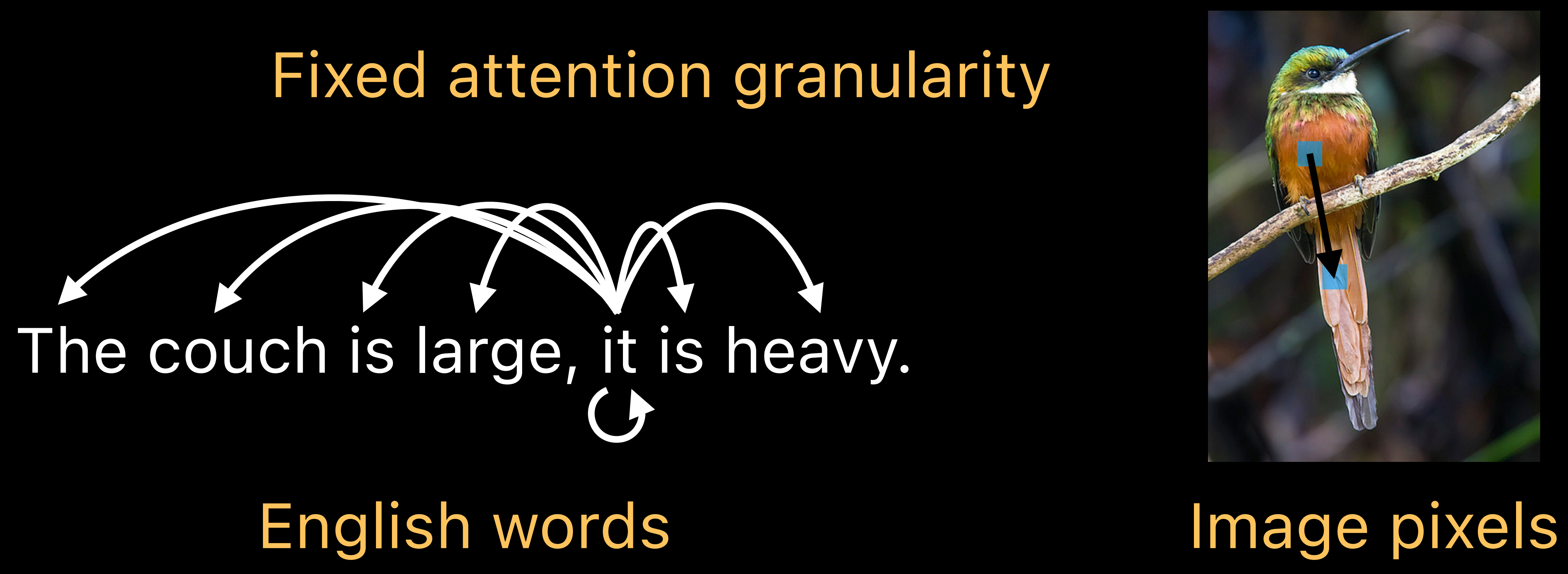
Efficient Representation Learning via Adaptive Context Pooling

Chen Huang, Walter Talbott, Navdeep Jaitly, Josh Susskind | ICML 2022
Apple Inc.

Motivation

Self-attention models capture long-range context by pairwise attention

- Assume fixed attention granularity defined by individual tokens
- Limited for modeling complex contextual dependencies
- Costly: may need many layers to make up for the fixed granularity



Literature

Hierarchical context in Transformers - fixed scaling scheme

- Swin transformer [ICCV 2021], PVT [ICCV 2021] …

Area attention [ICML 2019]

