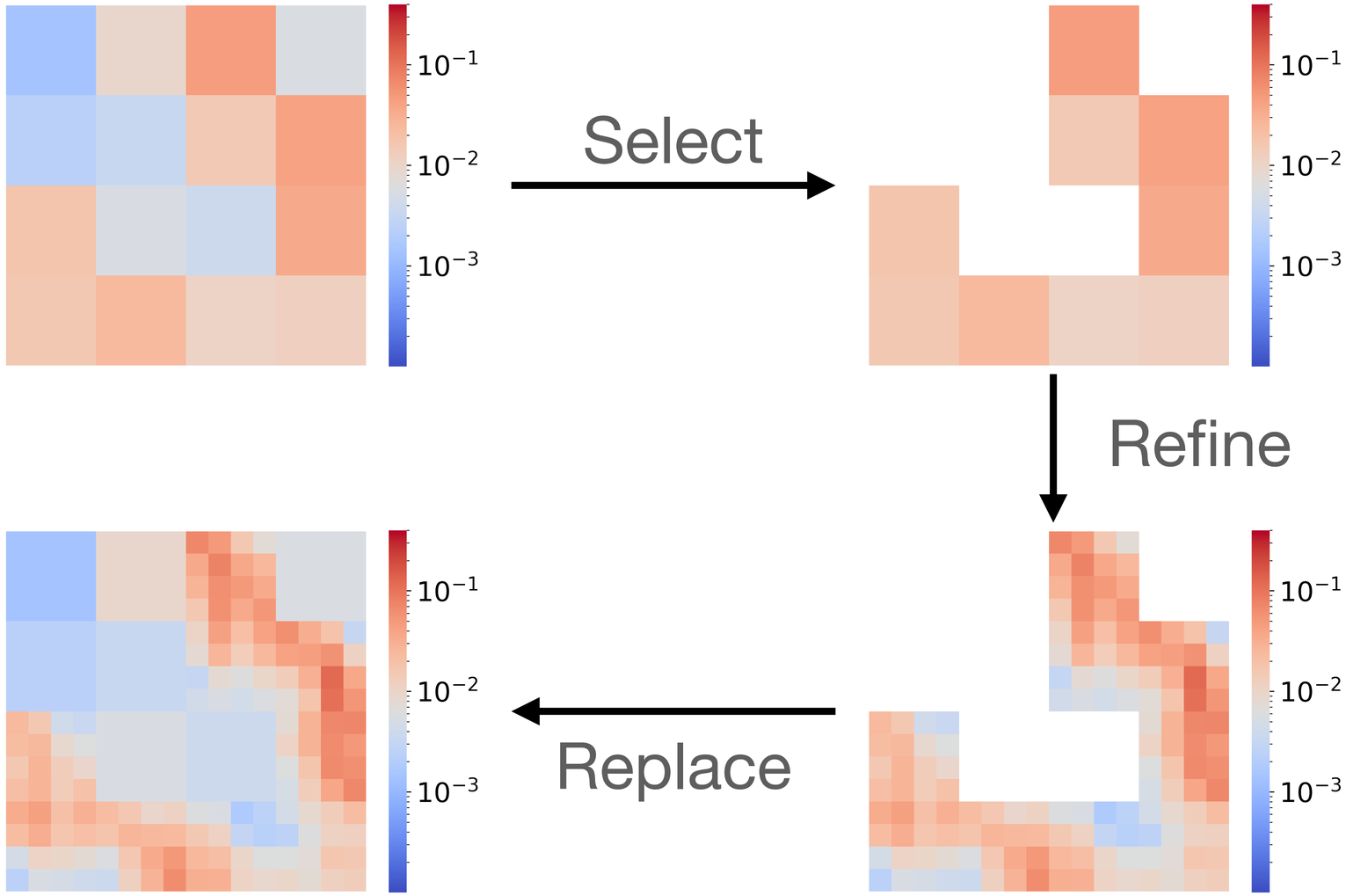
Multi-Resolution Approximation

Illustration



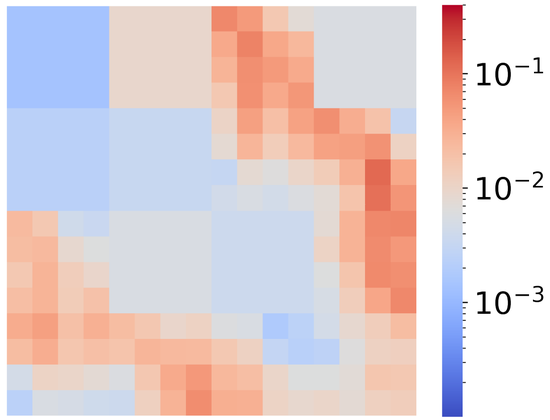
Multi-Resolution Approximation

Illustration



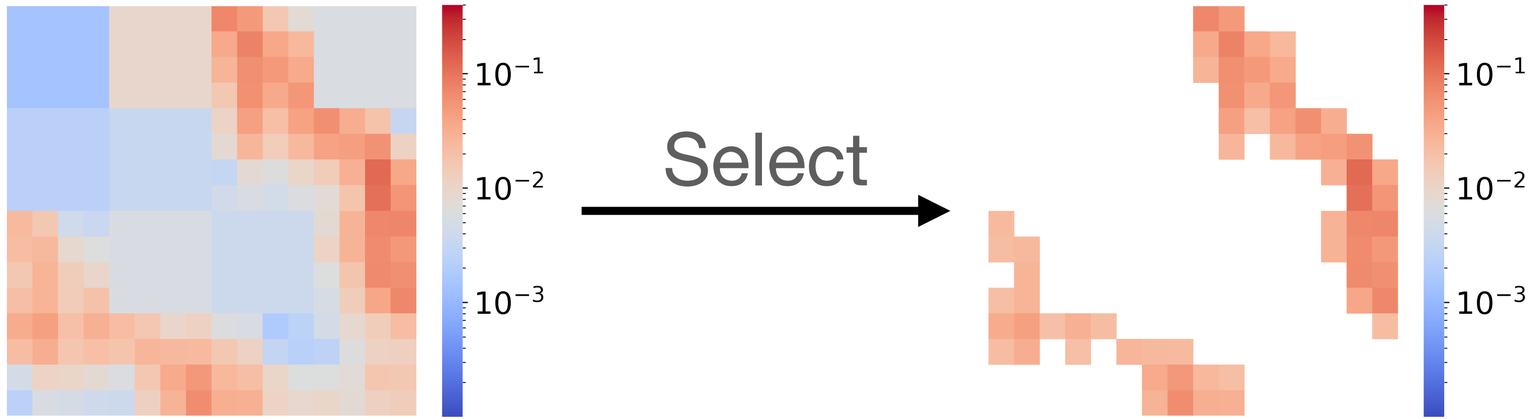
Multi-Resolution Approximation

Illustration



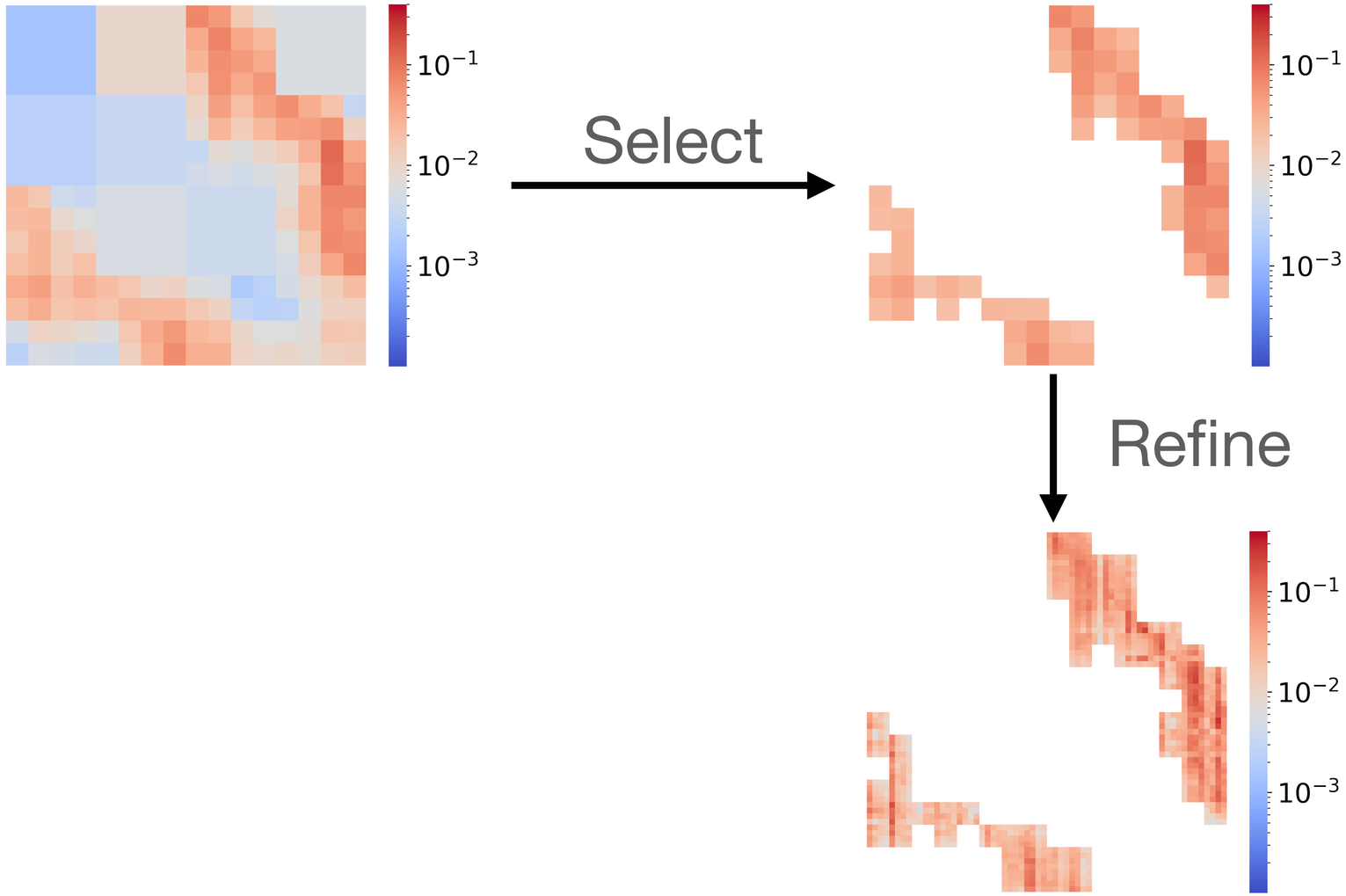
Multi-Resolution Approximation

Illustration



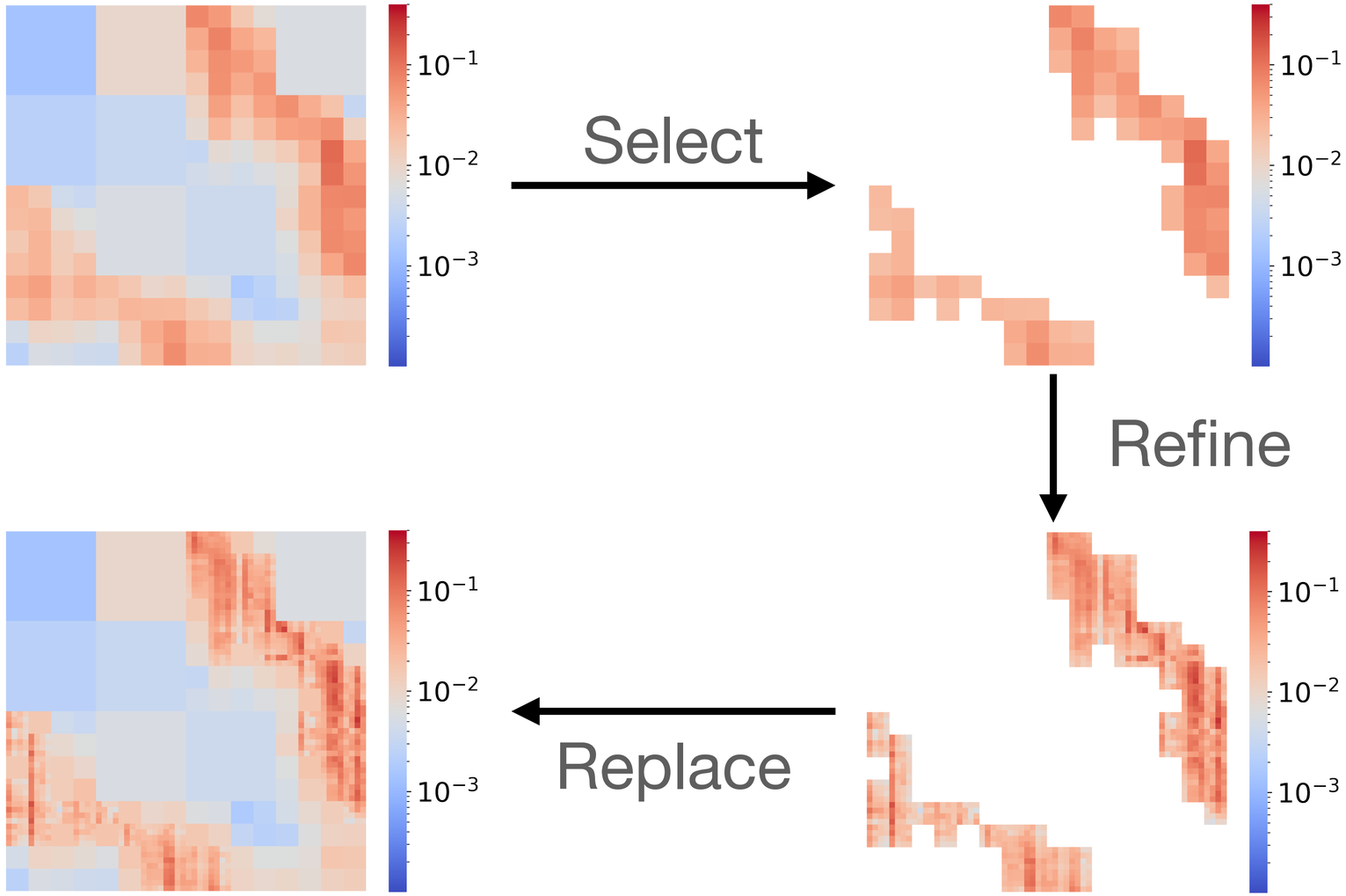
Multi-Resolution Approximation

Illustration



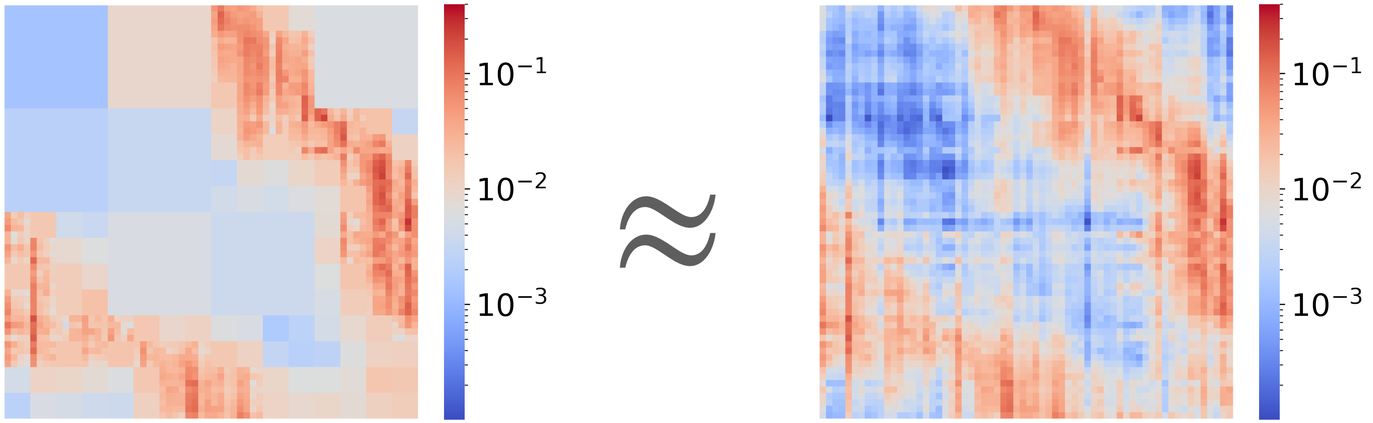
