Attention

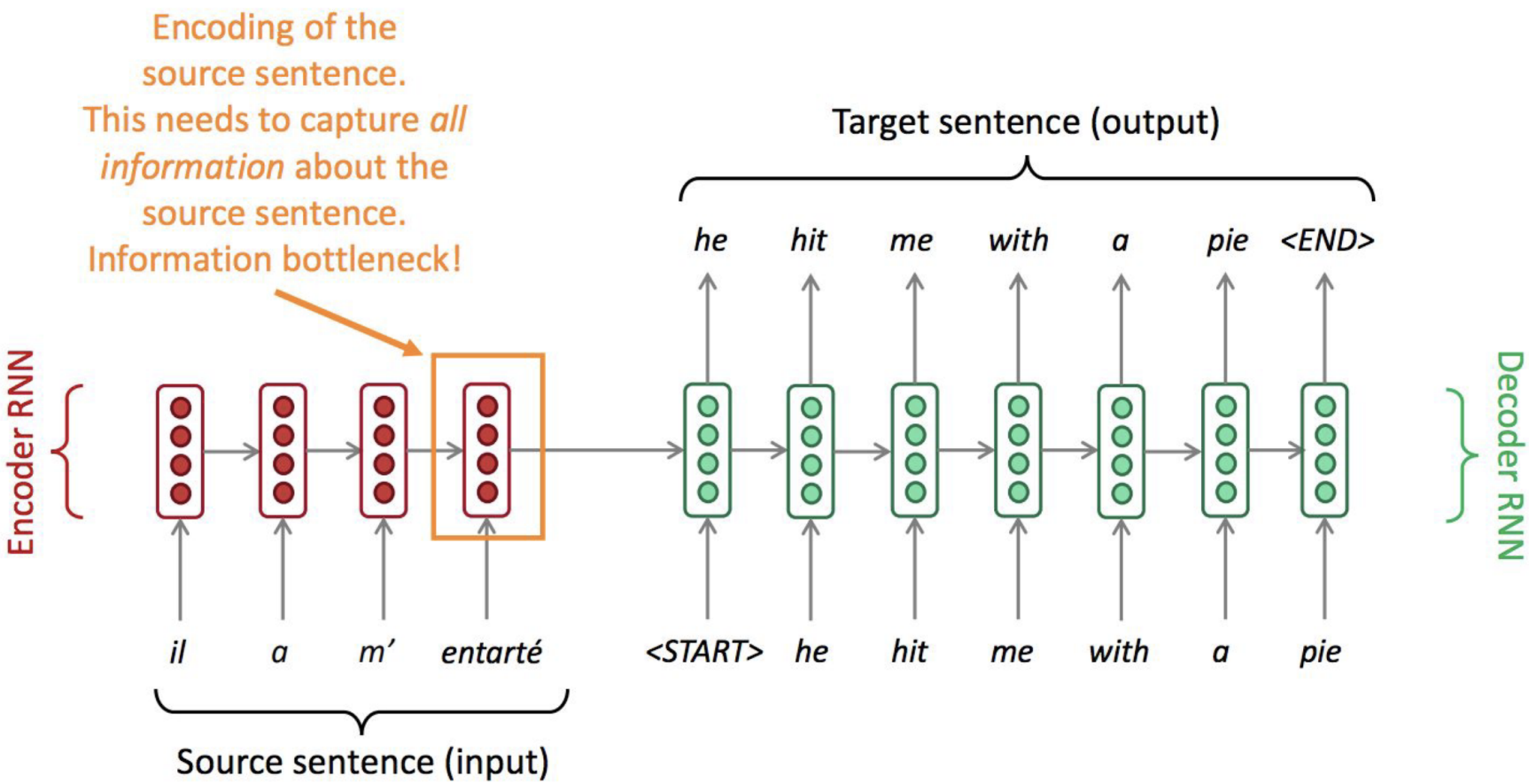
Artificial Intelligence @ Allegheny College

