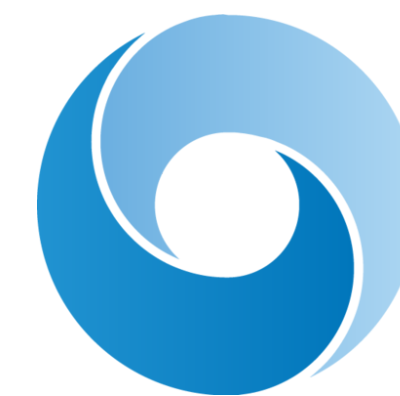


Random Feature Attention

Hao Peng, Nikos Pappas, Dani Yogatama, Roy Schwartz,
Noah Smith, Lingpeng Kong
@ICLR 2021



DeepMind

Transformers

State-of-the-art results in many sequence modeling tasks

- Machine translation ([Vaswani et al., 2017](#))
- Language modeling ([Ott et al., 2018](#))
- Pretraining ([DeLvin et al., 2019](#))

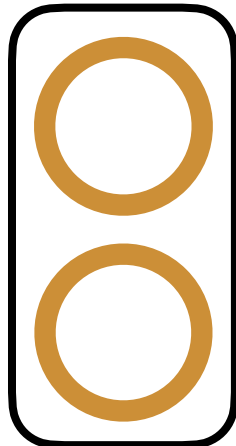
Transformers

State-of-the-art results in many sequence modeling tasks

- Machine translation ([Vaswani et al., 2017](#))
- Language modeling ([Ott et al., 2018](#))
- Pretraining ([DeLvin et al., 2019](#))
- Reinforcement learning ([Parisotto et al., 2019](#))
- Computer vision ([Parmar et al., 2018](#); [Dosovitskiy et al., 2020](#))
- Computational biology ([Choromanski et al., 2021](#))
- ...

Attention

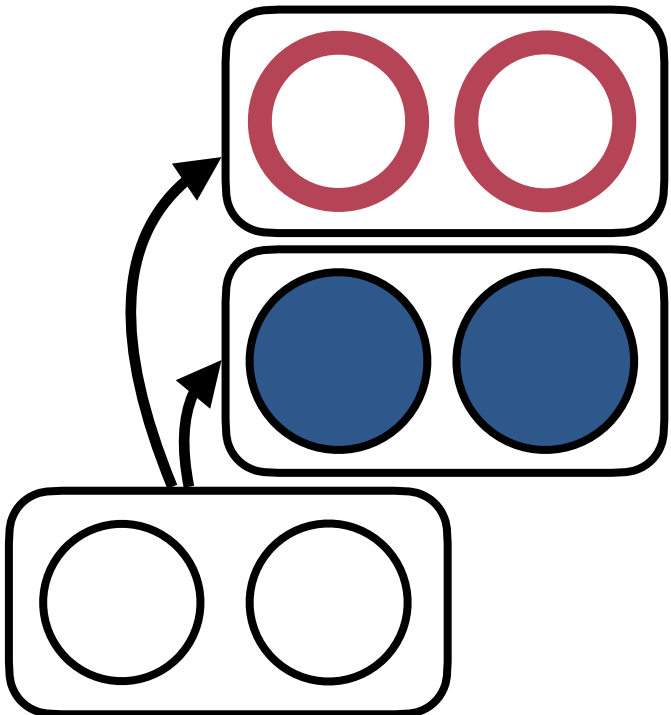
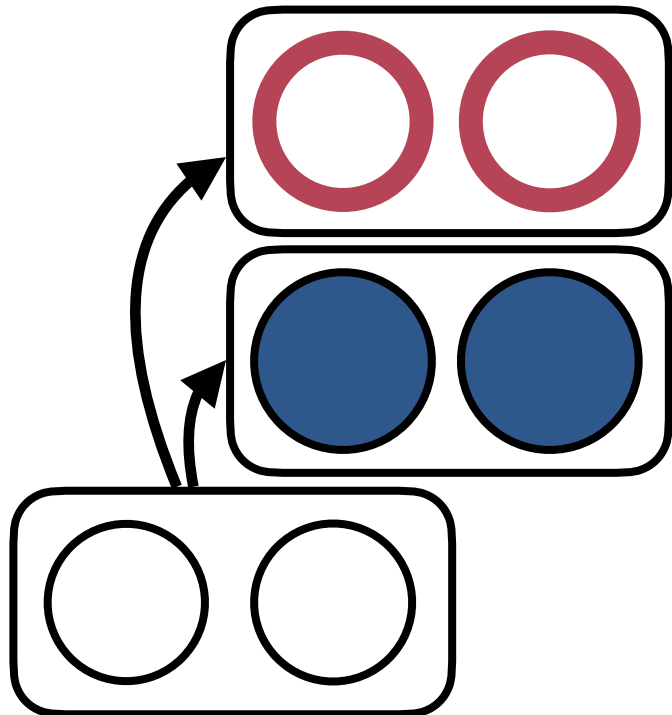
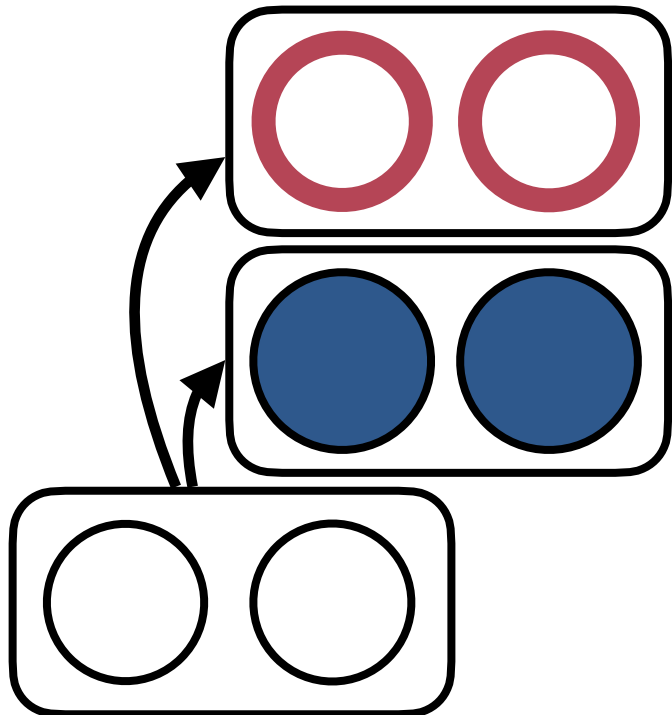
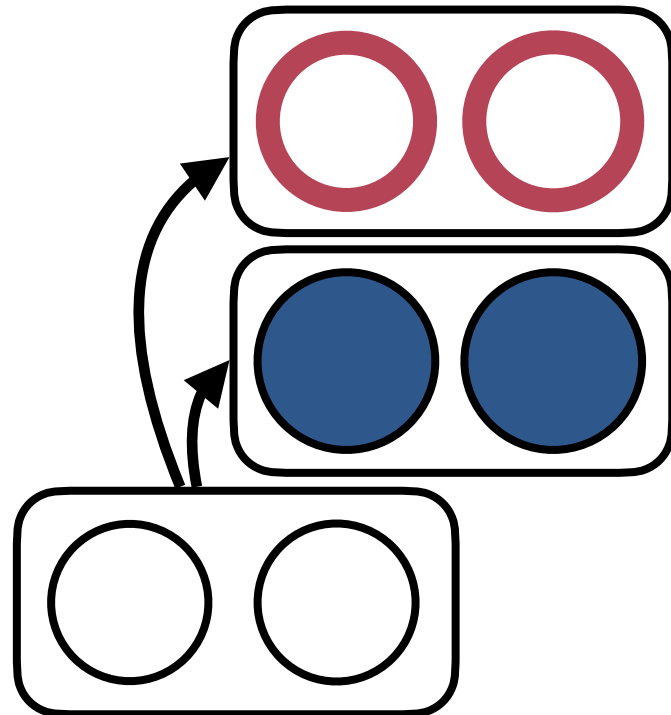
query