Multi-Resolution Approximation

Illustration



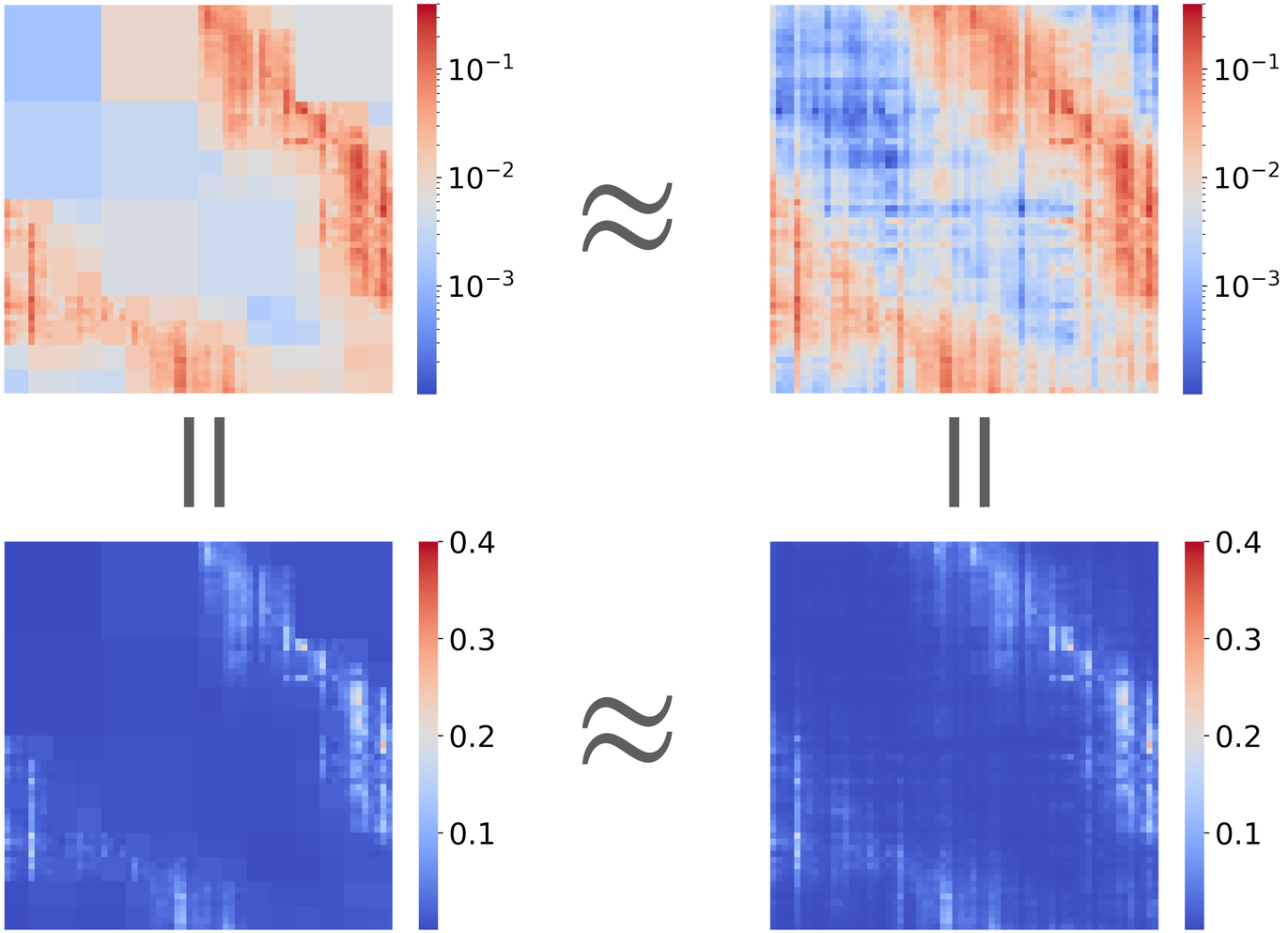
Multi-Resolution Approximation

Approximation Result



Multi-Resolution Approximation

Approximation Result

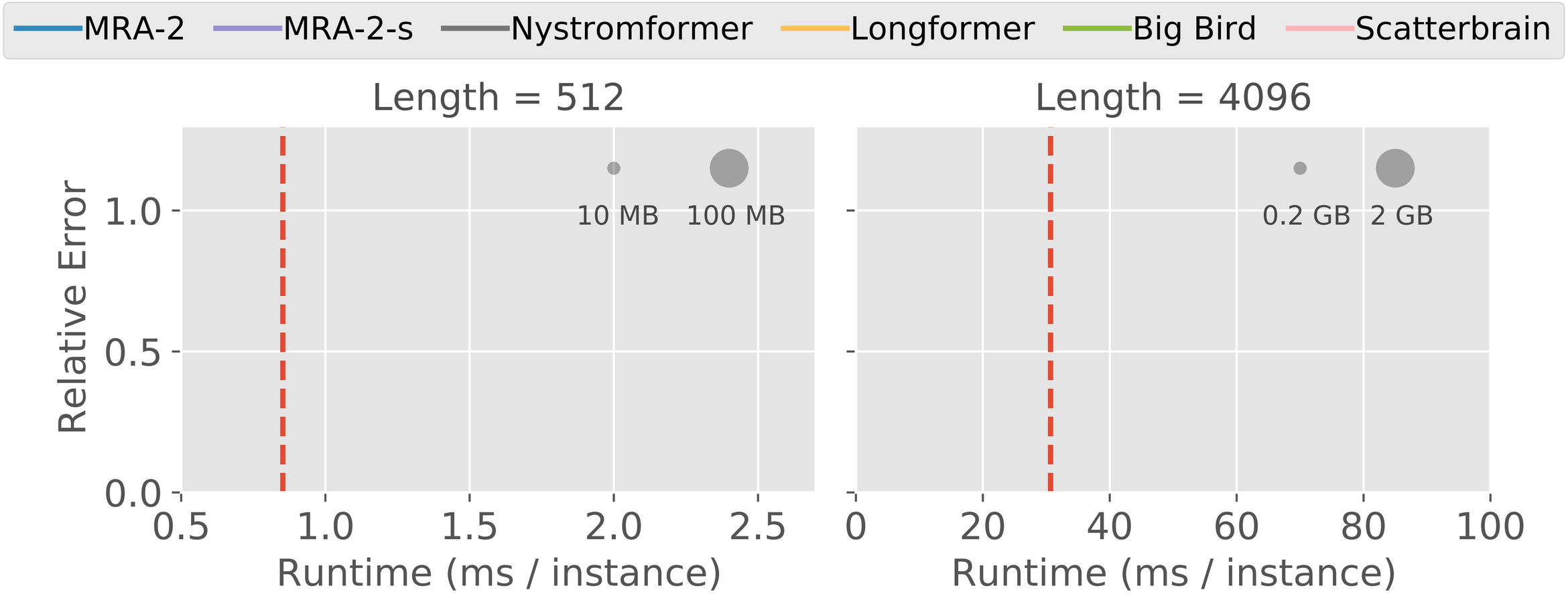


Evaluations

Efficiency versus Approximation

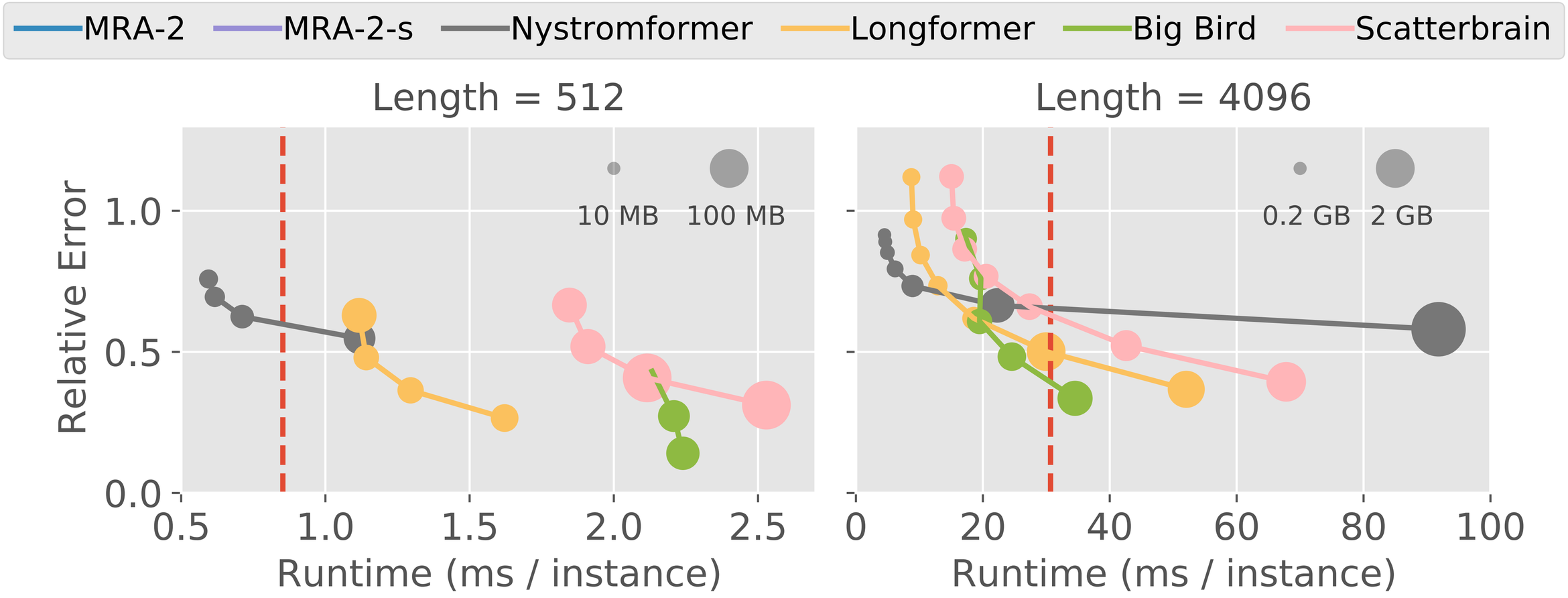
Evaluations

Efficiency v.s. Approximation



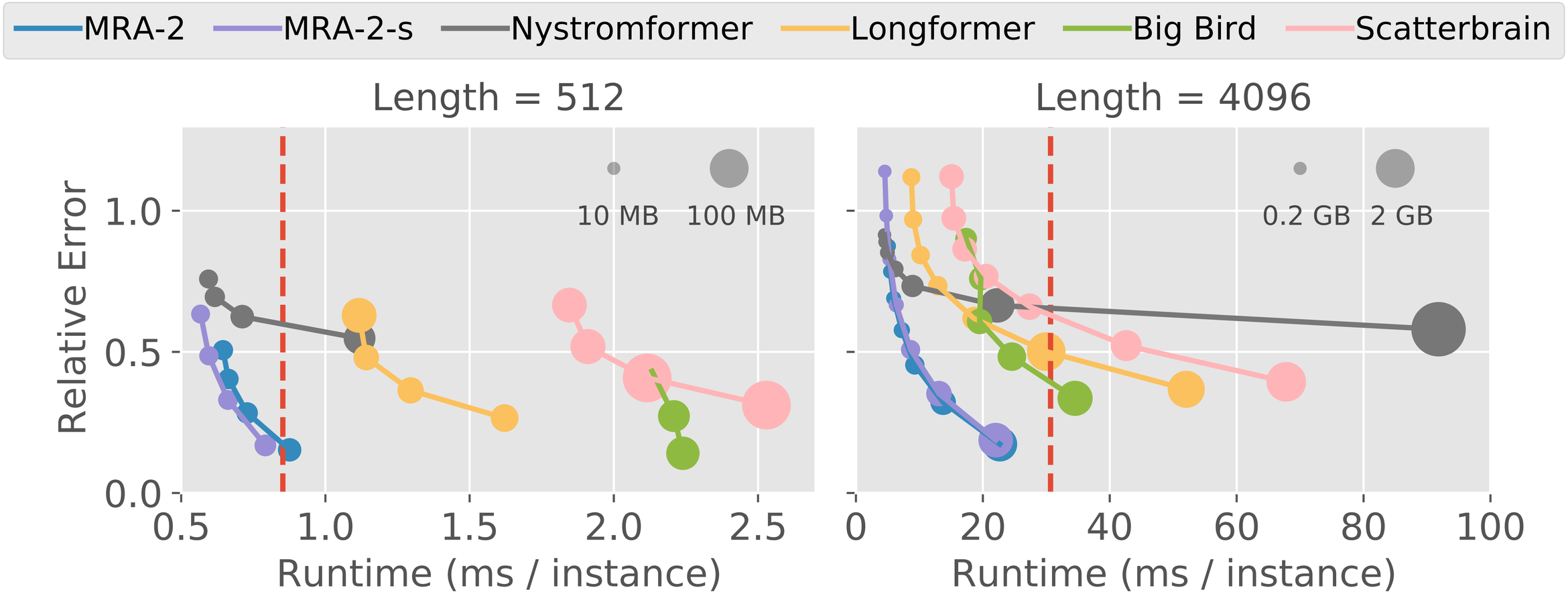
Evaluations

Efficiency v.s. Approximation