Janyl Jumadinova

April 10, 2023

Credit: Stanford's RNN Notes

RNN Example: Machine Translation



Attention

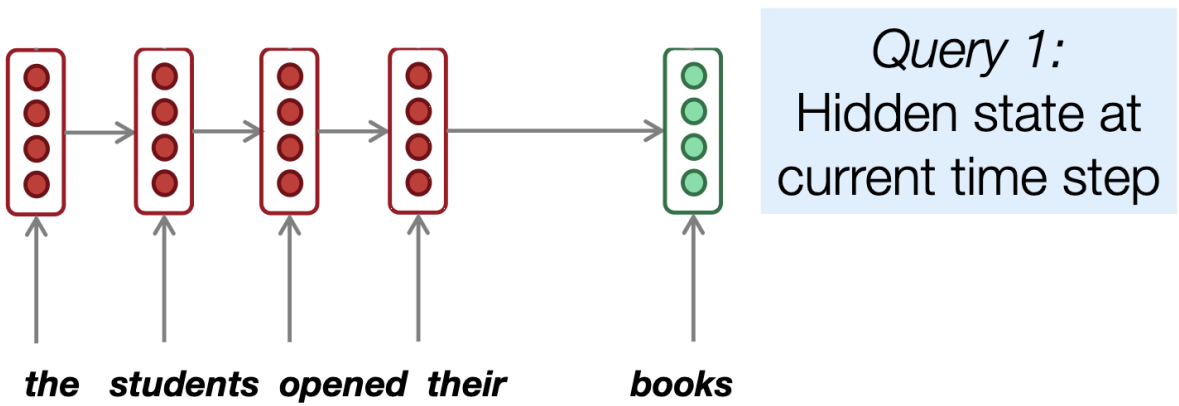
Idea: use new context vector at each step of decoder!

Attention

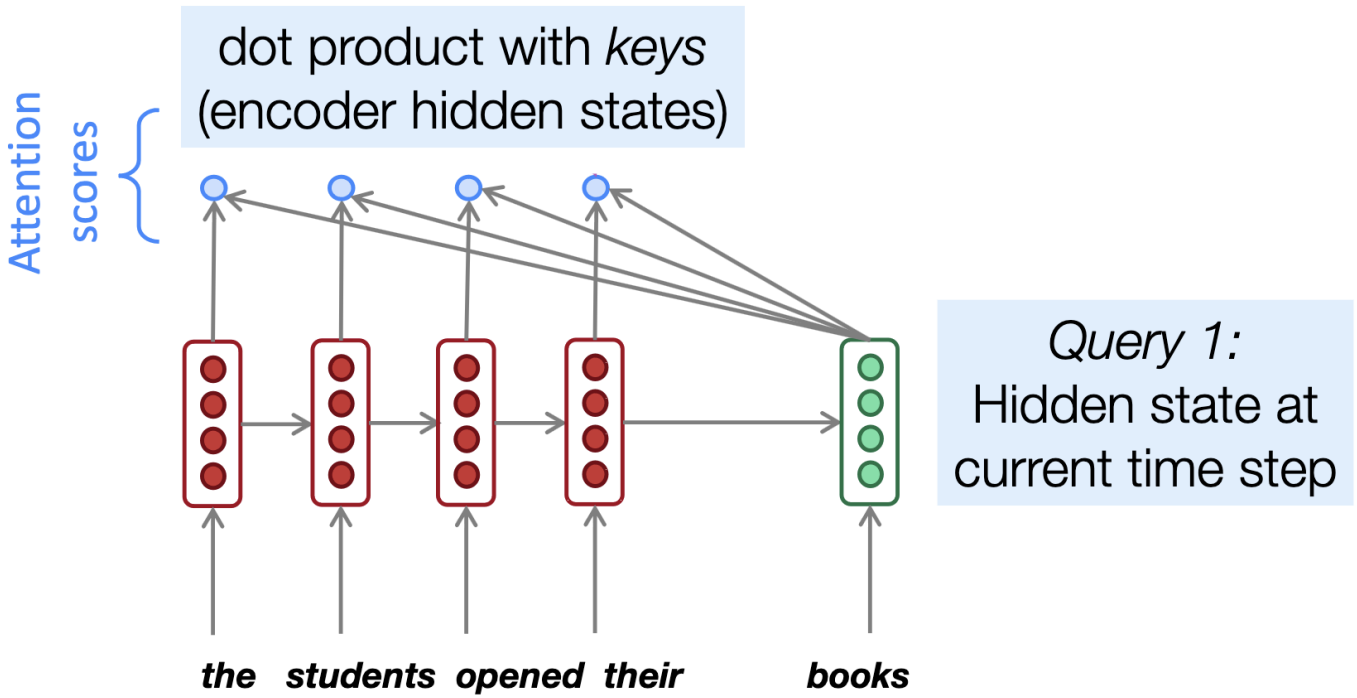
Idea: use new context vector at each step of decoder!