- Multi-scale memory captures rich context with fixed pooling sizes

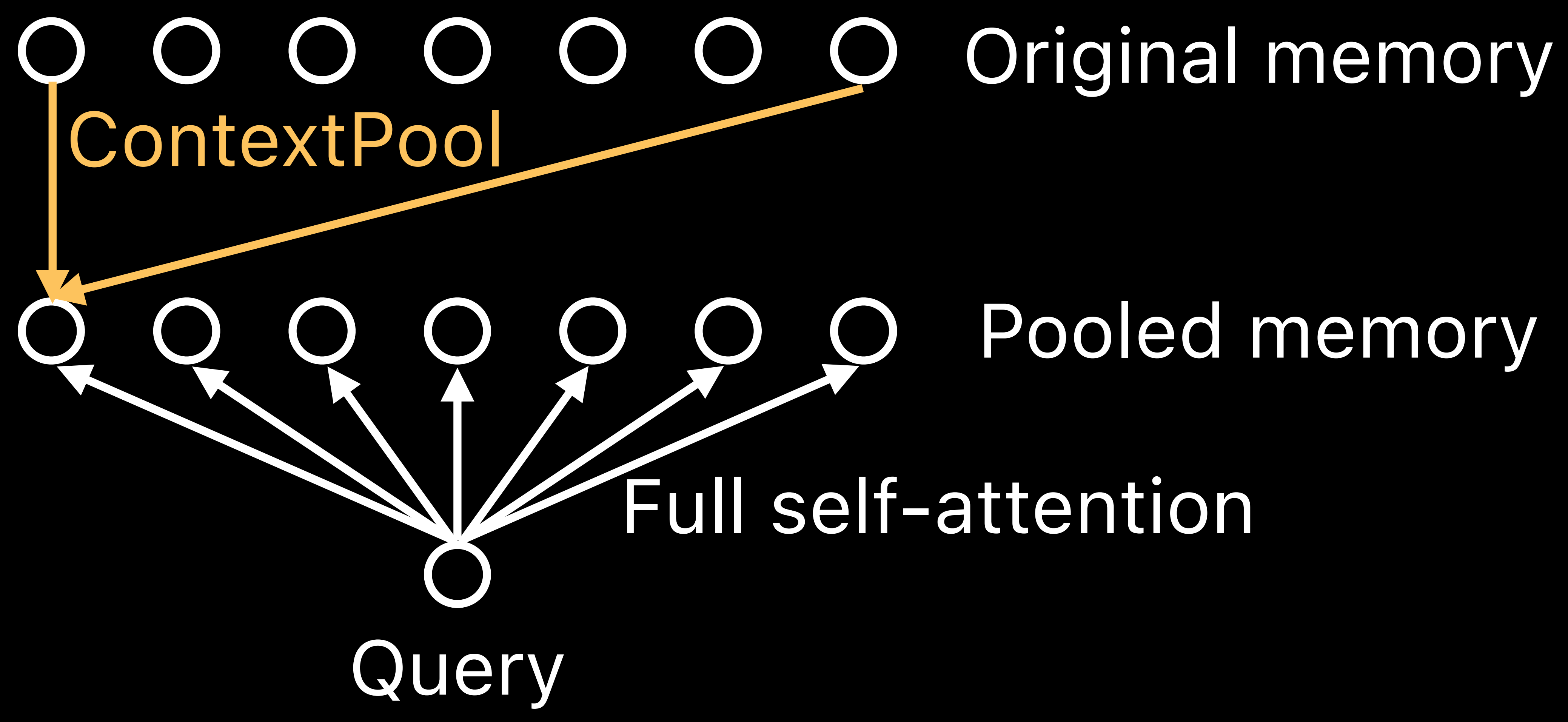
Efficient Transformers with sparse attention/context

- Local window [ACL 2019], blockwise [EMNLP 2020] …

Our Idea

ContextPool for each token

- Pool neighboring features in a memory in-place
- Input-adaptive pooling to encode meaningful context
- Adaptive attention granularity: item-wise→context-wise attention
- Generic mechanism across architectures



Context-wise attention example

ContextPool

Learning adaptive pooling function

$$\mathbf{y}_i = Pool(\mathbf{X}, \mathbf{w}, \mathbf{g}^i) = \sum_{j=1}^{n} x_j \cdot w_j \cdot g_j^i$$