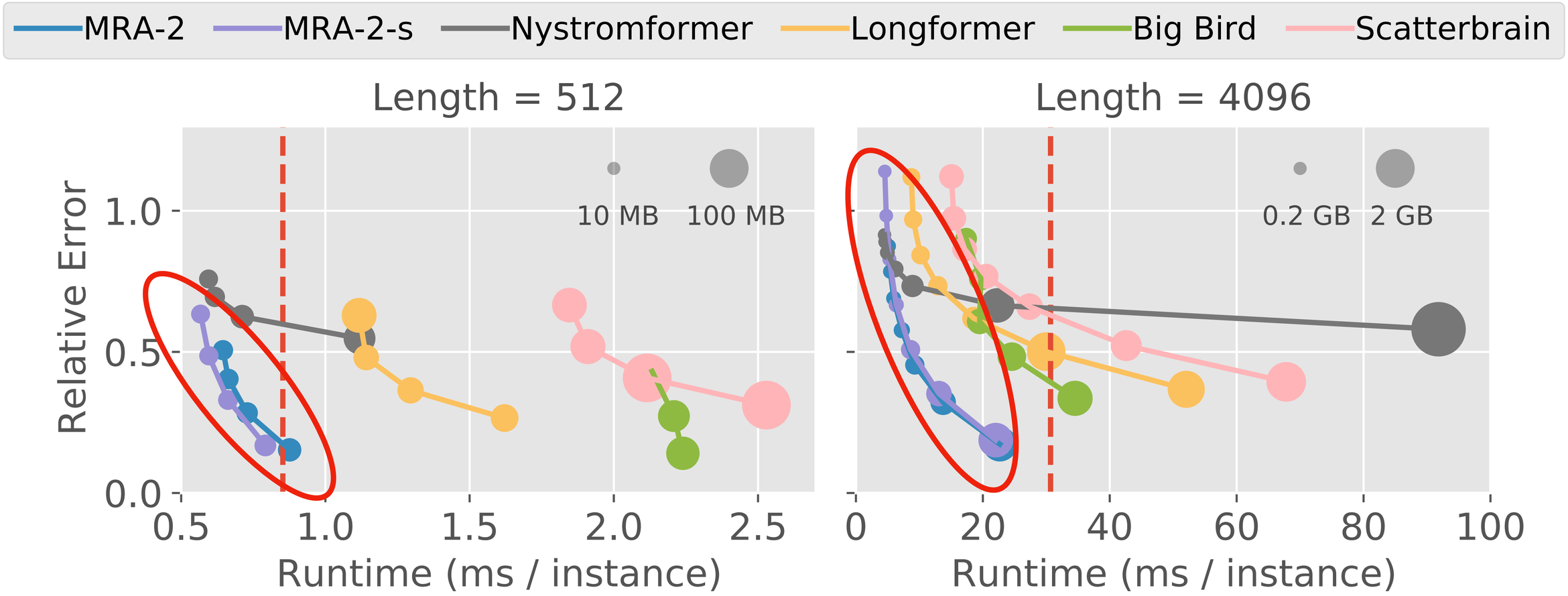
Evaluations

Efficiency v.s. Approximation



Evaluations

Efficiency v.s. Approximation

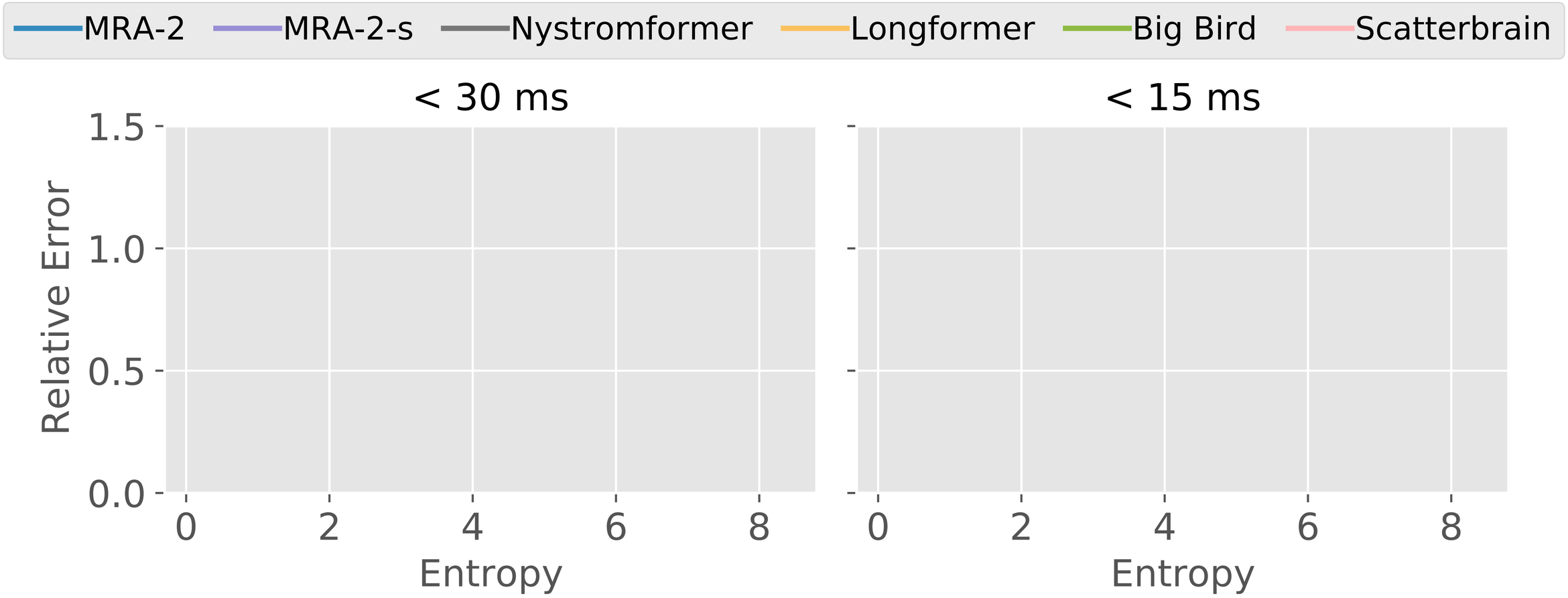


Evaluations

Entropy versus Approximation

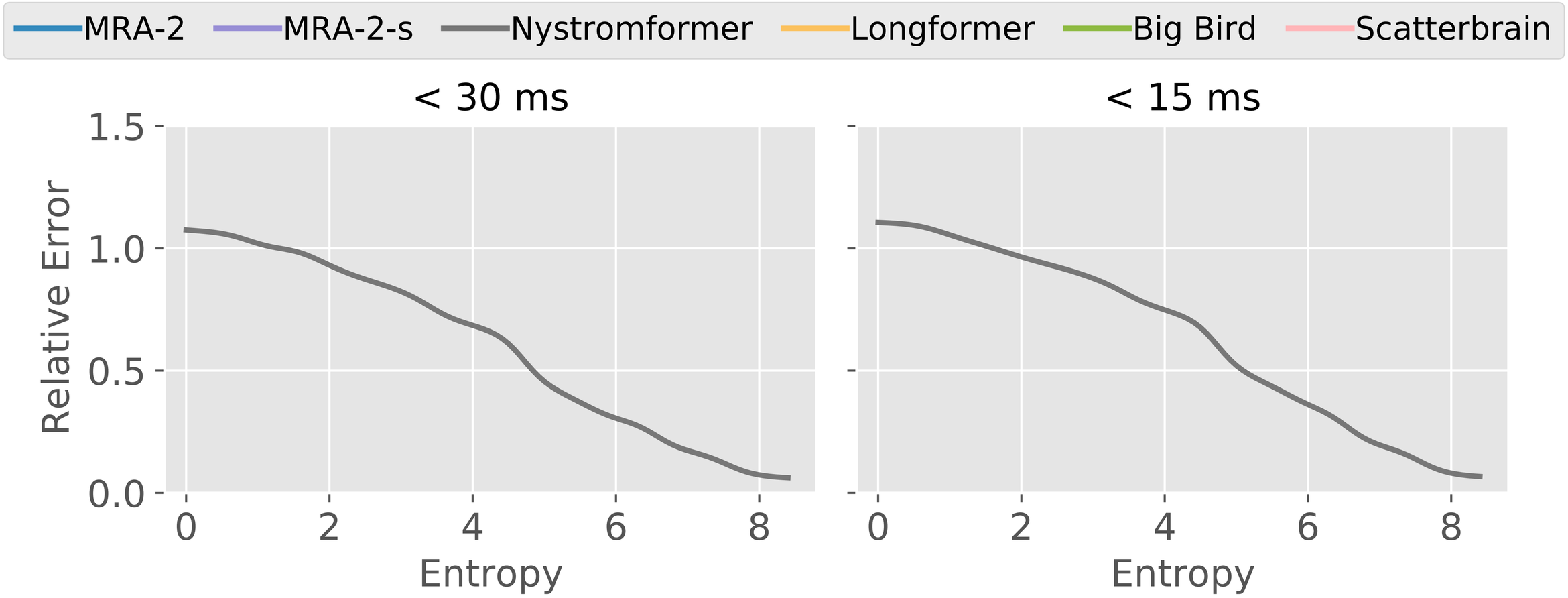
Evaluations

Entropy v.s. Approximation



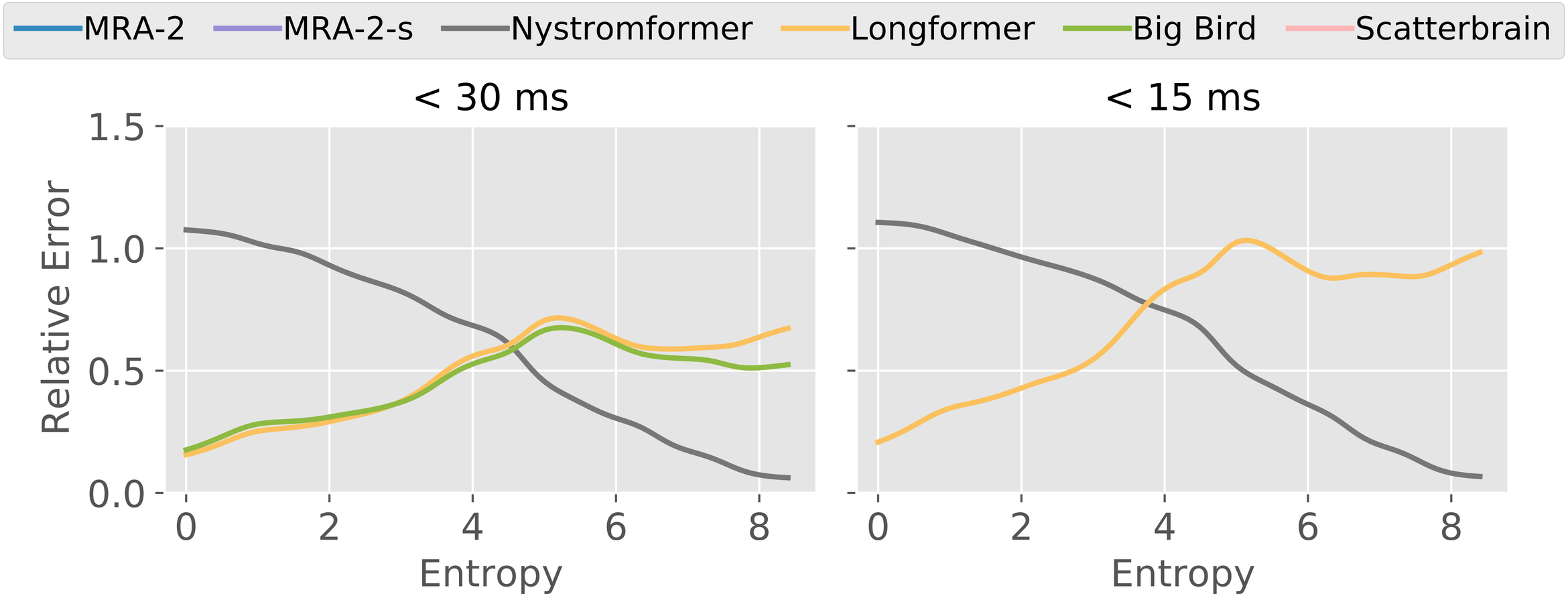
Evaluations

Entropy v.s. Approximation



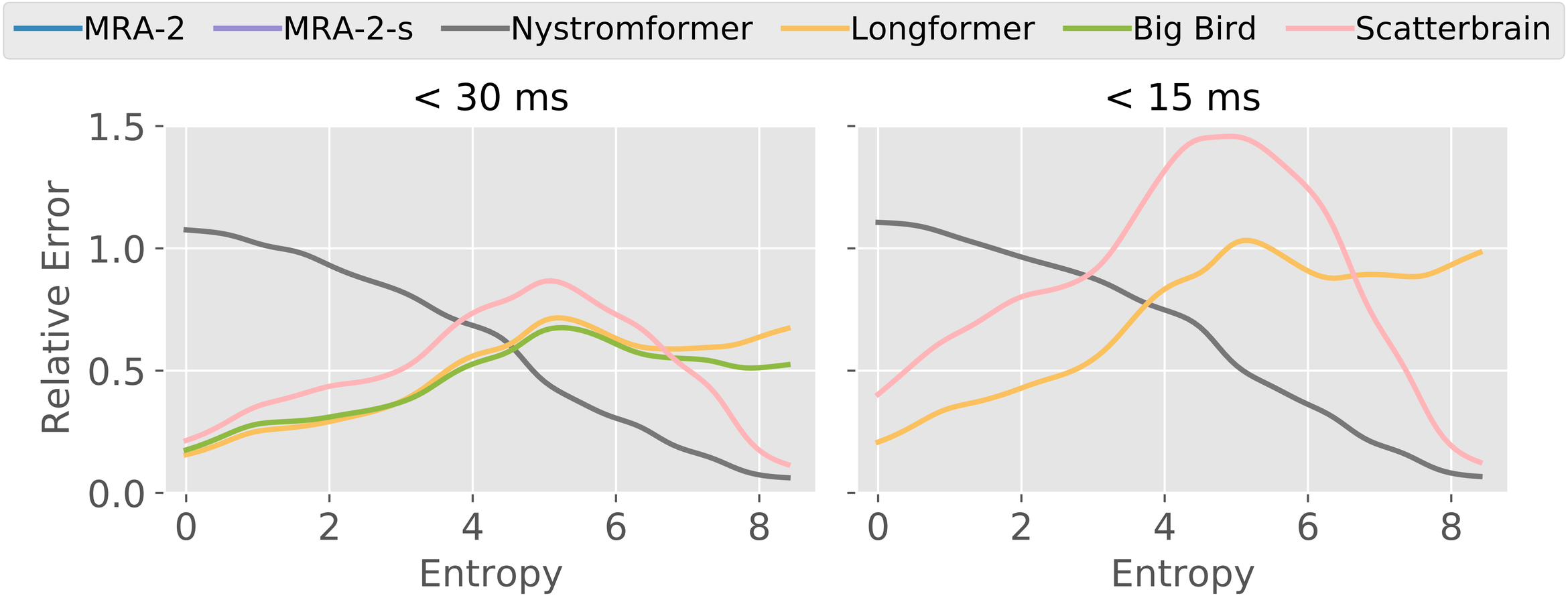
Evaluations

Entropy v.s. Approximation