- Attention mechanisms (Bahdanau et al., 2015) allow language models to focus on a particular part of the observed context at each time step.
- Originally developed for machine translation, and intuitively similar to word alignments between different languages.
- **Idea**: in general, we have a single query vector and multiple key vectors. We want to score each query-key pair.

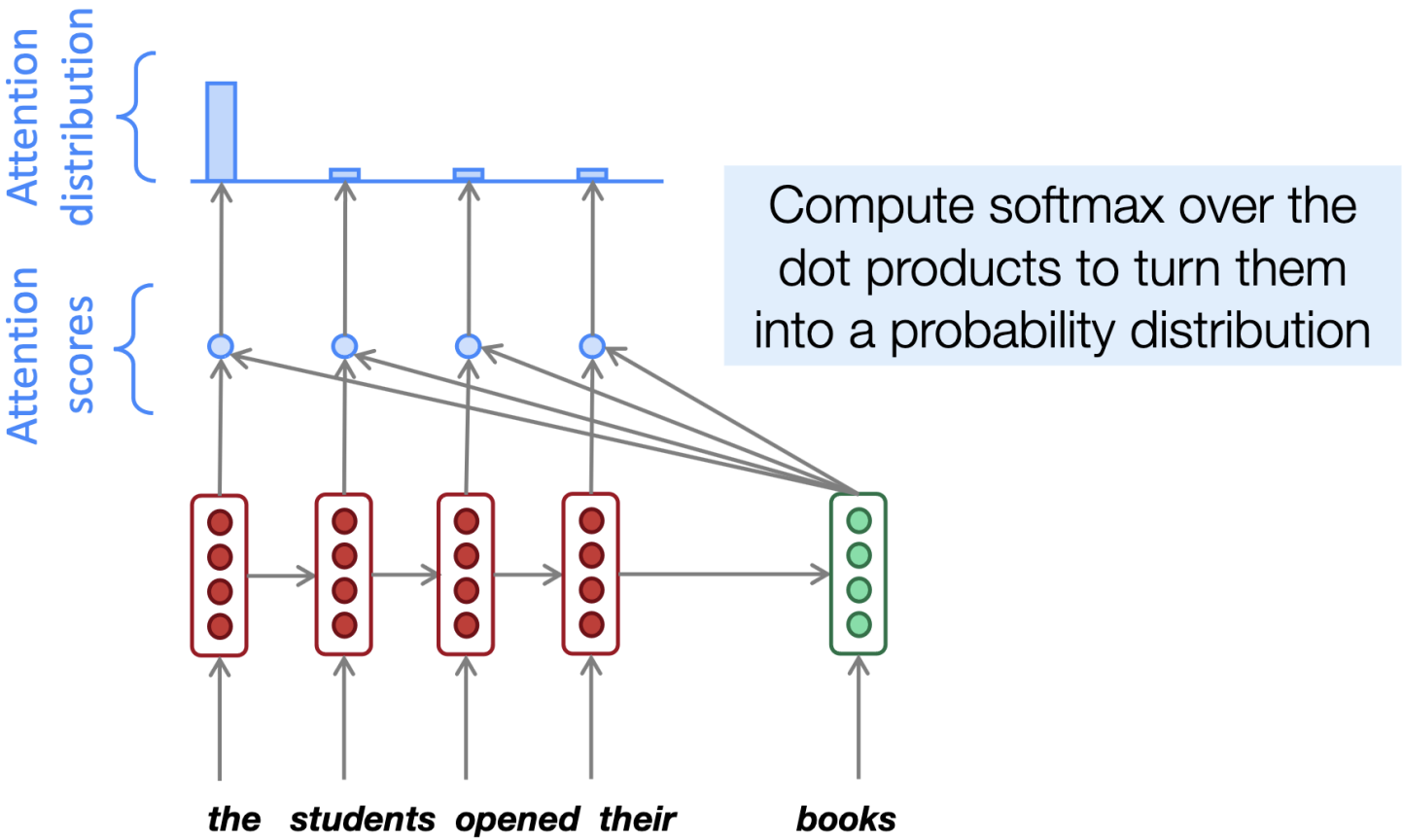
Attention


the
students
opened
their
books
Query 1:
Hidden state at
current time step

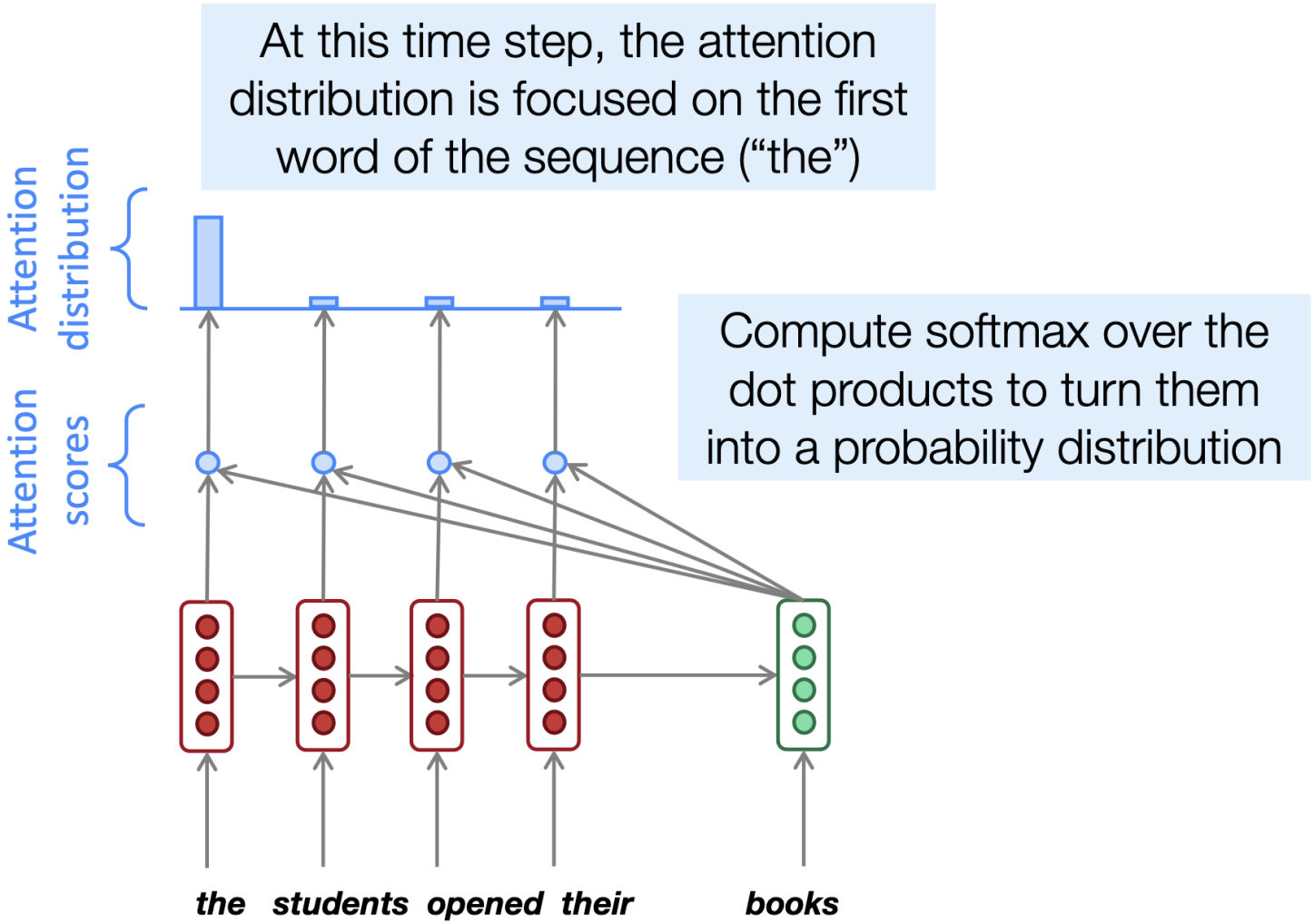
Attention


dot product with *keys*
(encoder hidden states)