Input feature matrix

$n \times d$

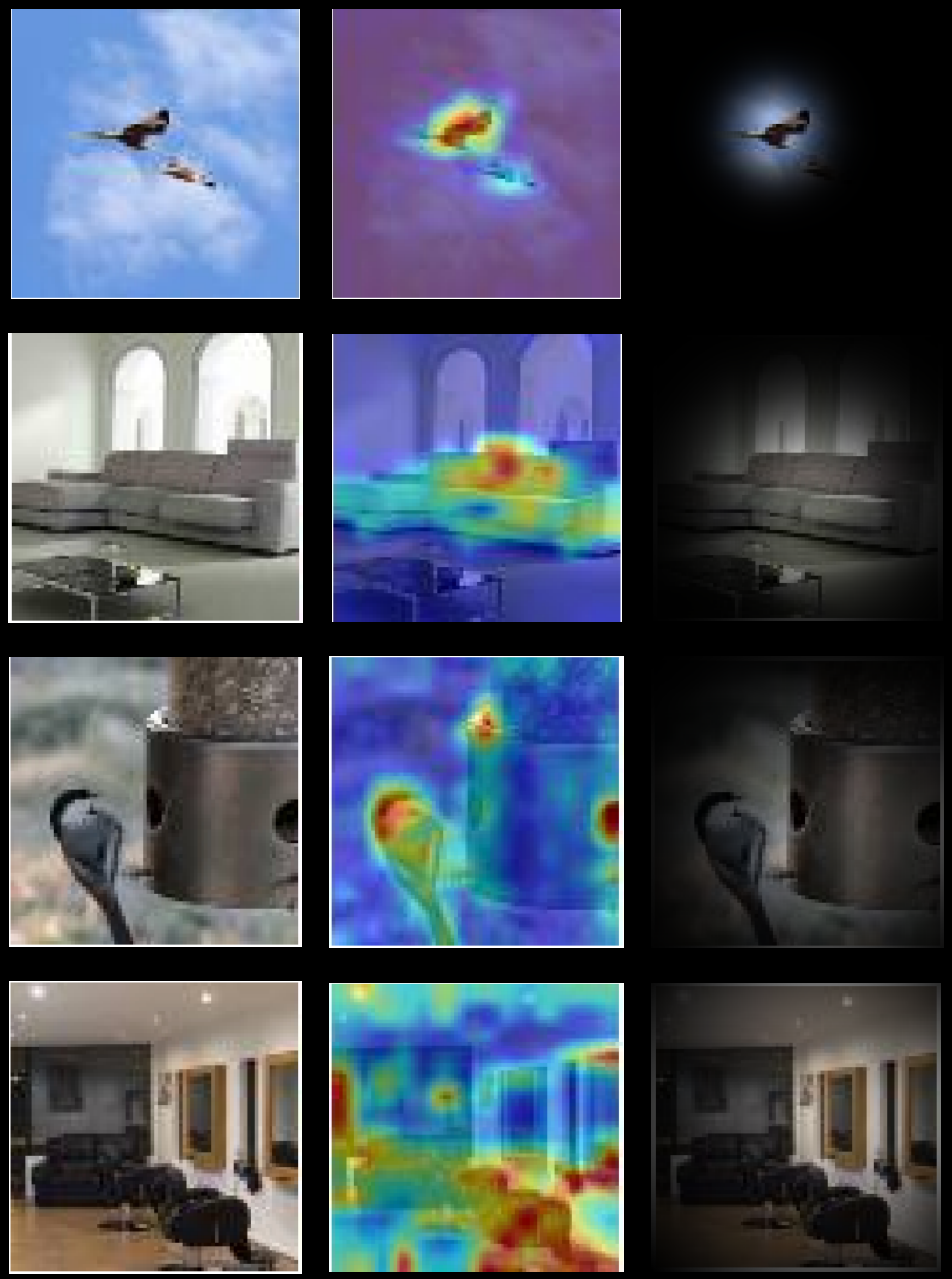