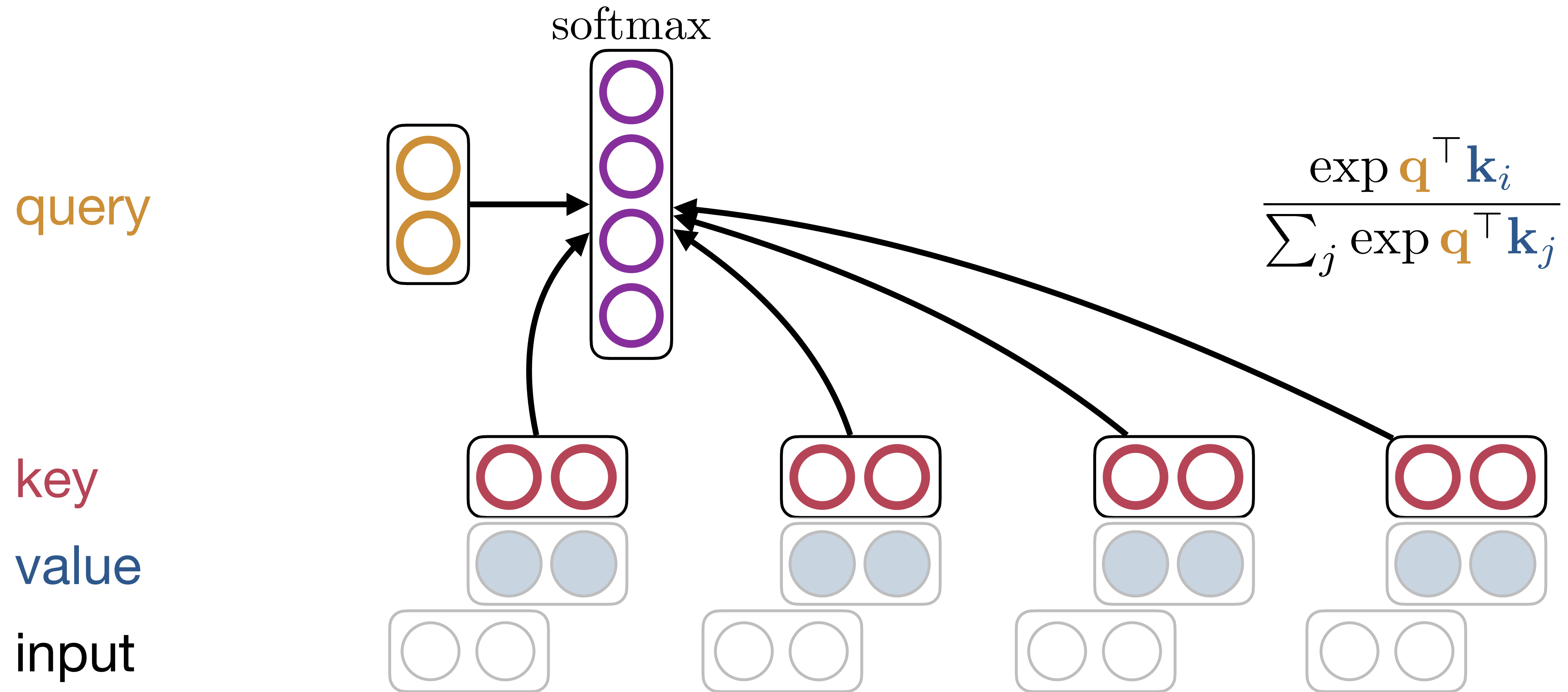
key

value

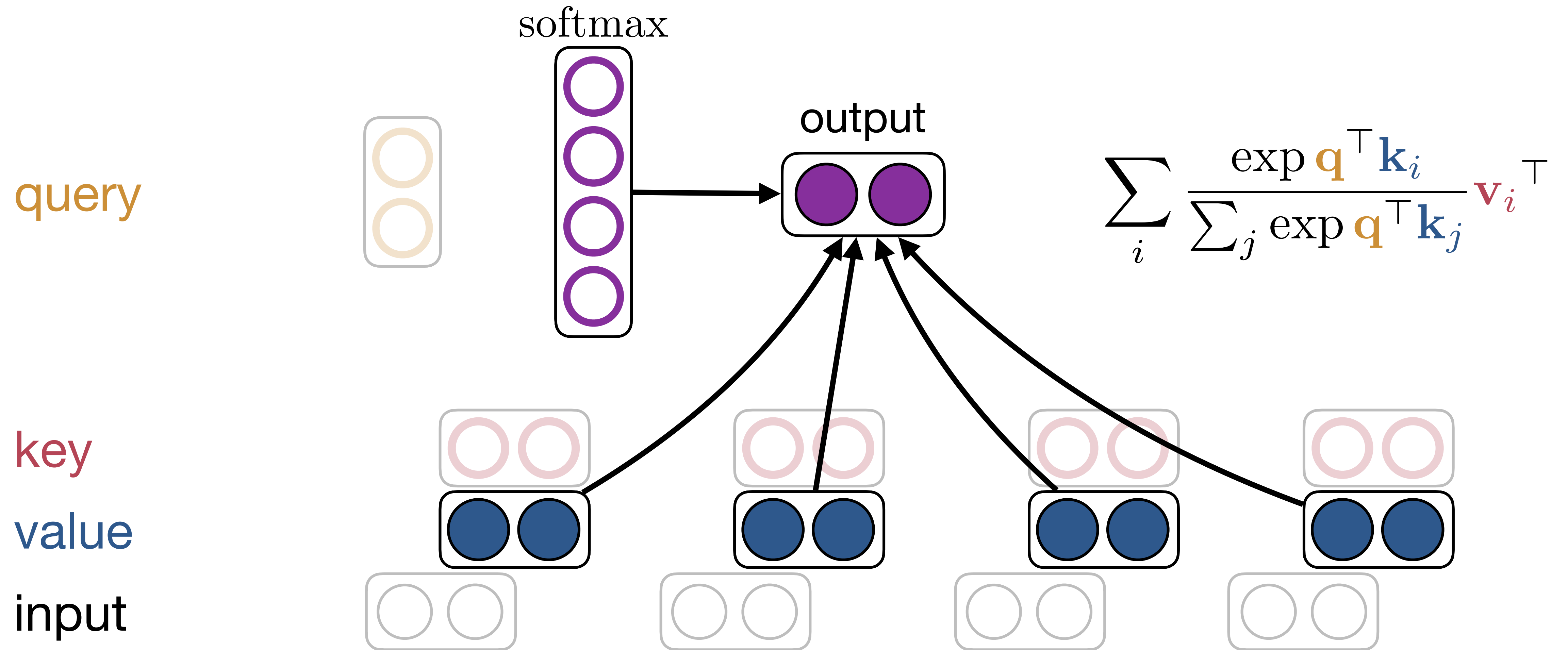
input



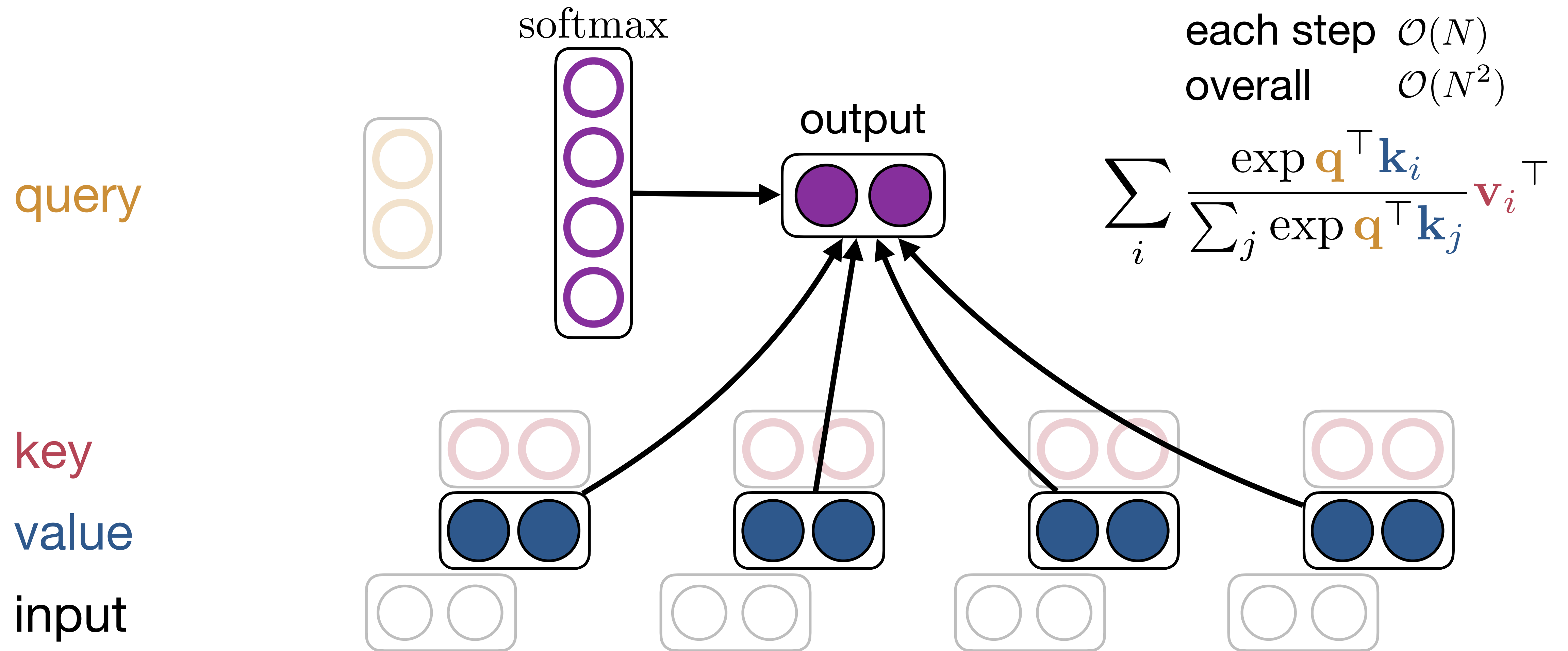
Attention



Attention



Attention



Overview

Transformers: quadratic overhead, limited in

- Character-level language modeling
- Document-level machine translation
- Speech
- ...

Overview

Transformers

- State-of-the-art results in many sequence modeling tasks
- Quadratic complexity, less well-suited for long sequences



This Work: Random Feature Attention

- Strong performance
- Scales linearly in sequence length



From Attention to Random Feature Attention

$$\sum_i \frac{\exp \mathbf{q}^\top \mathbf{k}_i}{\sum_j \exp \mathbf{q}^\top \mathbf{k}_j} \mathbf{v}_i^\top$$

$\mathbf{q}$ query

$\mathbf{k}_i$ keys

$\mathbf{v}_i$ values

From Attention to Random Feature Attention

$\mathbf{q}$ query
 $\mathbf{k}_i$ keys
 $\mathbf{v}_i$ values