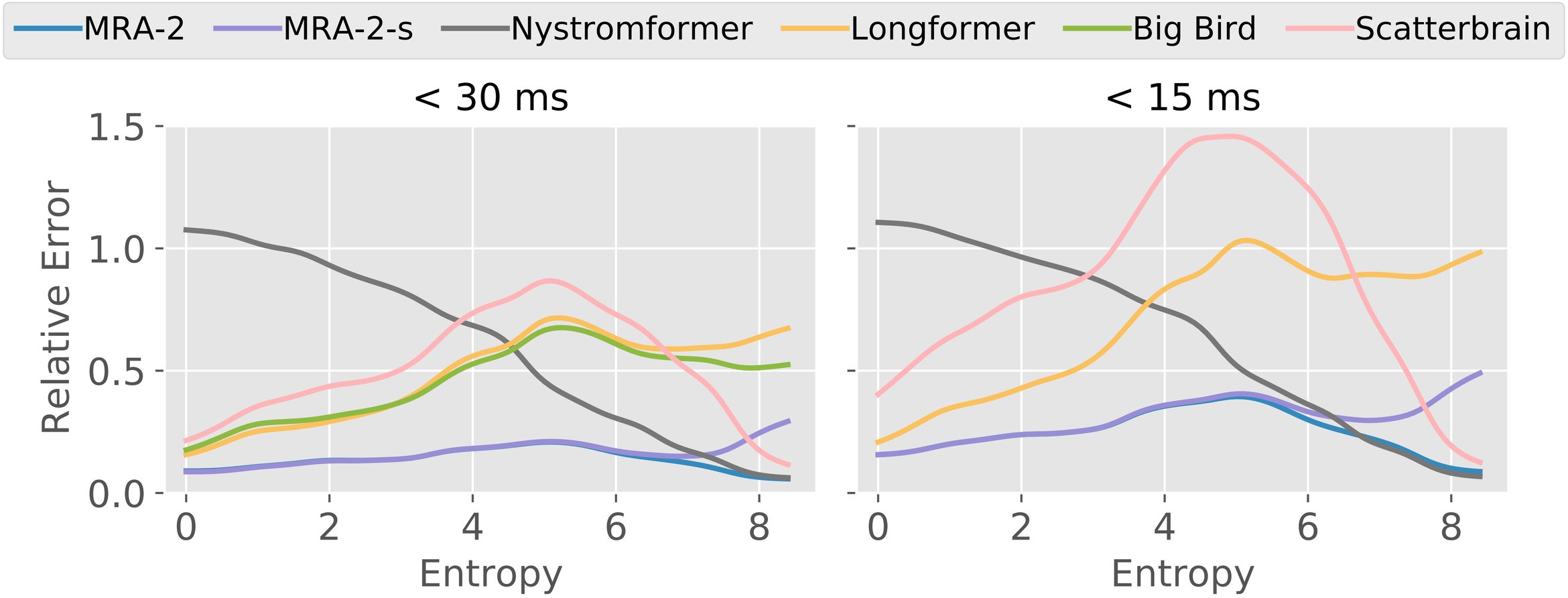
Evaluations

Entropy v.s. Approximation



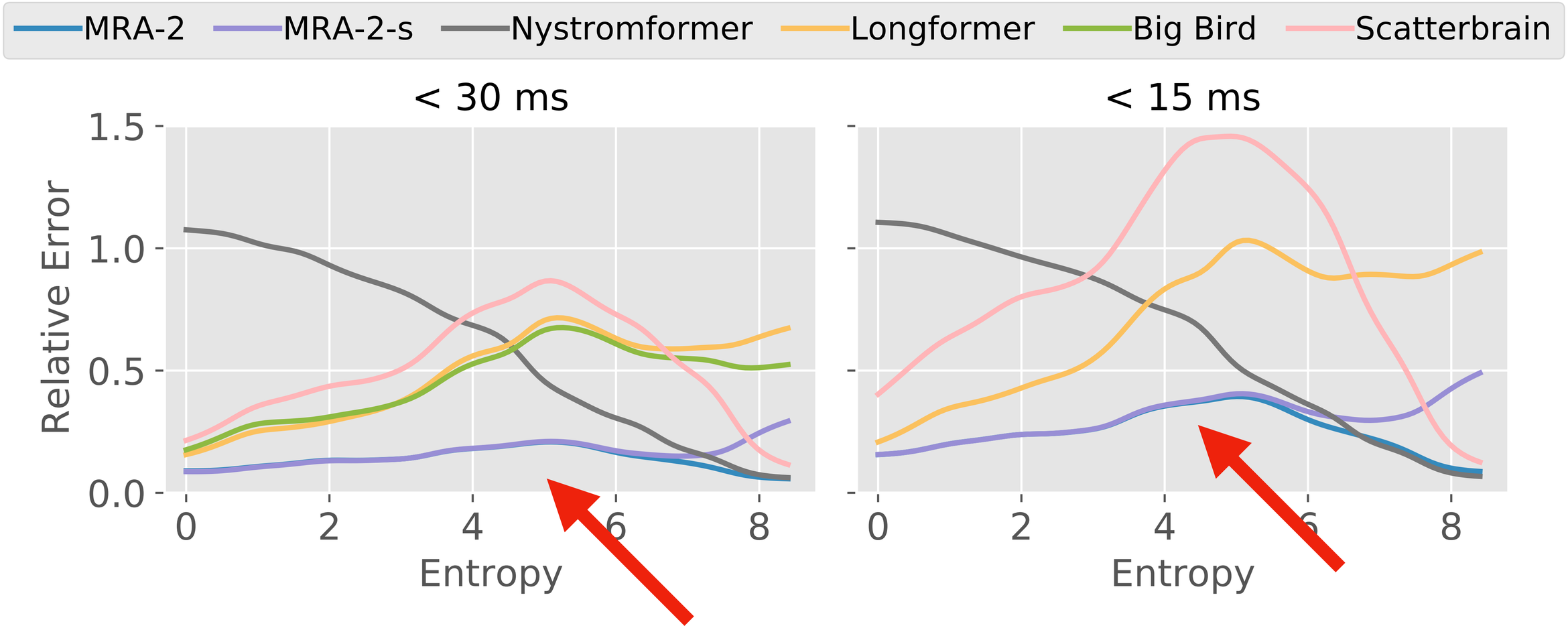
Evaluations

Entropy v.s. Approximation



Evaluations

Entropy v.s. Approximation

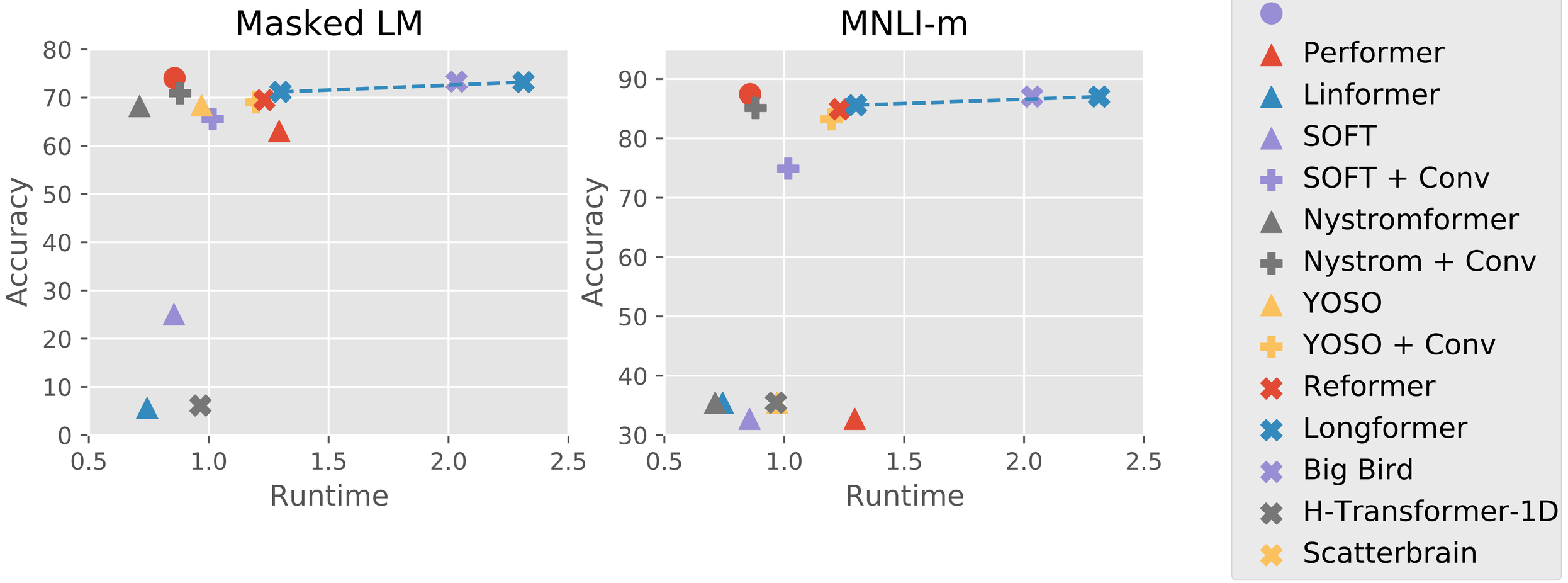


Evaluations

RoBERTa-base-512

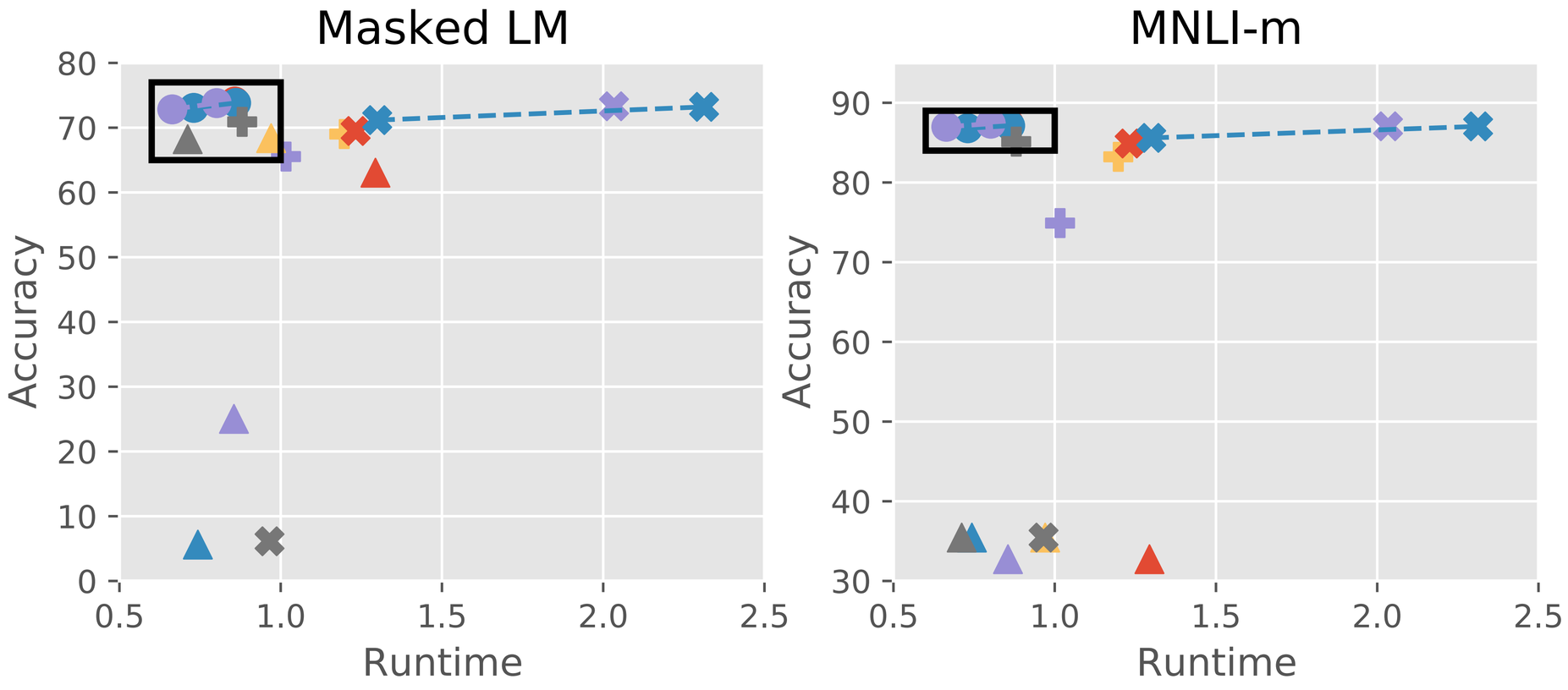
Evaluations

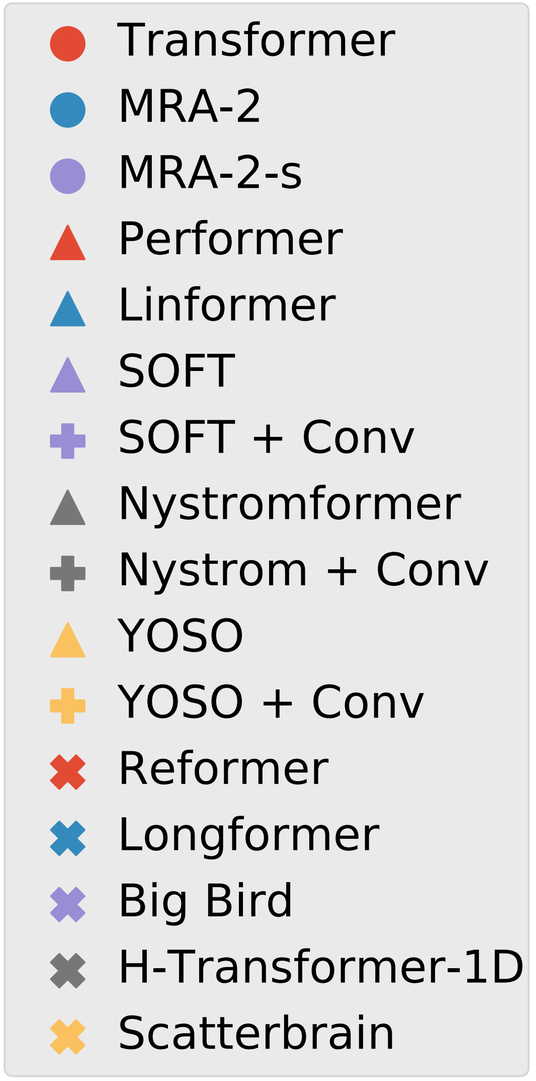
RoBERTa-base-512



Evaluations

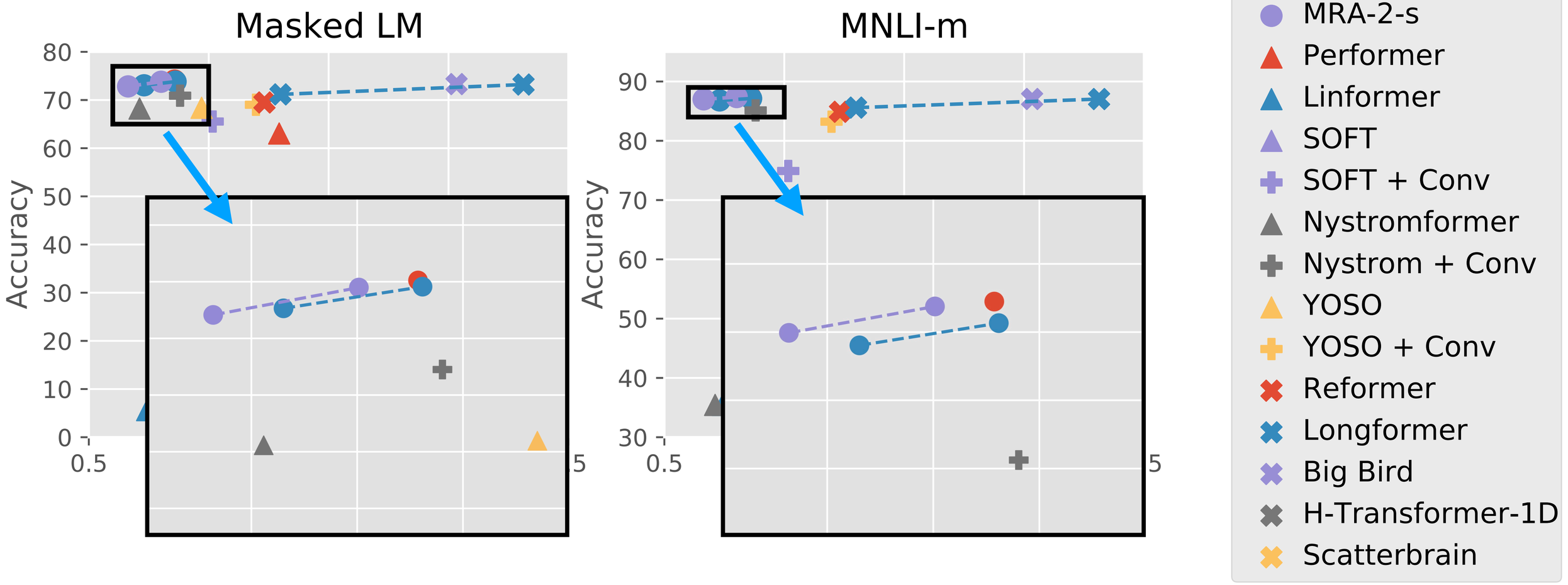