Learned weights

$n \times 1$

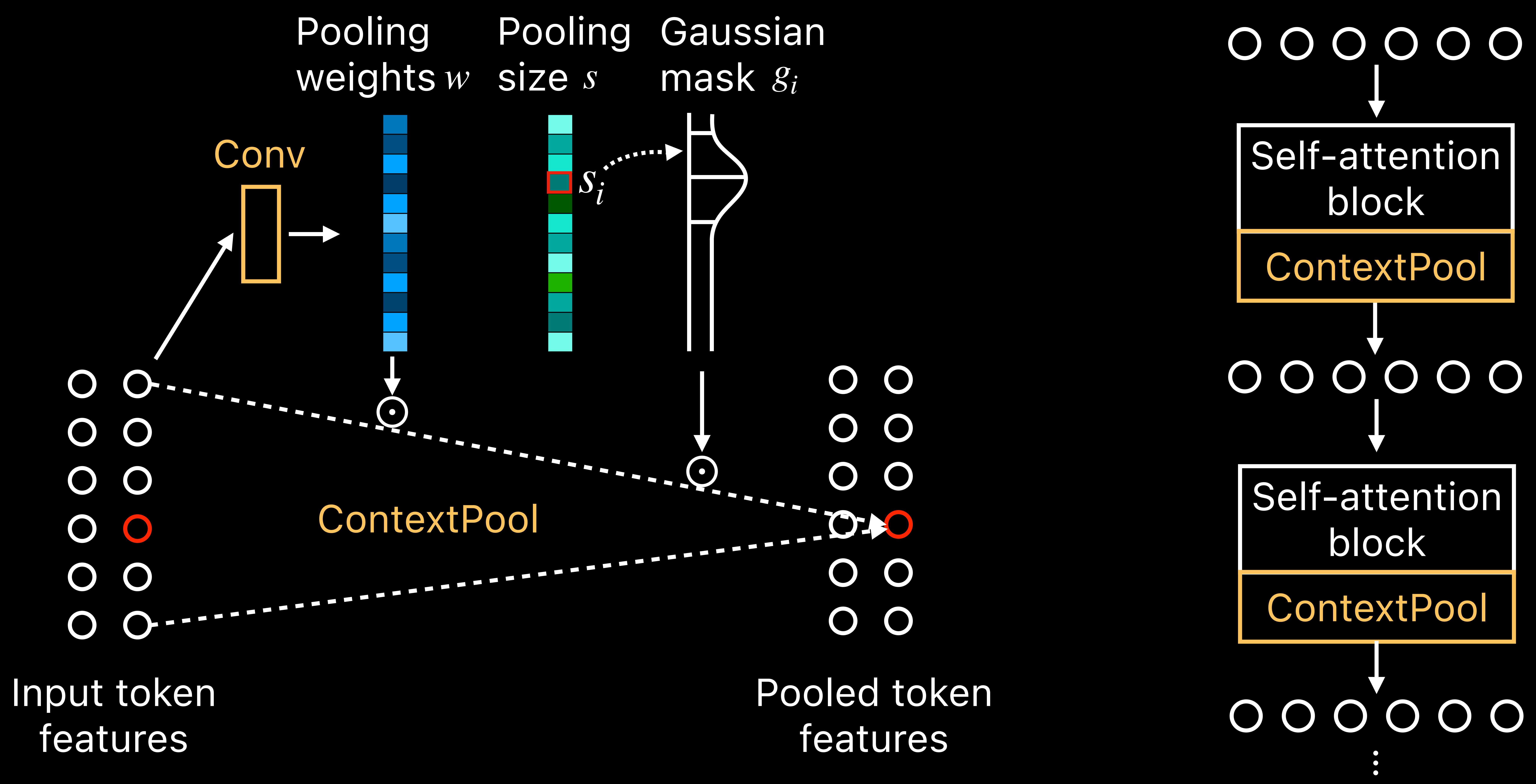
Learned Gaussian mask