$$\sum_i \frac{\exp \mathbf{q}^\top \mathbf{k}_i}{\sum_j \exp \mathbf{q}^\top \mathbf{k}_j} \mathbf{v}_i^\top$$
$$\approx \sum_i \frac{\phi(\mathbf{q})^\top \phi(\mathbf{k}_i) \otimes \mathbf{v}_i}{\sum_j \phi(\mathbf{q})^\top \phi(\mathbf{k}_j)}$$

$$\mathbb{E} [\phi(\mathbf{q})^\top \phi(\mathbf{k})] = \exp \mathbf{q}^\top \mathbf{k}$$

Random Fourier features

Rahimi and Recht (2007)

From Attention to Random Feature Attention

$\mathbf{q}$ query
 $\mathbf{k}_i$ keys
 $\mathbf{v}_i$ values

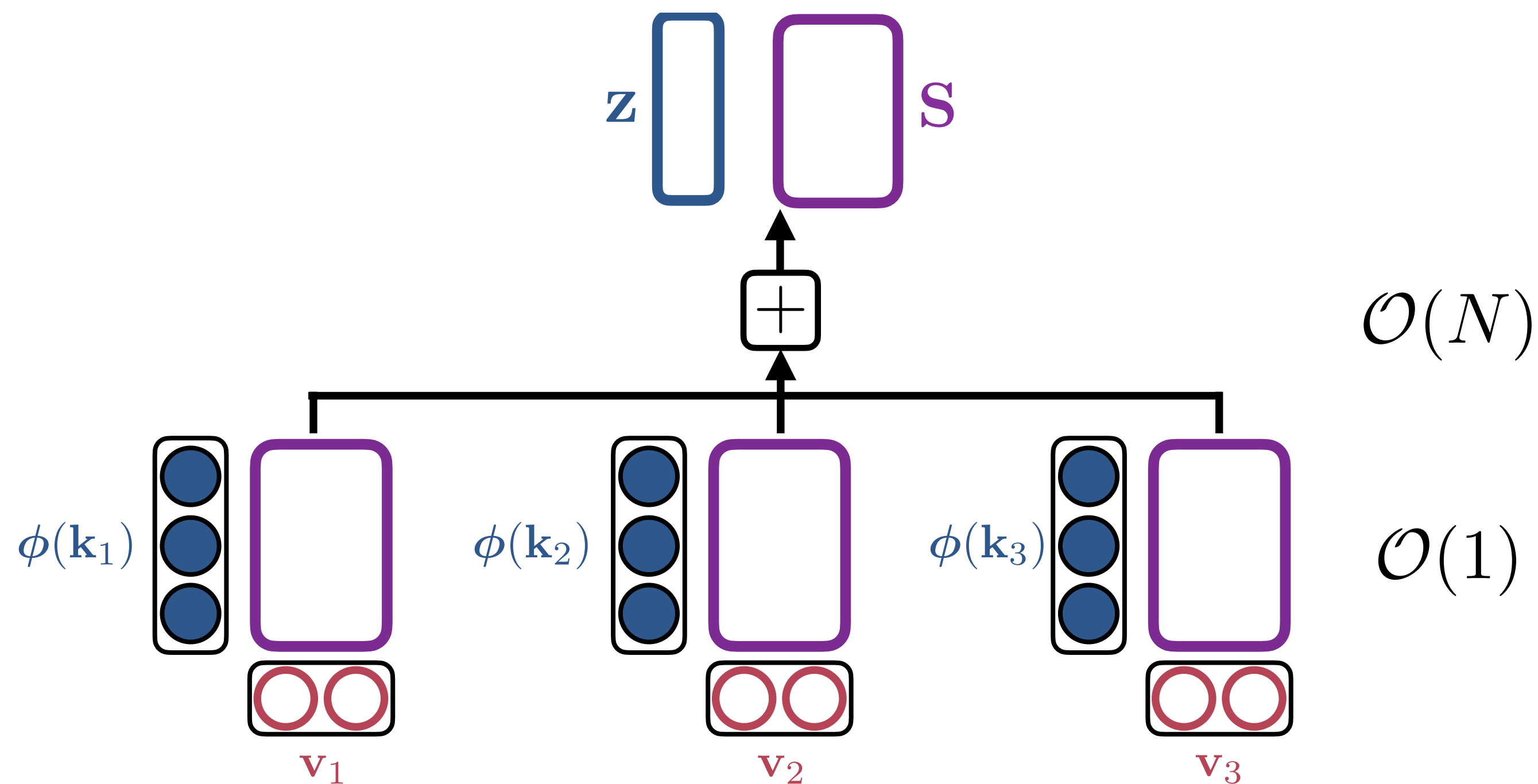
$$\begin{aligned} & \sum_i \frac{\exp \mathbf{q}^\top \mathbf{k}_i}{\sum_j \exp \mathbf{q}^\top \mathbf{k}_j} \mathbf{v}_i^\top \\ & \approx \sum_i \frac{\phi(\mathbf{q})^\top \phi(\mathbf{k}_i) \otimes \mathbf{v}_i}{\sum_j \phi(\mathbf{q})^\top \phi(\mathbf{k}_j)} \\ & = \frac{\phi(\mathbf{q})^\top \sum_i \phi(\mathbf{k}_i) \otimes \mathbf{v}_i}{\phi(\mathbf{q})^\top \sum_j \phi(\mathbf{k}_j)} \end{aligned}$$