RoBERTa-base-512





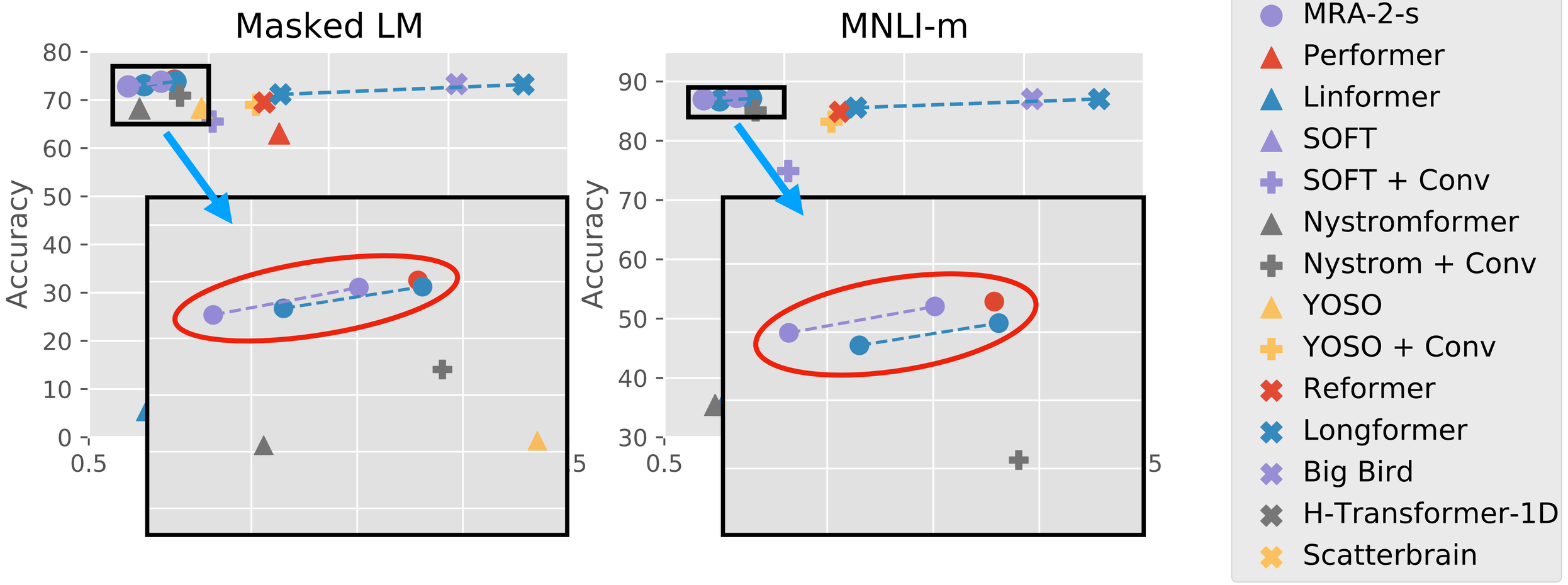
Evaluations

RoBERTa-base-512



Evaluations

RoBERTa-base-512

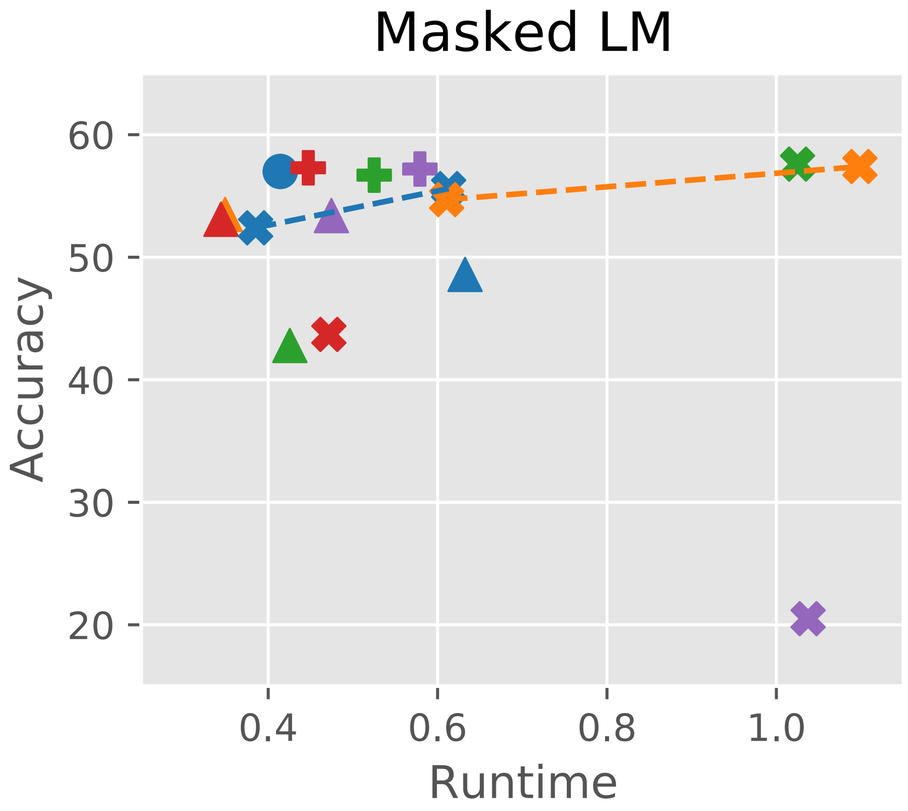


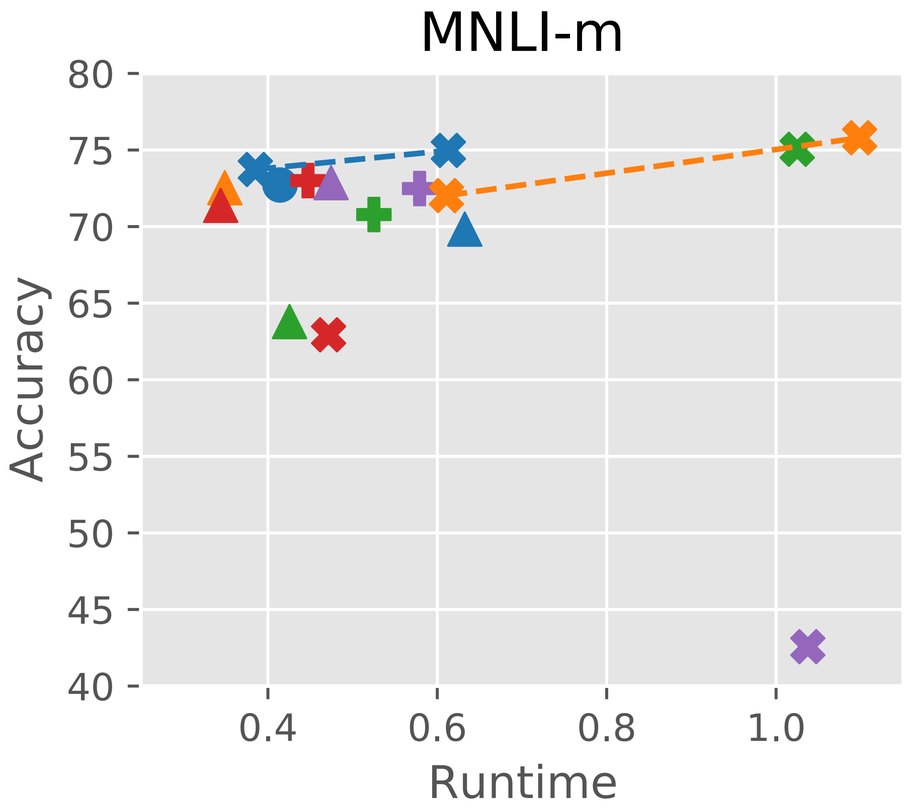
Evaluations

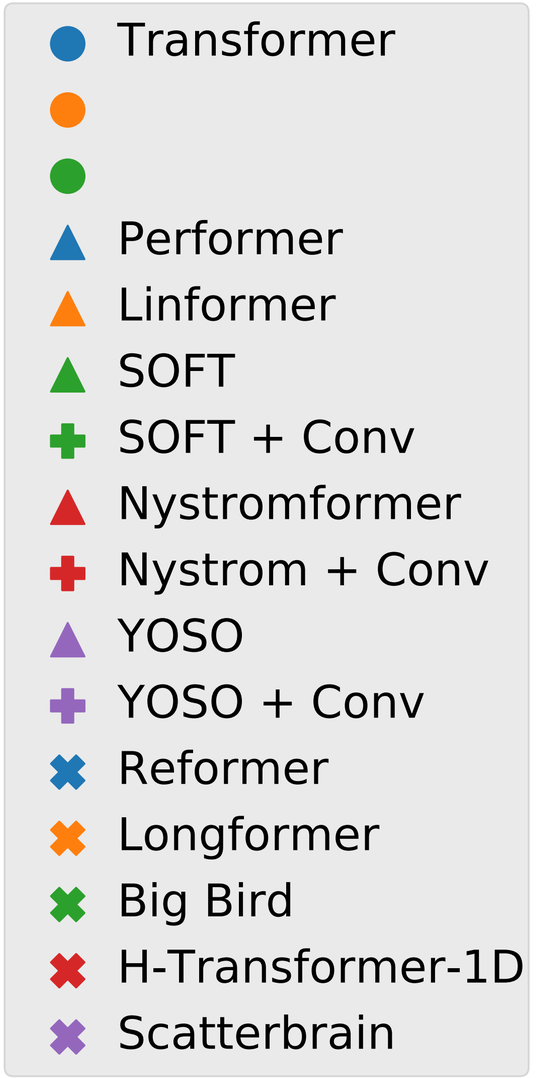
RoBERTa-small-512

Evaluations

RoBERTa-small-512

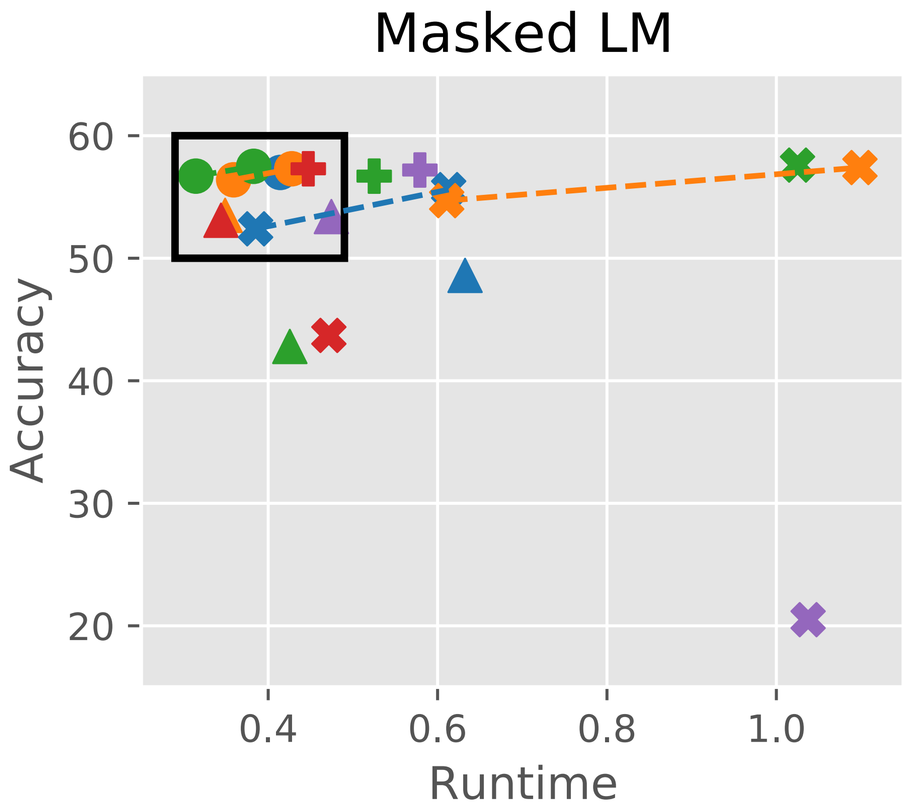


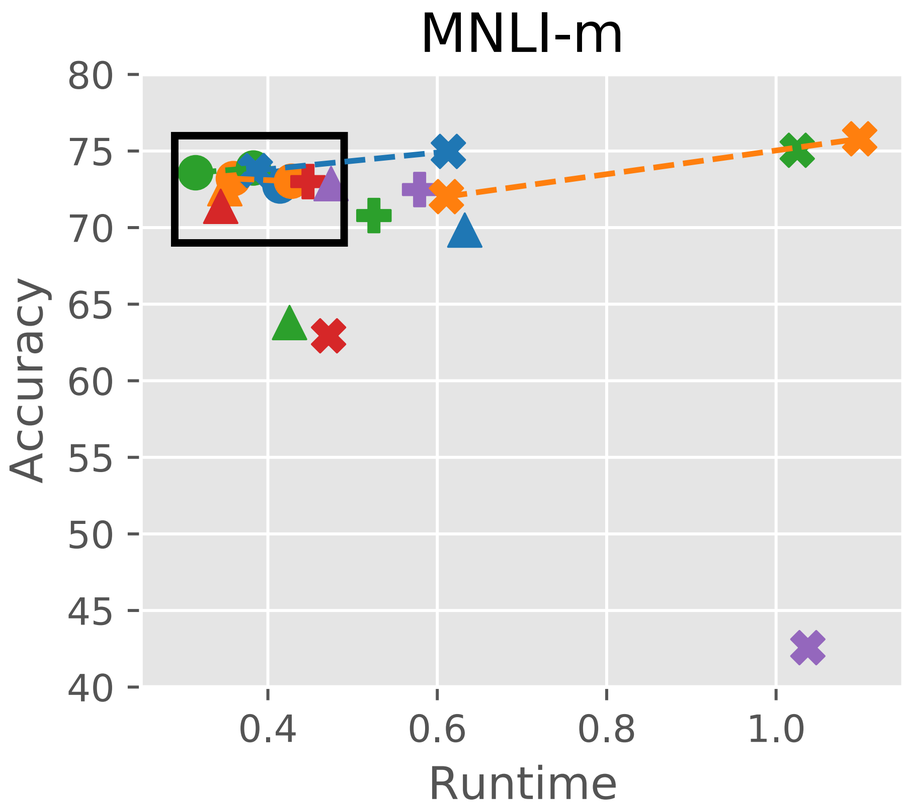


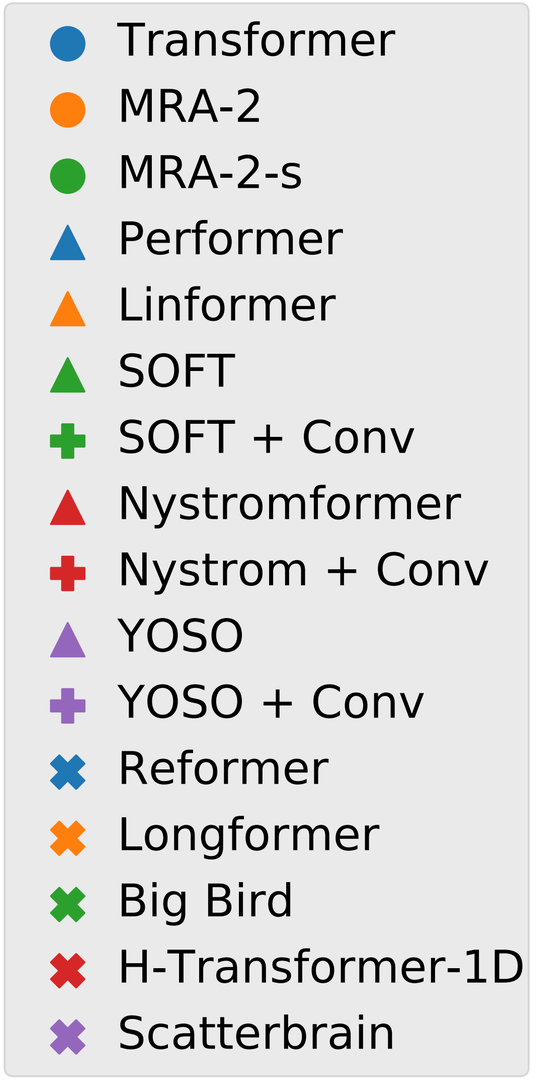


Evaluations

RoBERTa-small-512