$n \times 1$

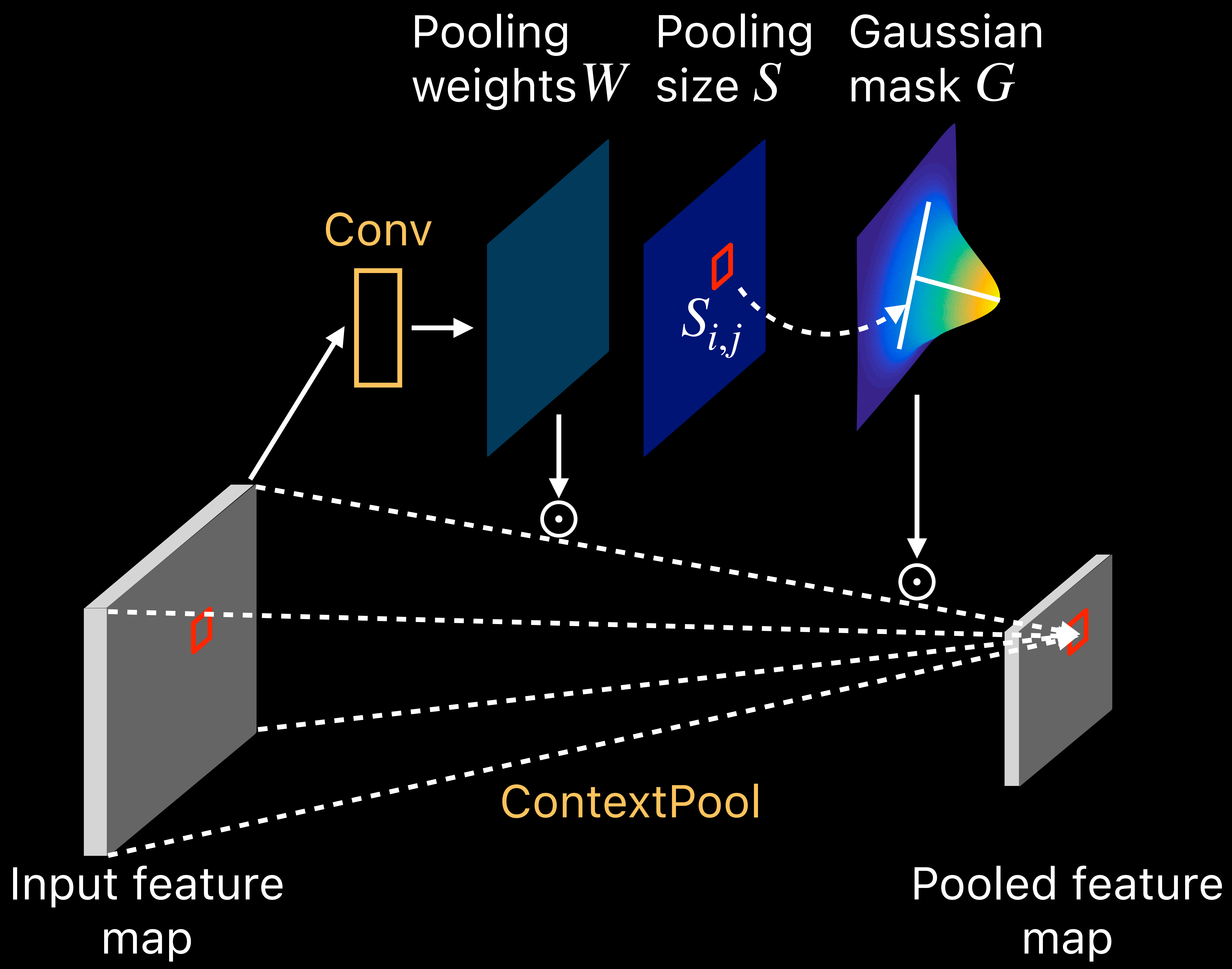
(Encodes pooling size s_i)