Moving $\phi(\mathbf{q})$
out of the sum

Random Feature Attention

$$\mathbf{S} = \sum_i \phi(\mathbf{k}_i) \otimes \mathbf{v}_i$$

$$\mathbf{z} = \sum_j \phi(\mathbf{k}_j)$$

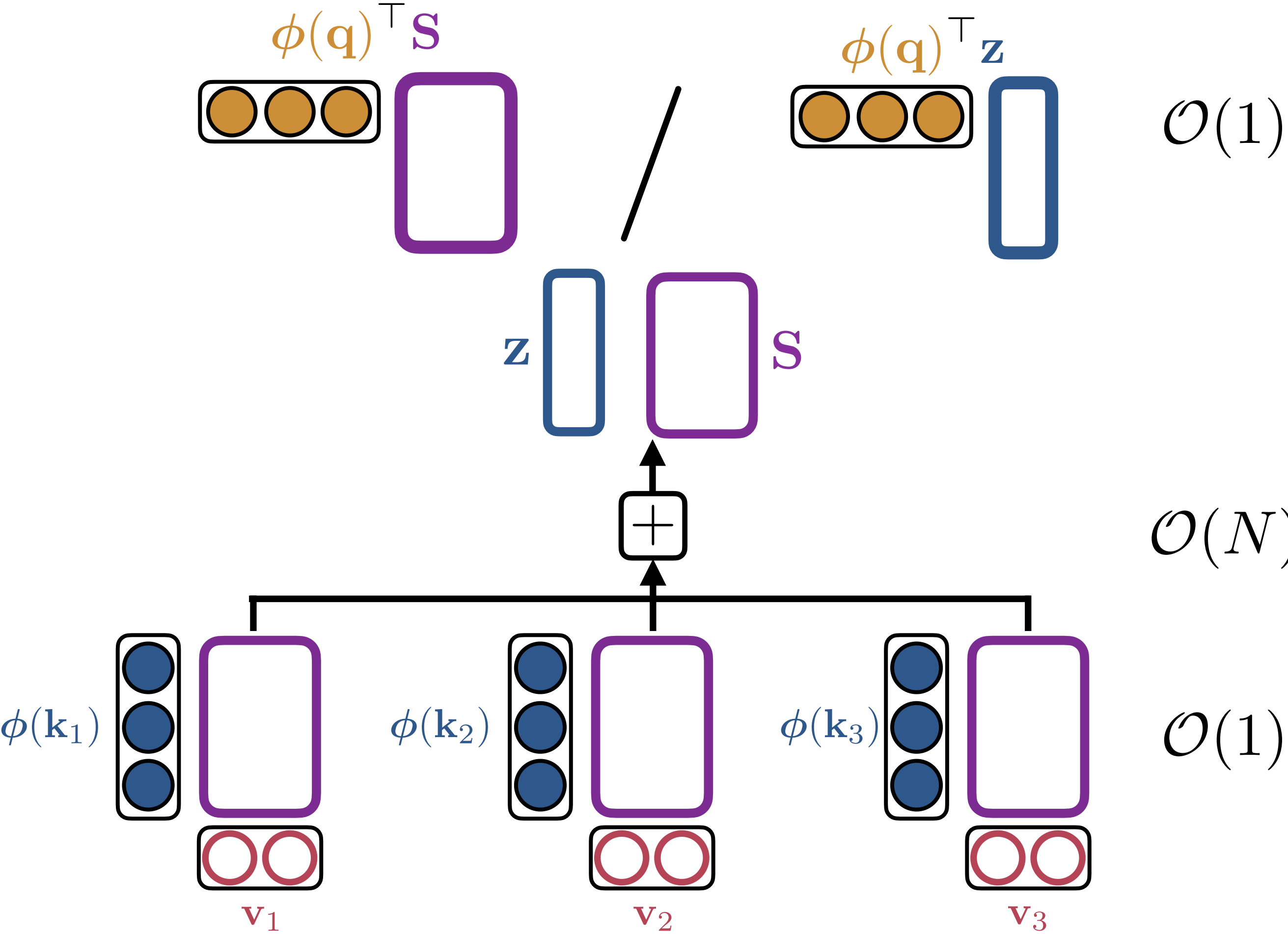


Random Feature Attention

$$\text{output} = \phi(\mathbf{q})^\top \mathbf{S} / (\phi(\mathbf{q})^\top \mathbf{z})$$