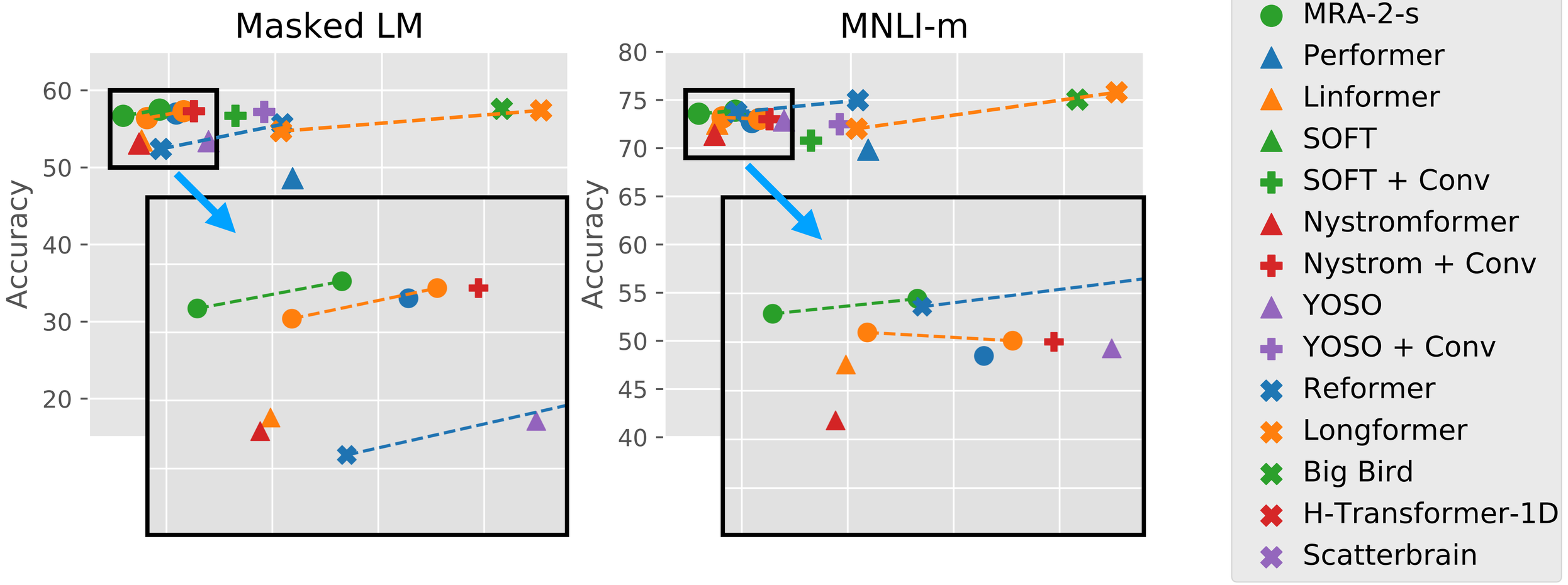






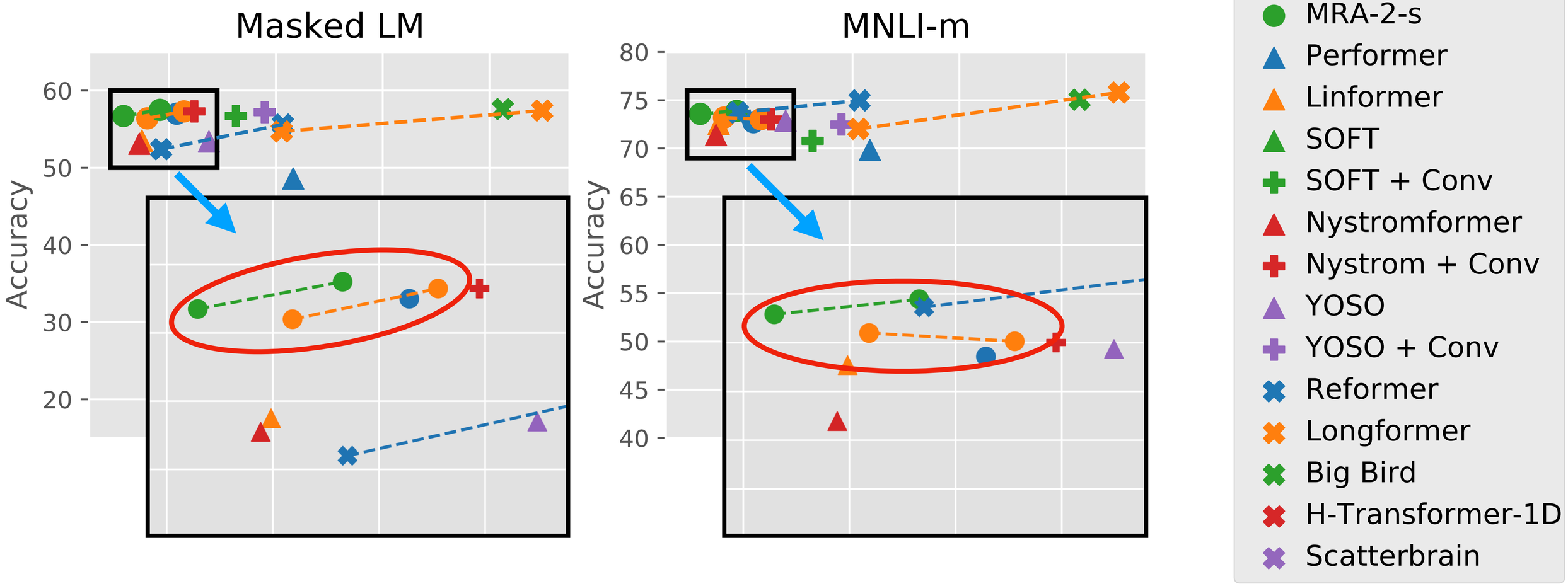
Evaluations

RoBERTa-small-512



Evaluations

RoBERTa-small-512

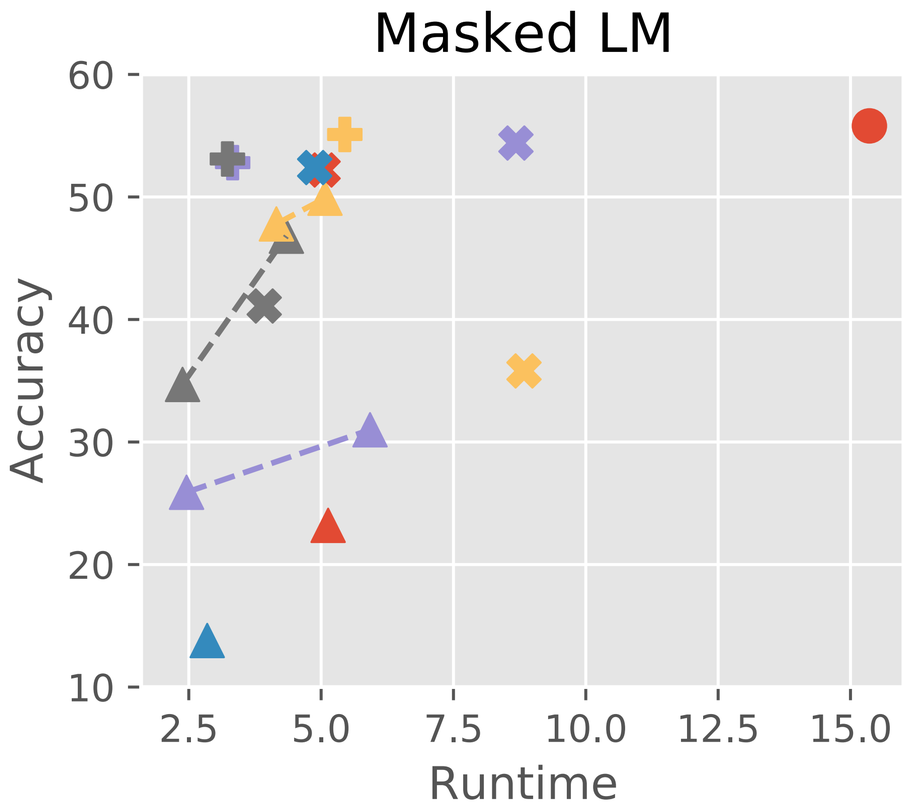


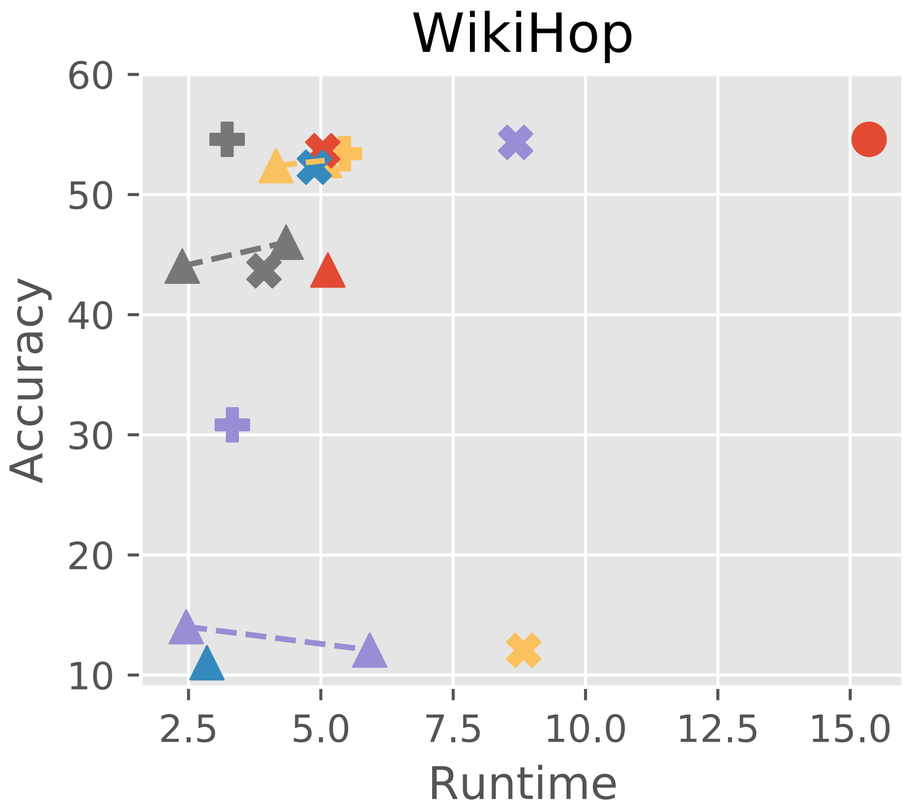
Evaluations

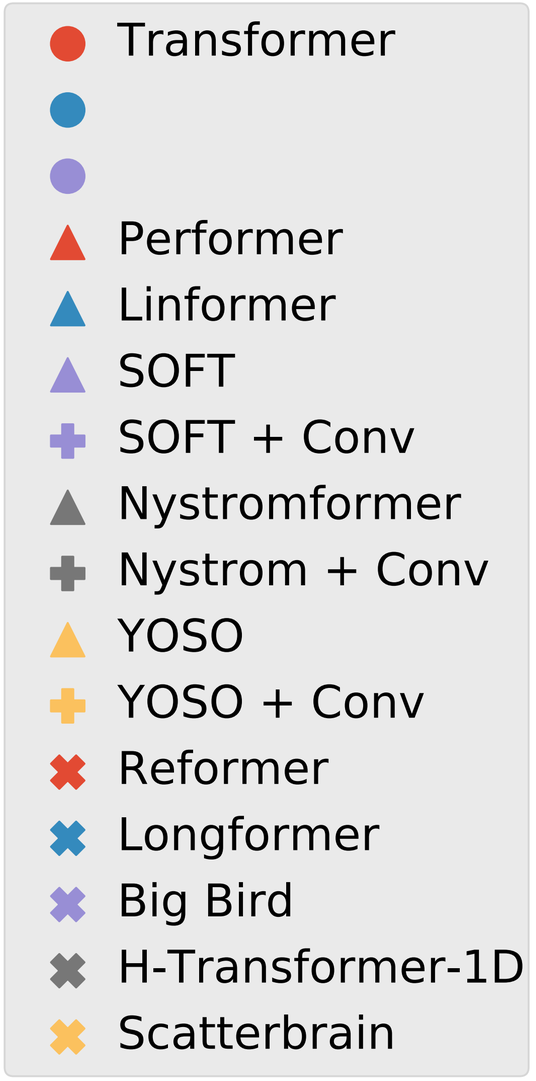
RoBERTa-small-4096

Evaluations

RoBERTa-small-4096

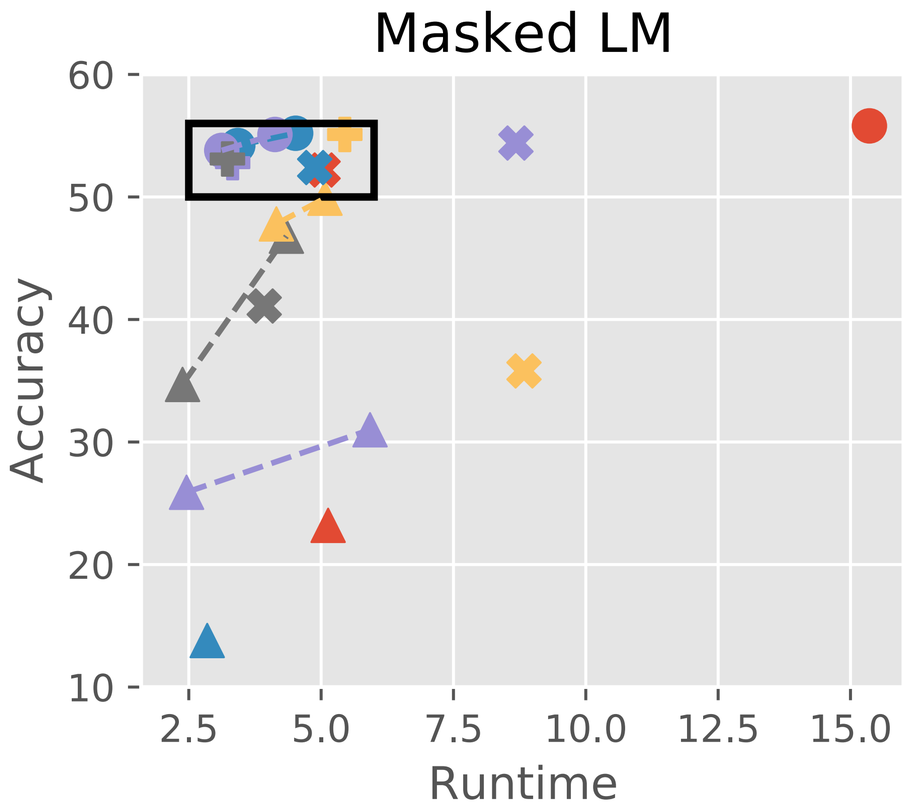


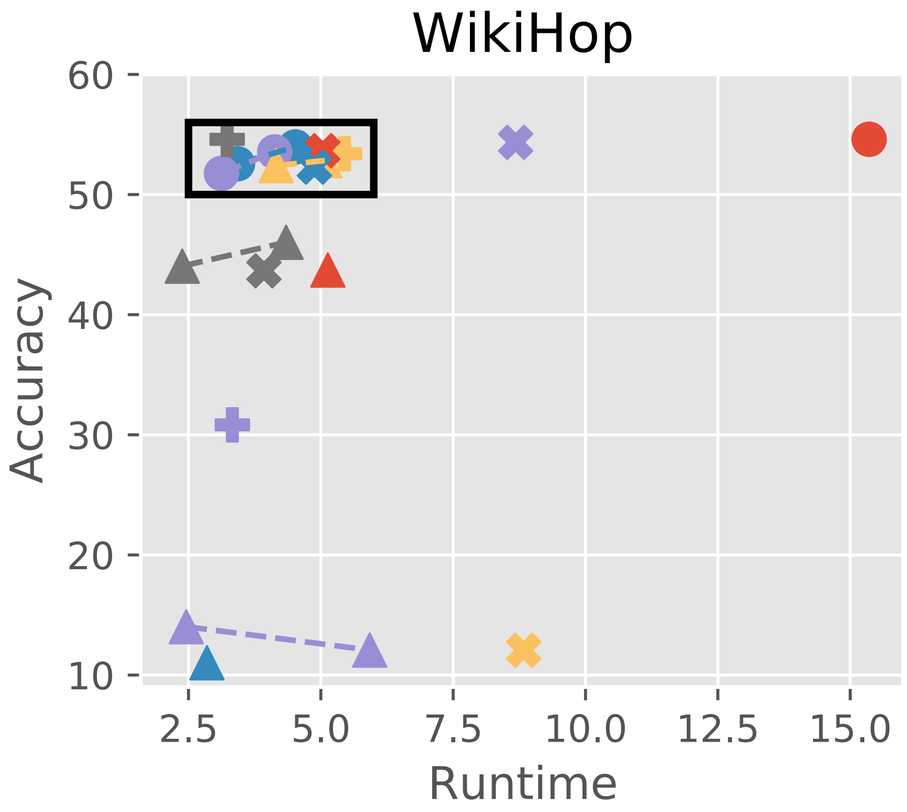


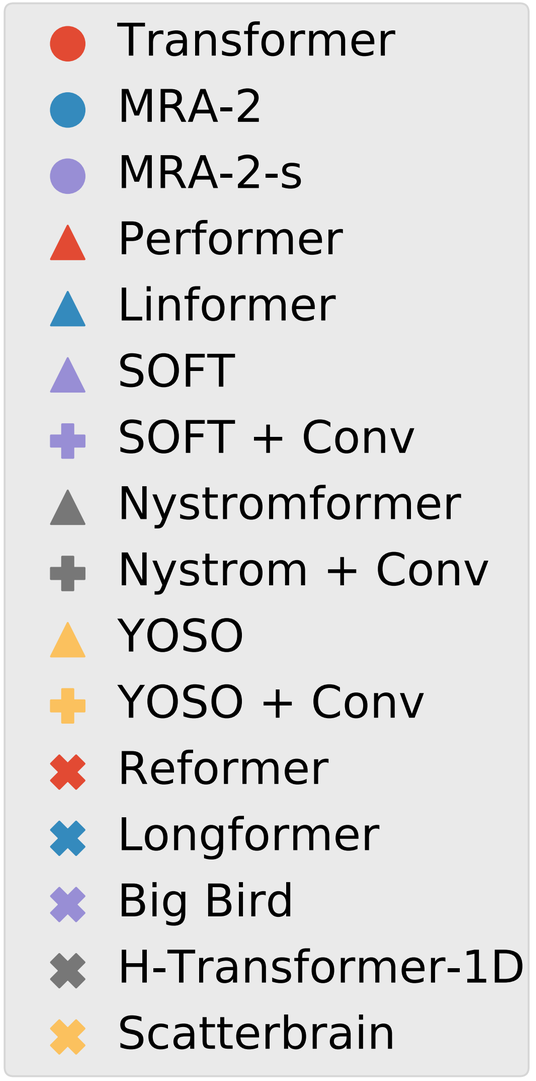