Attention scores
Query 1:
Hidden state at
current time step
the *students* *opened* *their* *books*

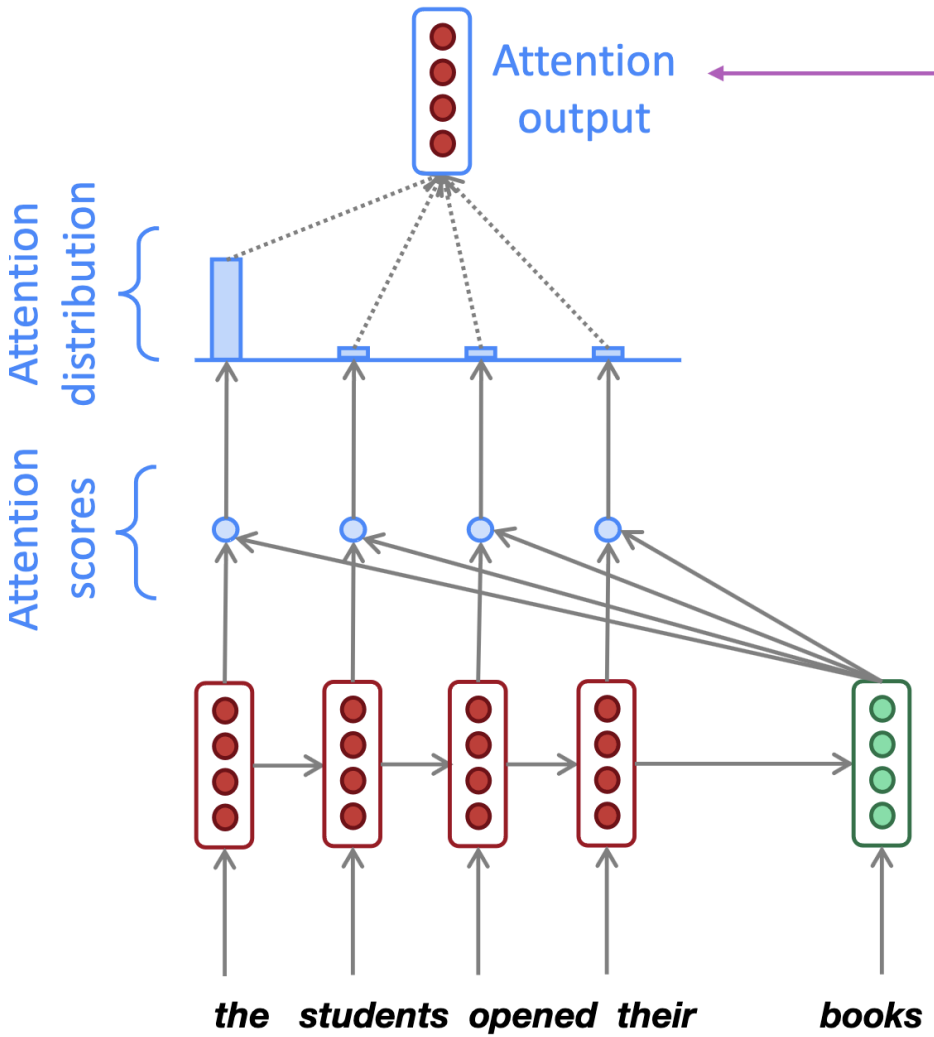
Attention


Attention distribution
Attention scores
Compute softmax over the dot products to turn them into a probability distribution
the
students
opened
their
books