$$\mathbf{S} = \sum_i \phi(\mathbf{k}_i) \otimes \mathbf{v}_i$$

$$\mathbf{z} = \sum_j \phi(\mathbf{k}_j)$$



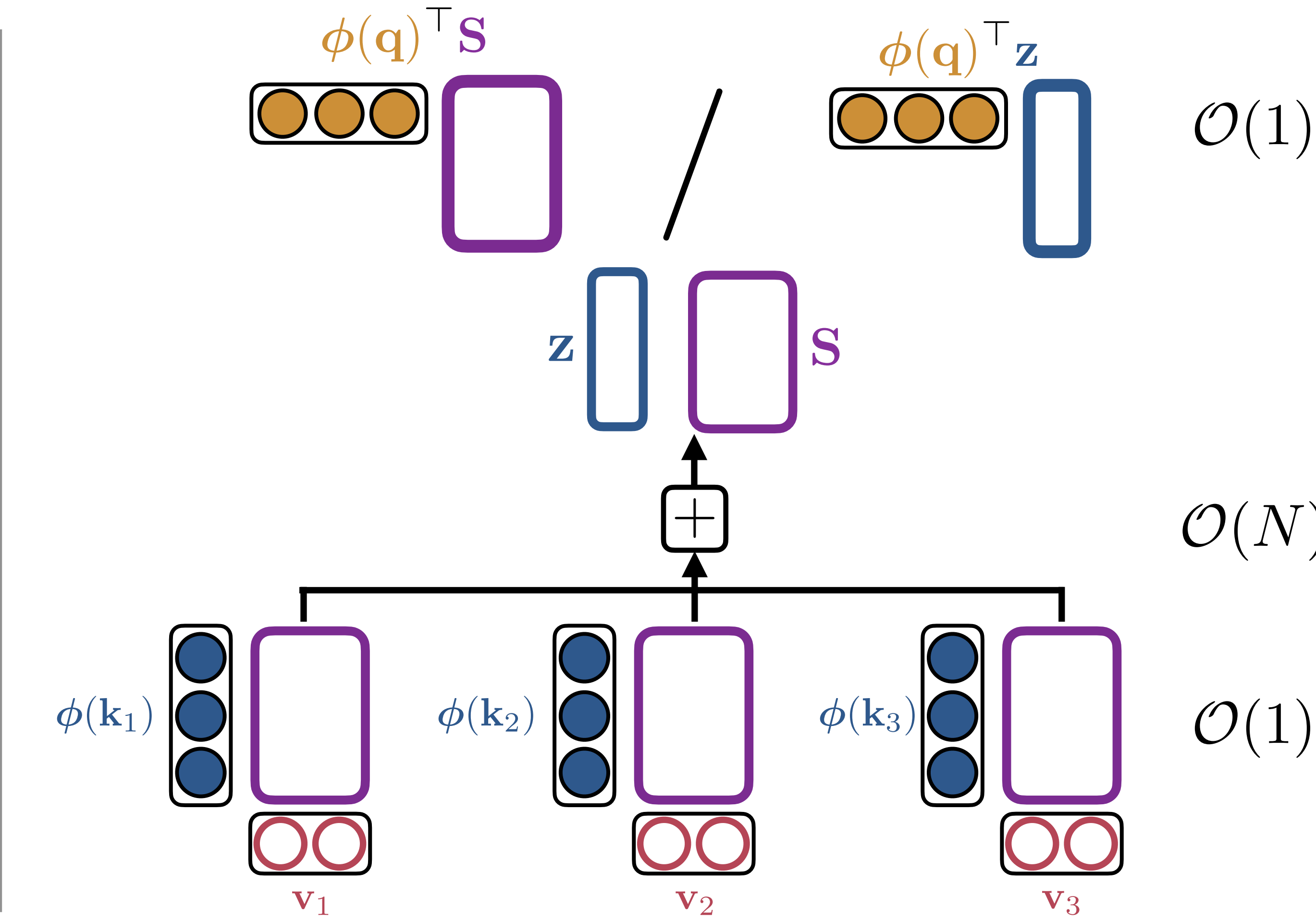
Random Feature Attention

$$\text{output} = \phi(\mathbf{q})^\top \mathbf{S} / (\phi(\mathbf{q})^\top \mathbf{z})$$

$$\mathbf{S} = \sum_i \phi(\mathbf{k}_i) \otimes \mathbf{v}_i$$

$$\mathbf{z} = \sum_j \phi(\mathbf{k}_j)$$

per step: $\mathcal{O}(1)$
 overall: $\mathcal{O}(N)$



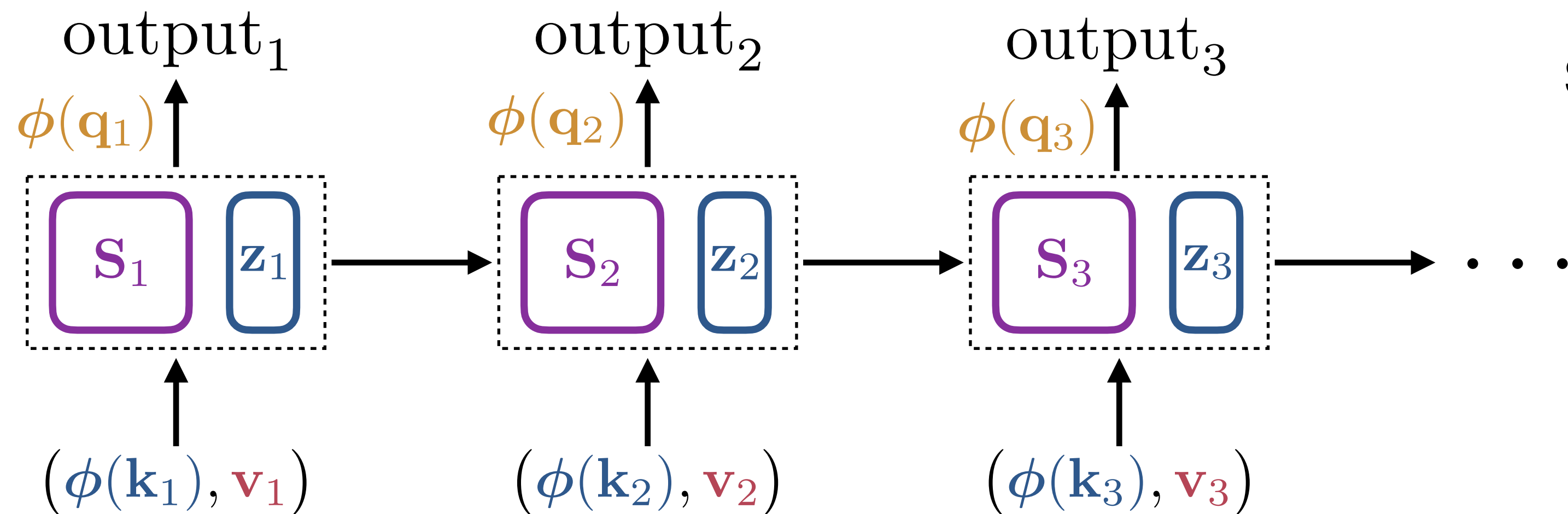
Random Feature Attention

Recurrent Updates

$$\mathbf{S}_t = \mathbf{S}_{t-1} + \phi(\mathbf{k}_t) \otimes \mathbf{v}_t$$

$$\mathbf{z}_t = \mathbf{z}_{t-1} + \phi(\mathbf{k}_t)$$

$$\text{output}_t = \phi(\mathbf{q}_t)^\top \mathbf{S}_t / \phi(\mathbf{q}_t)^\top \mathbf{z}_t$$