Evaluations

RoBERTa-small-4096

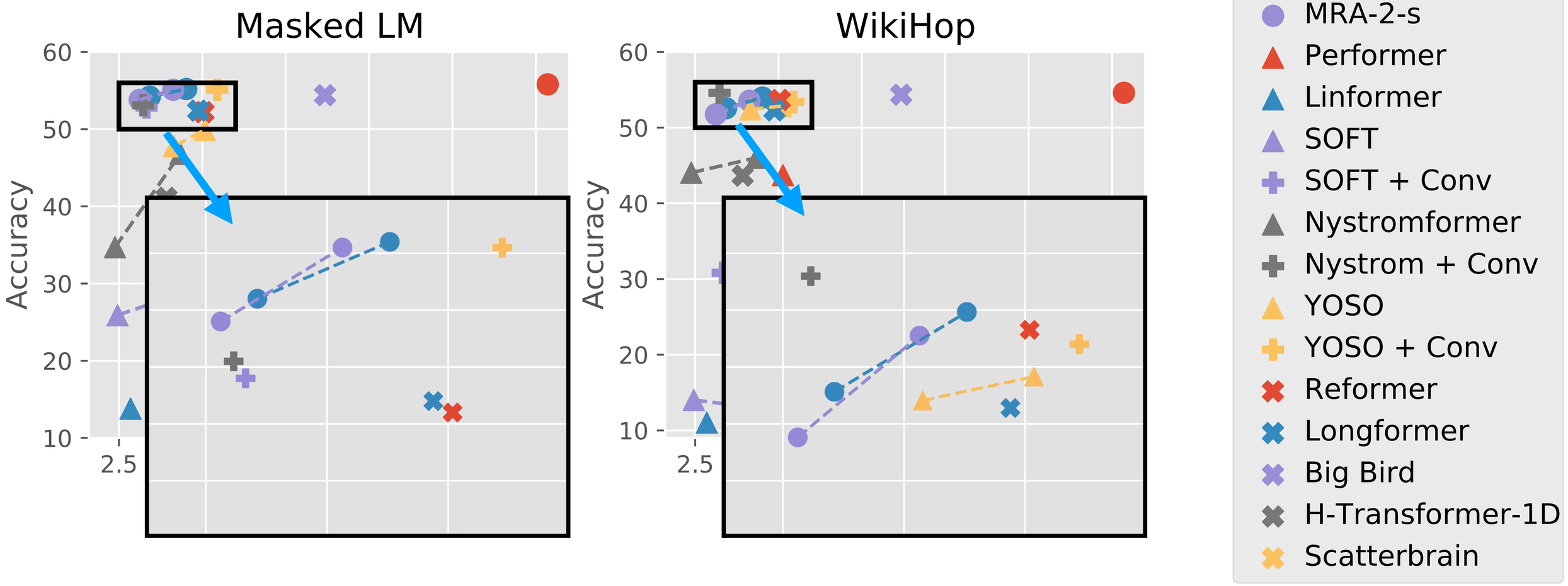






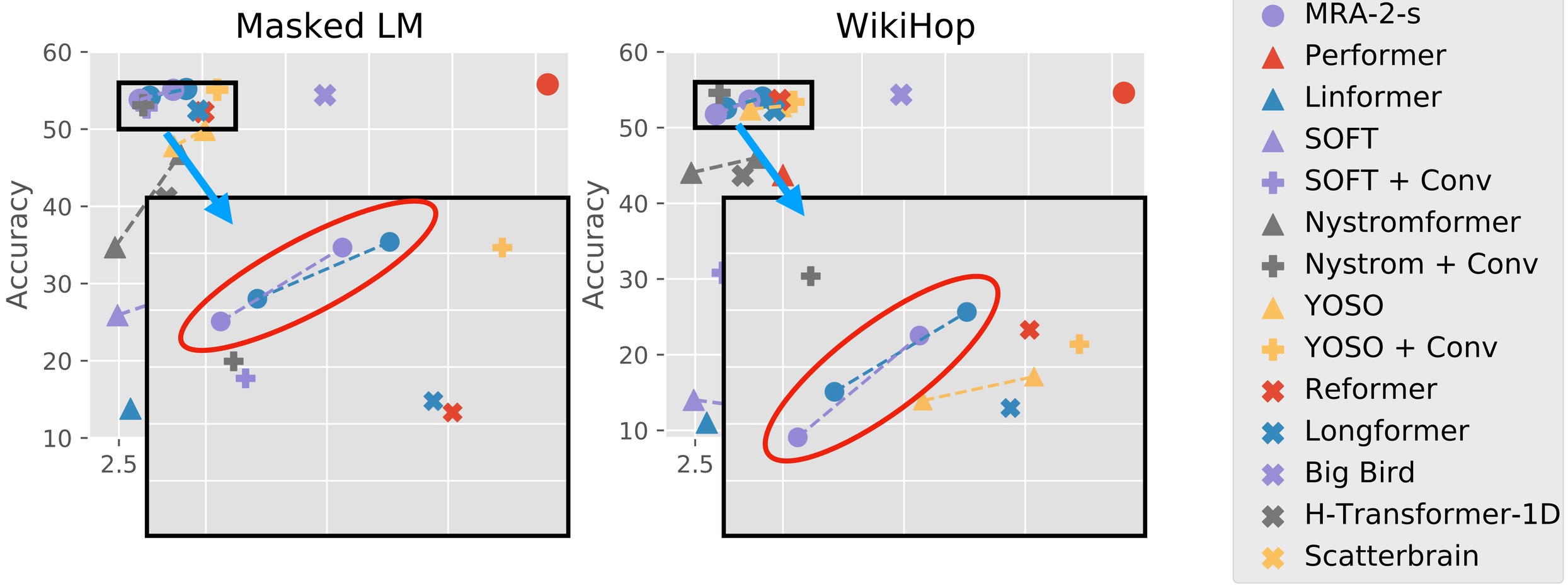
Evaluations

RoBERTa-small-4096



Evaluations

RoBERTa-small-4096

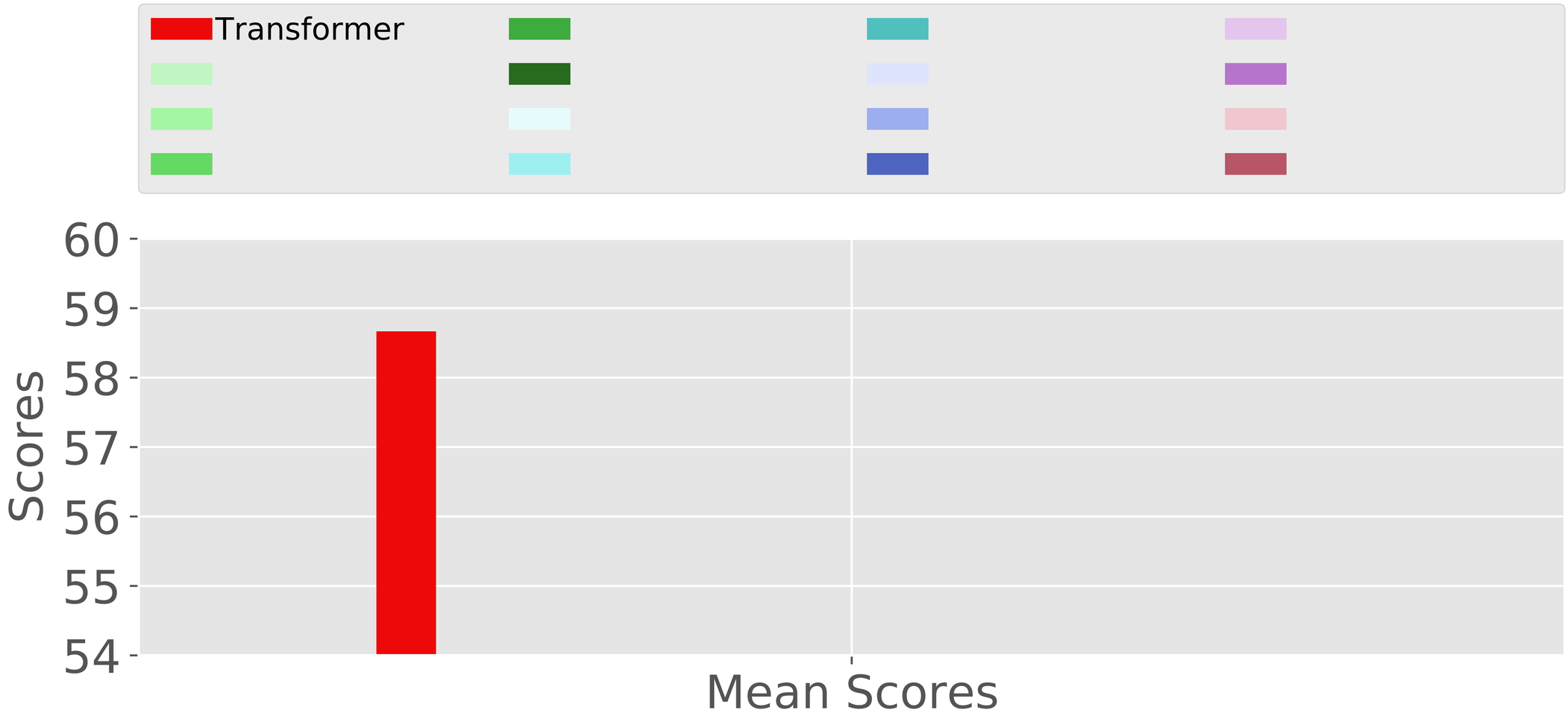


Evaluations

Long Range Arena Benchmark

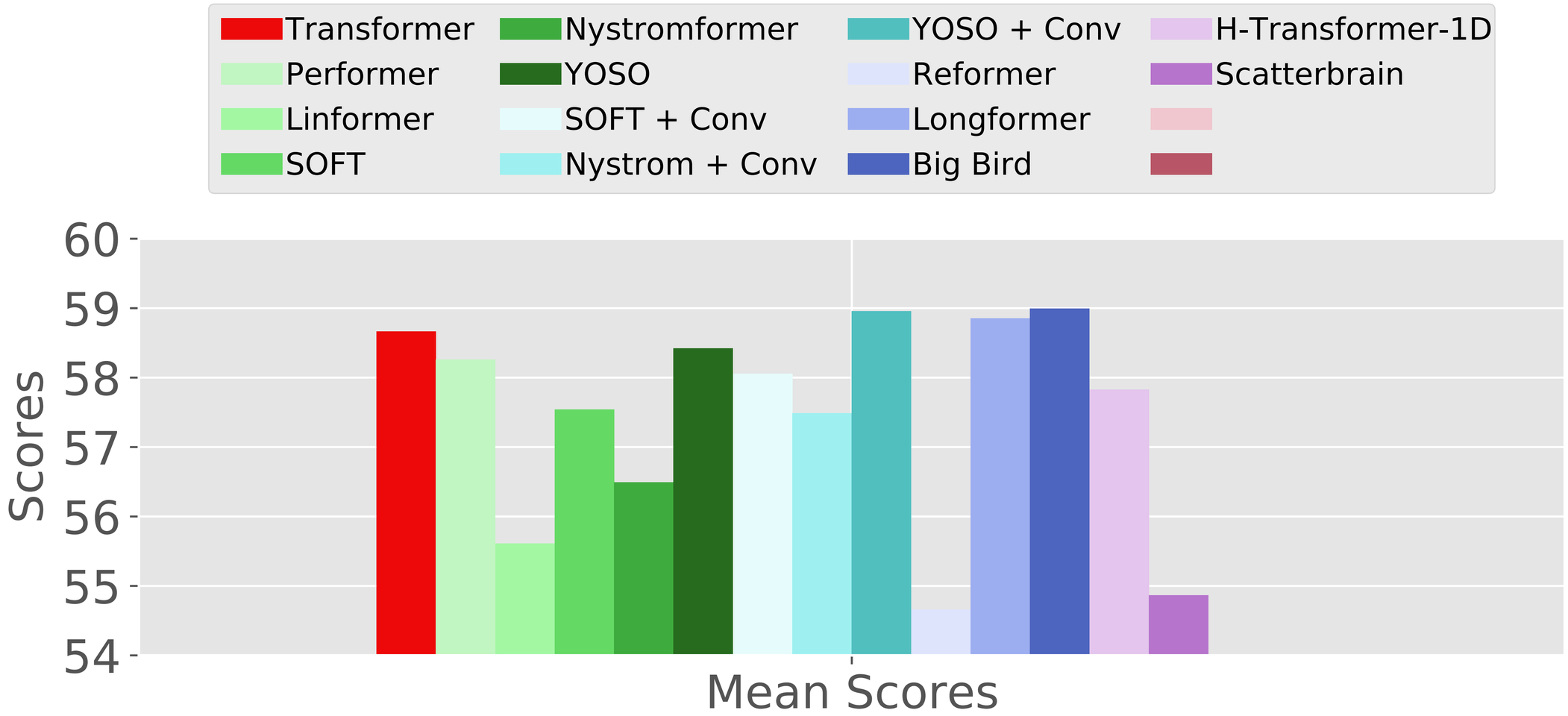
Evaluations

Long Range Arena Benchmark



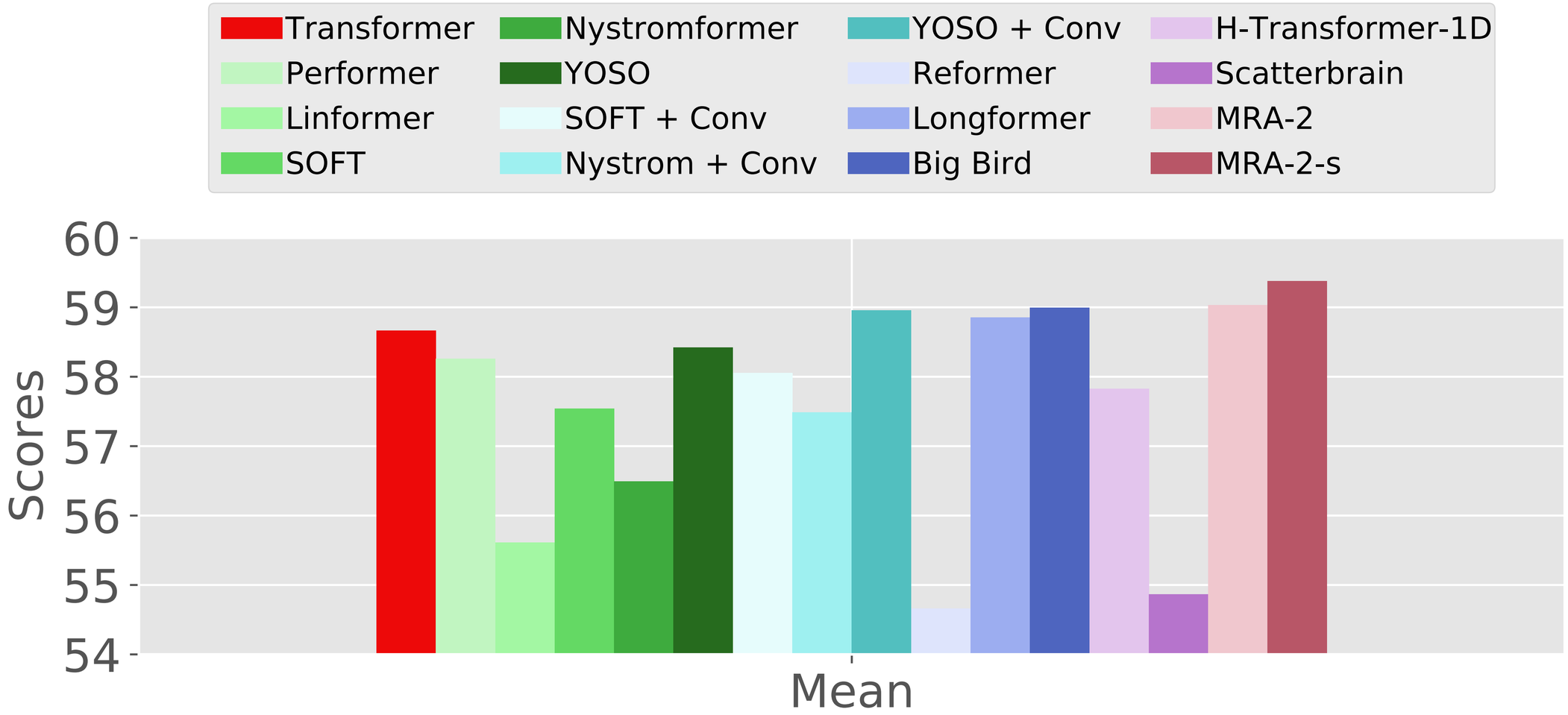
Evaluations

Long Range Arena Benchmark