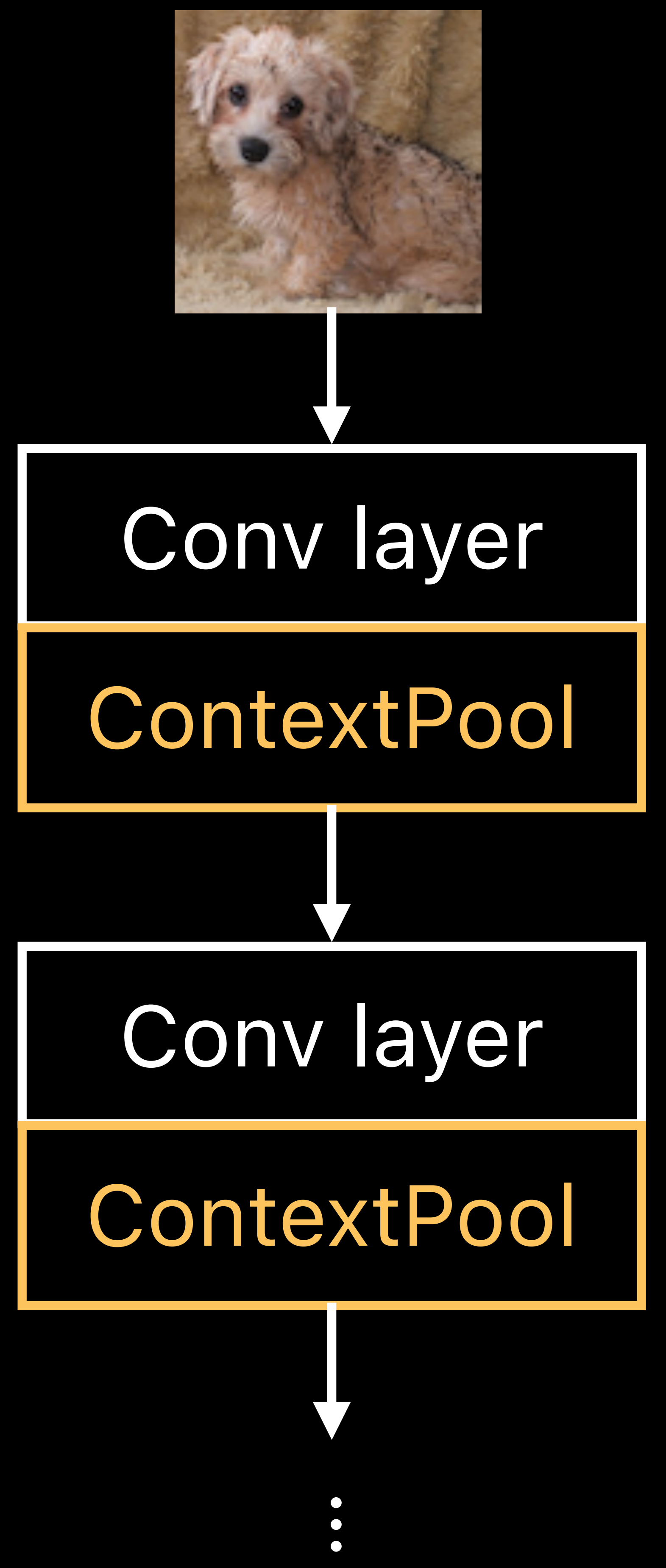


ContextPool for Transformer



ContextPool for ConvNet

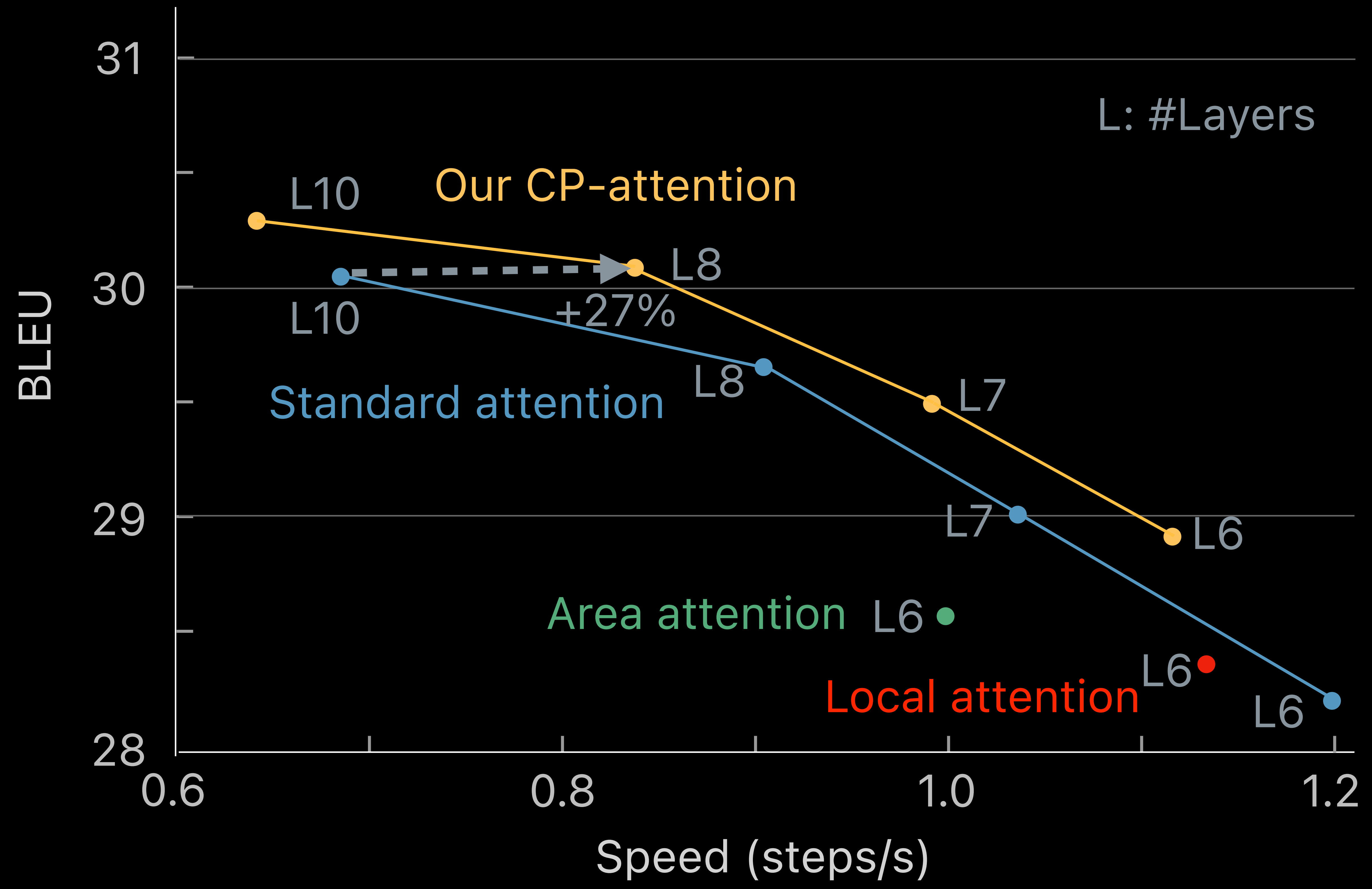




Results

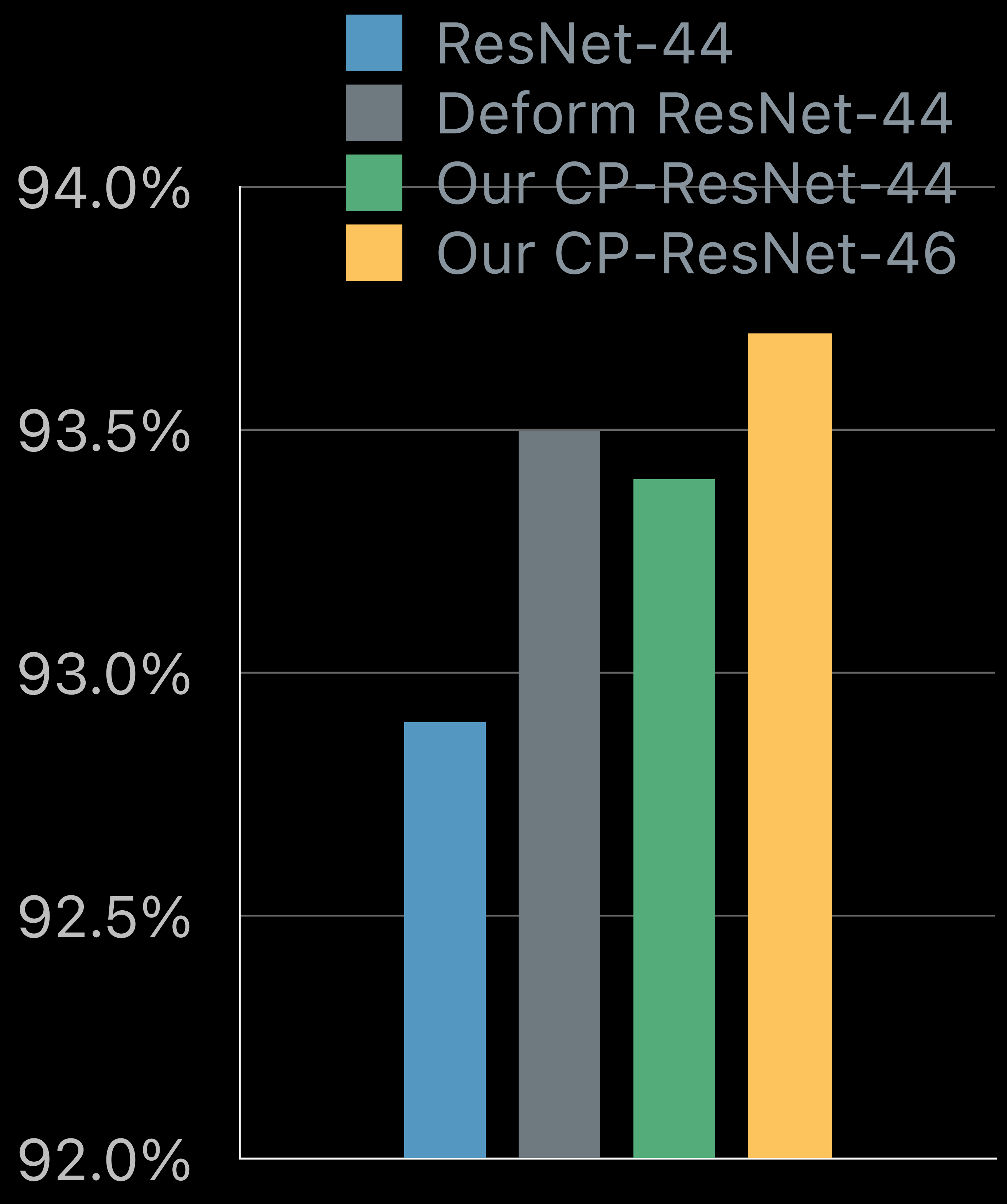
Transformers

Machine translation (EN-DE task)



ConvNets

CIFAR-10 Accuracy



Conclusions

- Introduce ContextPool to model dynamic context and adapt attention granularity
- Improves transformer models in performance-cost trade-off
- Also applicable to ConvNets for efficient but strong representation learning

Paper ID 5218