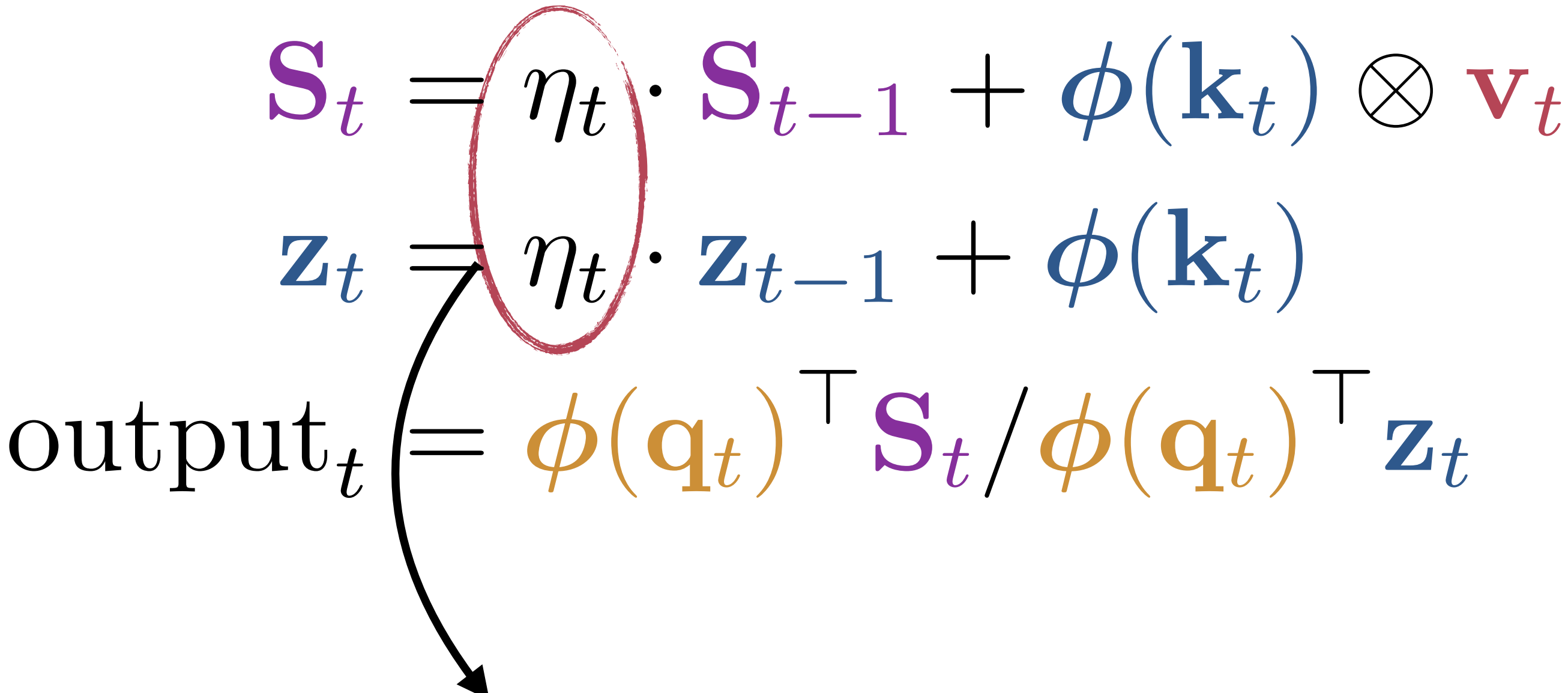


Applications

- Language model
- Decoder self attention in a sequence-to-sequence model

Random Feature Attention

Recency Bias with Learned Gates

$$\begin{aligned}\mathbf{S}_t &= \eta_t \cdot \mathbf{S}_{t-1} + \phi(\mathbf{k}_t) \otimes \mathbf{v}_t \\ \mathbf{z}_t &= \eta_t \cdot \mathbf{z}_{t-1} + \phi(\mathbf{k}_t) \\ \text{output}_t &= \phi(\mathbf{q}_t)^\top \mathbf{S}_t / \phi(\mathbf{q}_t)^\top \mathbf{z}_t\end{aligned}$$


learned sigmoid gate

$$\eta_t = \sigma(\mathbf{w}^\top \mathbf{x} + b)$$

Experiments with Language Modeling

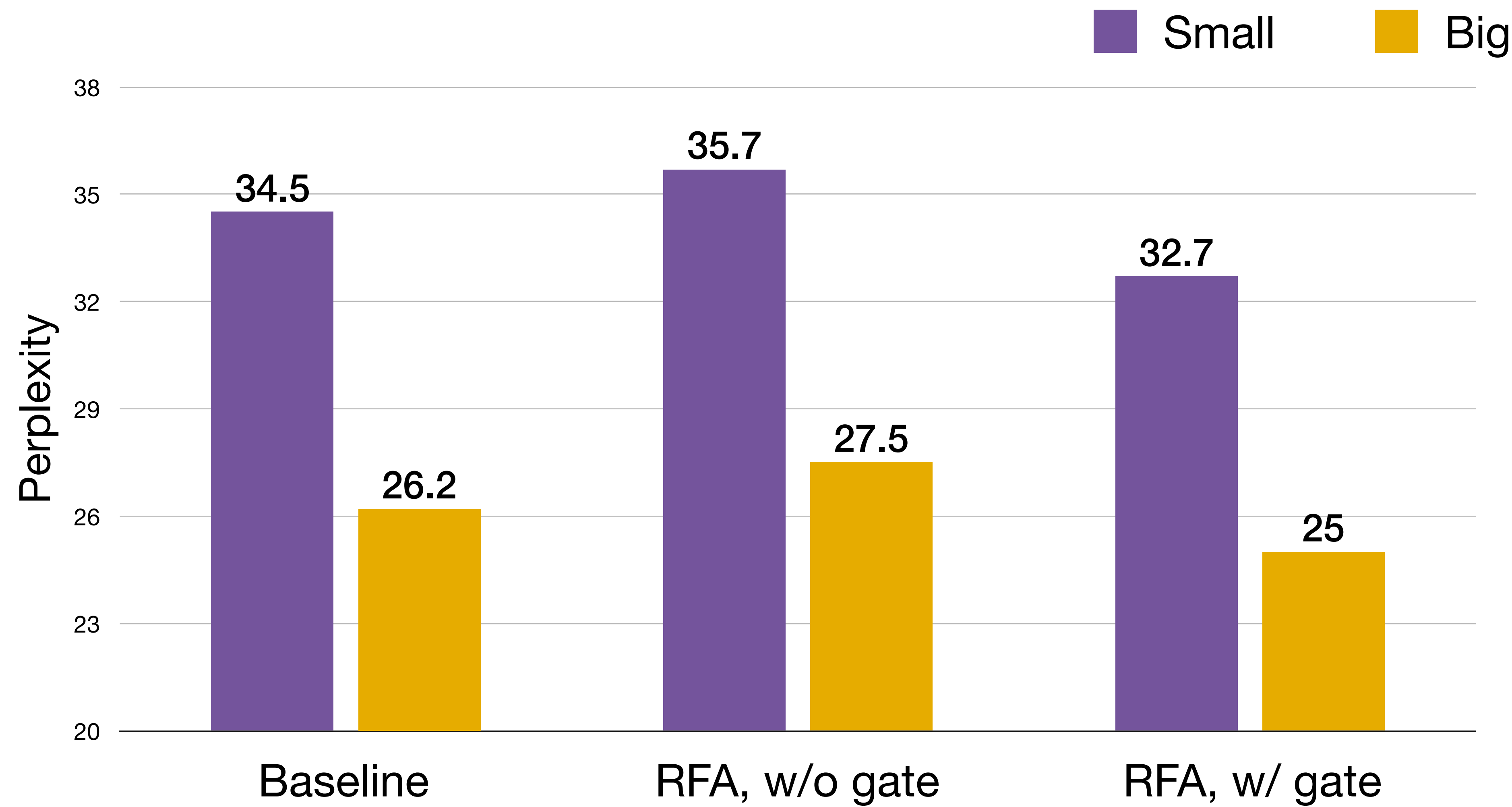
Dataset:

- WikiText-103 ([Merity et al., 2016](#)). 103M training data, 268K vocab size

Implementation:

- Based on [Baevski and Auli, 2019](#)
- Replace all self attention with random feature attention
- Random feature size: 64; context window 512, not “stateful”
- All models trained for 150K steps

Wikitext-103 test set perplexity (lower is better)