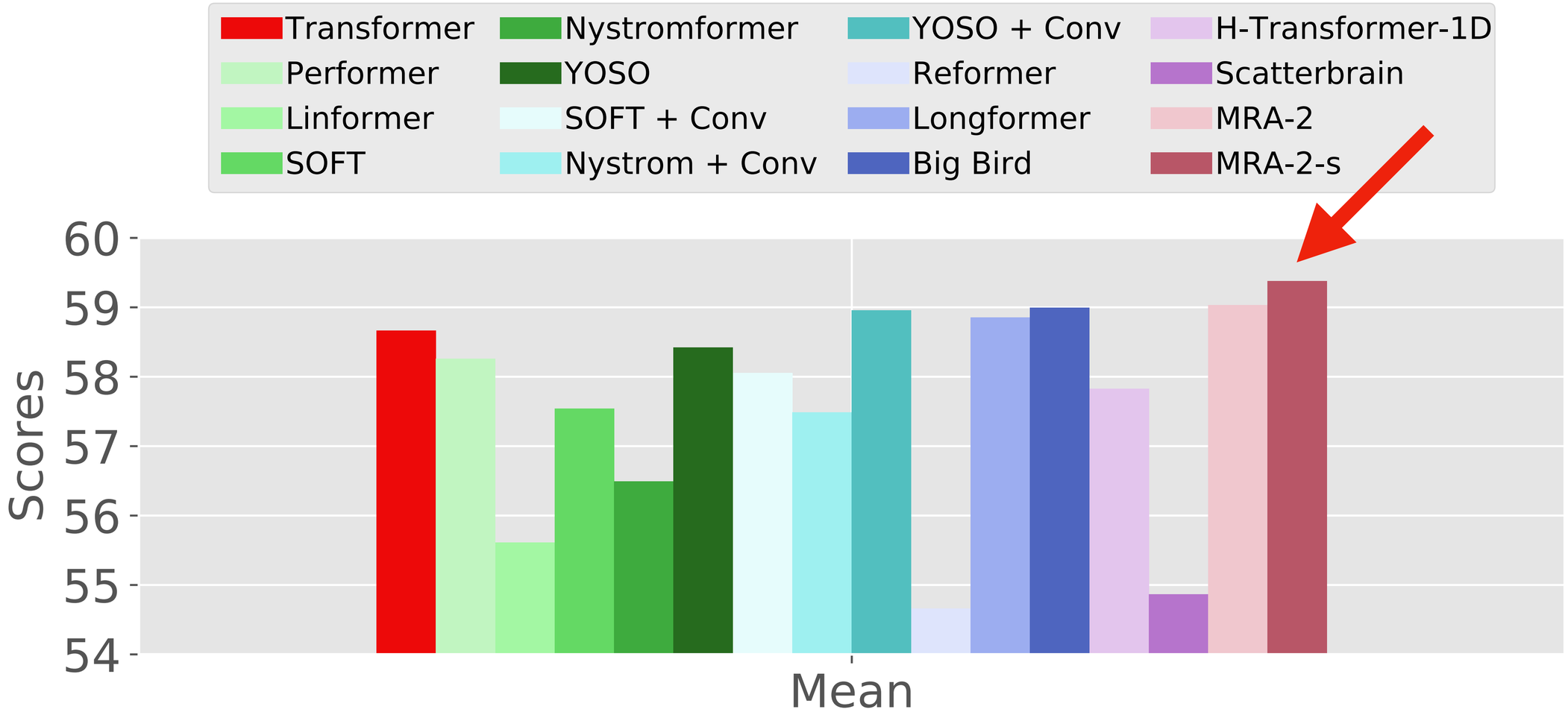
Evaluations

Long Range Arena Benchmark



Evaluations

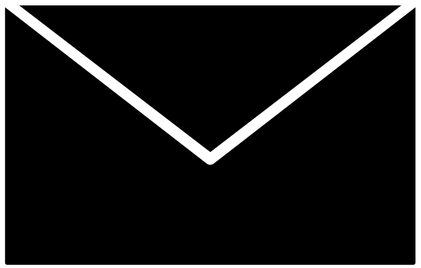
Long Range Arena Benchmark



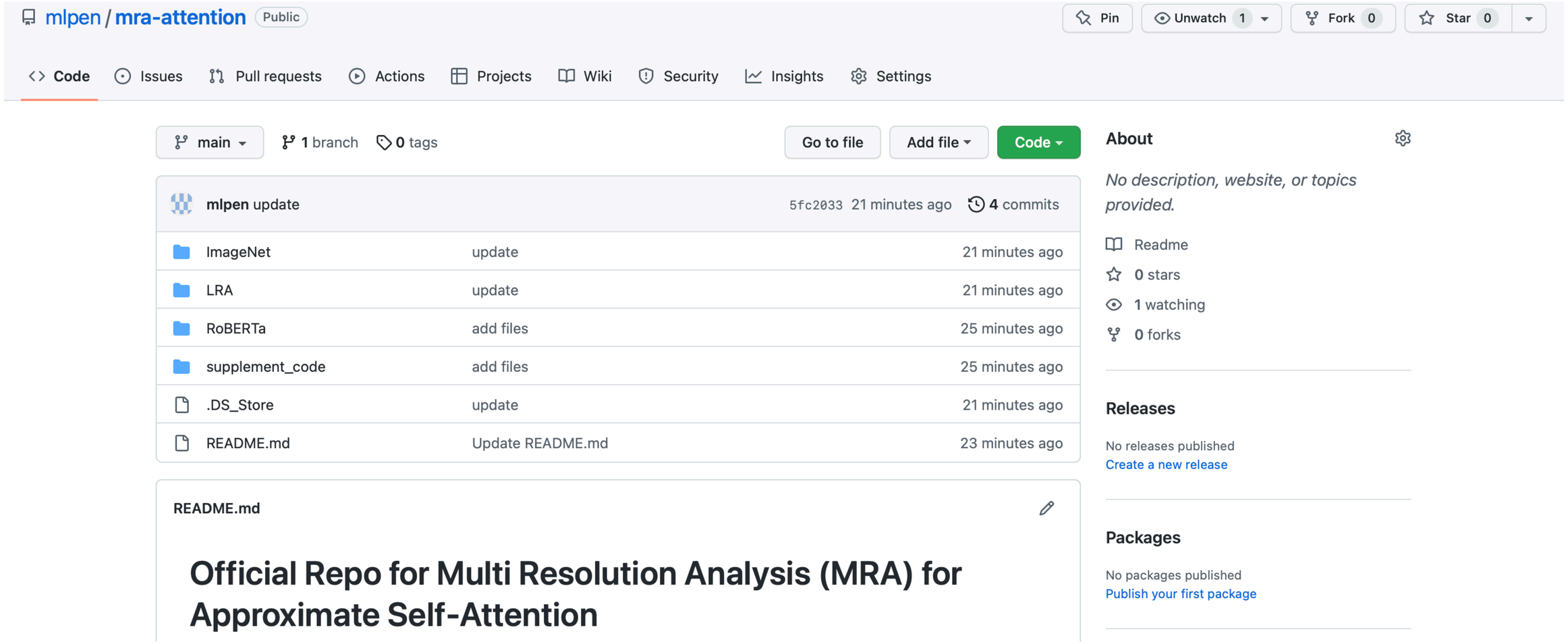
Takeaway

- Multi-resolution analysis (MRA) ideas offer a win-win for efficient self-attention
  - potential access to a rich MRA theory
  - immediate practical benefits on efficiency and accuracy

mlpen/mra-attention



zzeng38@wisc.edu



https://github.com/mlpen/mra-attention