Experiments: Machine Translation

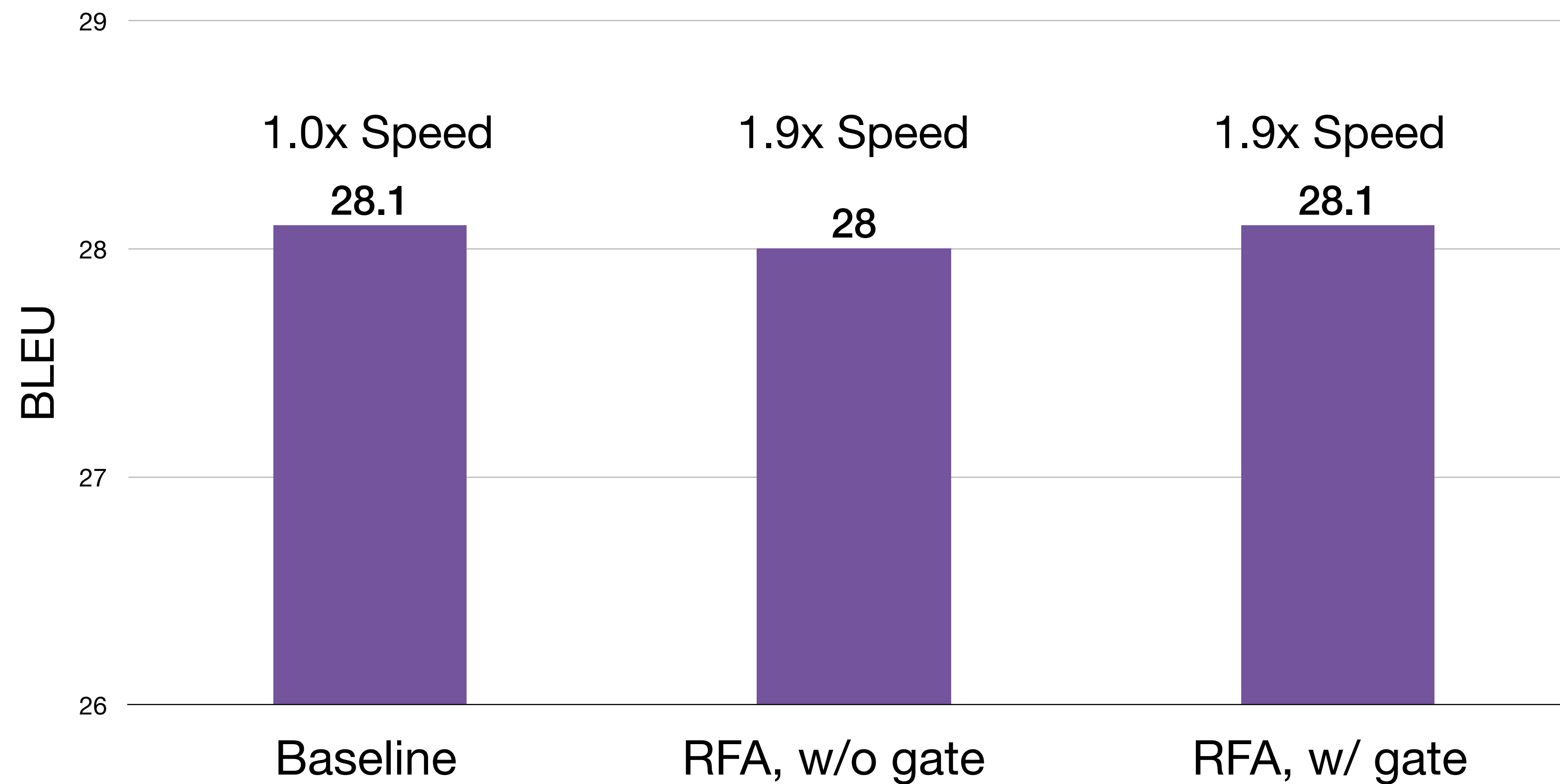
Dataset: WMT'14 ([Bojar et al., 2014](#))

- EN->DE, 4.5M training instances
- EN->FR, 35.8M training instances

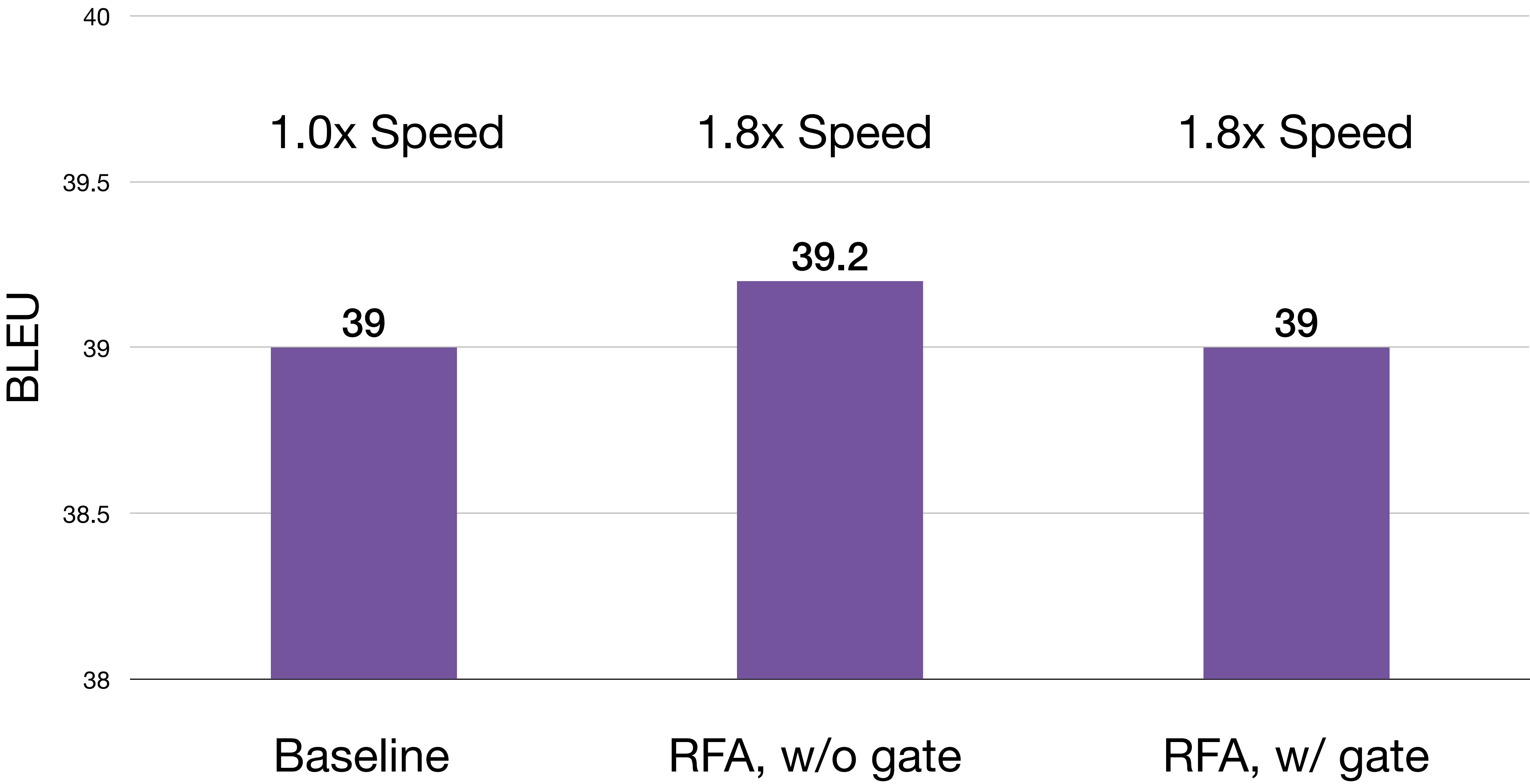
Implementation:

- Based on transformer base ([Vaswani et al., 2017](#))
- Replace **decoder** causal and cross attention with random feature attention
- Random feature size: 64 causal, 128 cross
- Trained for up to 350K steps; beam size 4; average 10 checkpoints

Test set BLEU on WMT'14 EN->DE

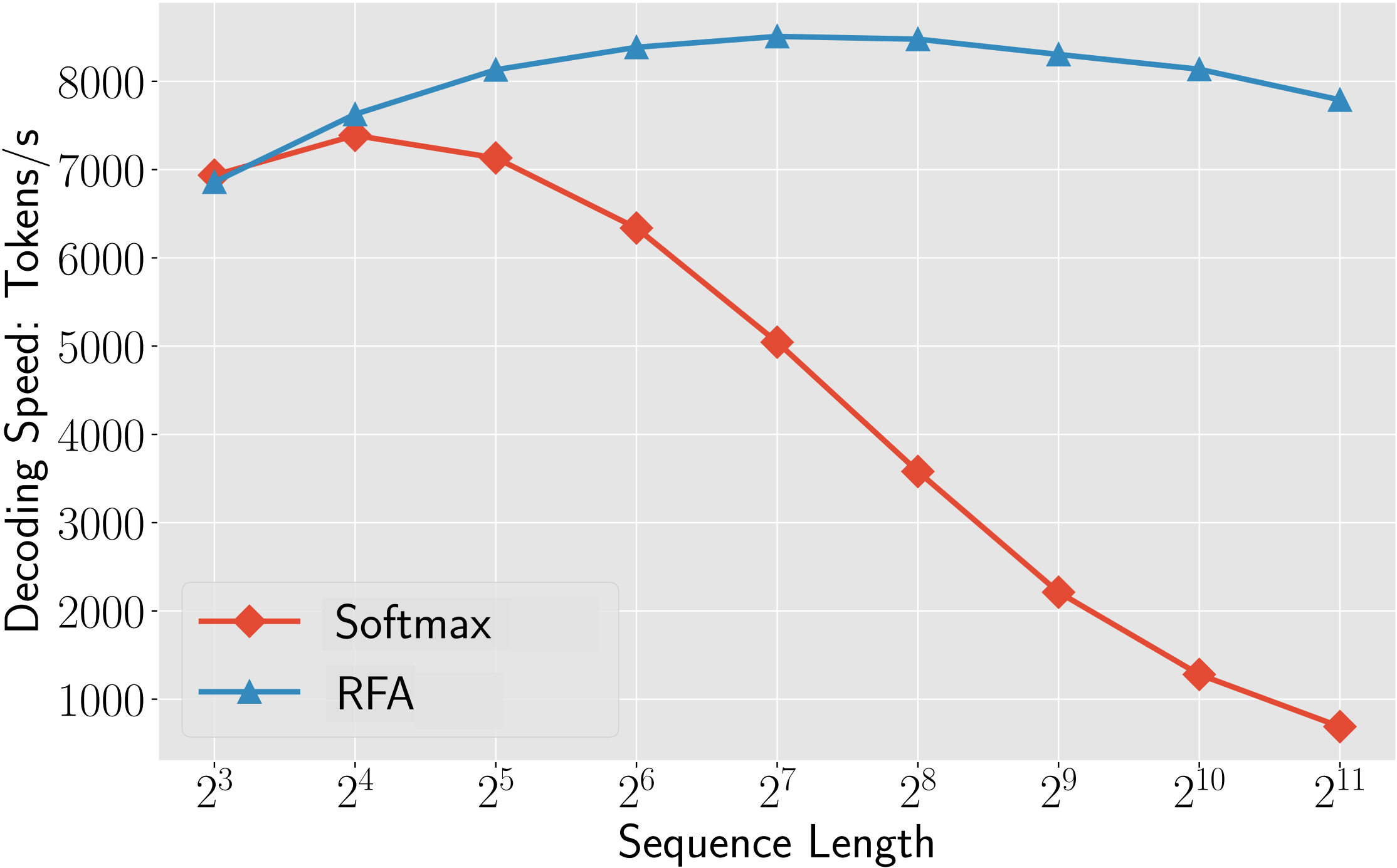


Test set BLEU on WMT'14 EN->FR

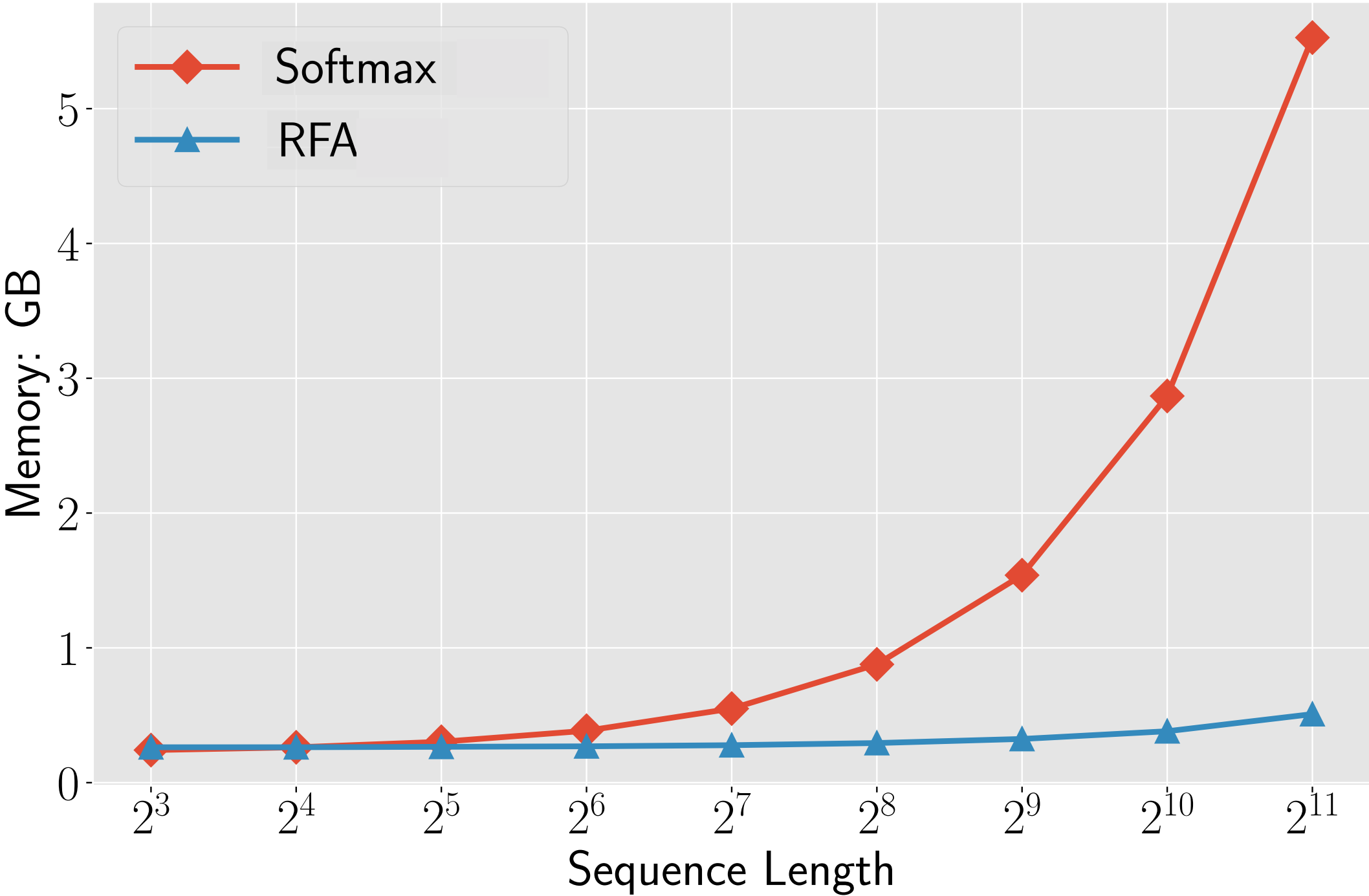


Decoding Speed & Memory vs. Lengths

Speed



Memory



Wrap-up

RFA:

- Linear complexity attention with random feature methods
- Well-suited for tasks involving long sequences
- Recurrent style update; intuitive ways to connect to gated RNNs

Experiments:

- Strong performance on language modeling and machine translation
- 1.9x speed up in MT decoding; more for longer text
- Competitive in both efficiency and accuracy on character-level text classification ([Tay et al., 2021](#))

Thank You!

Please check out our poster

collaborators

paper



bit.ly/3rnYRTw

code



bit.ly/3ISzisS



DeepMind