Attention


At this time step, the attention distribution is focused on the first word of the sequence (“the”)
Attention distribution
Attention scores
Compute softmax over the dot products to turn them into a probability distribution
the
students
opened
their
books

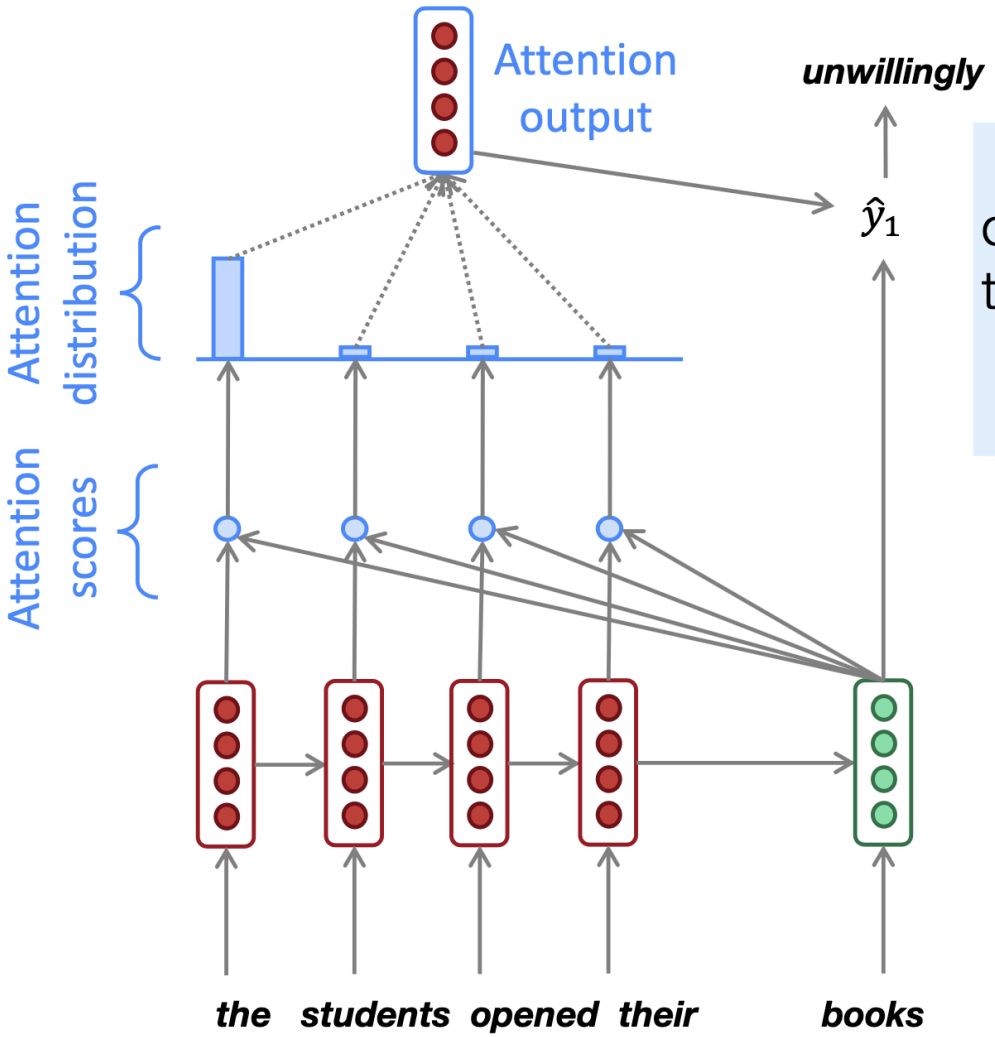
Attention



We use the attention distribution to compute a weighted average of the hidden states.

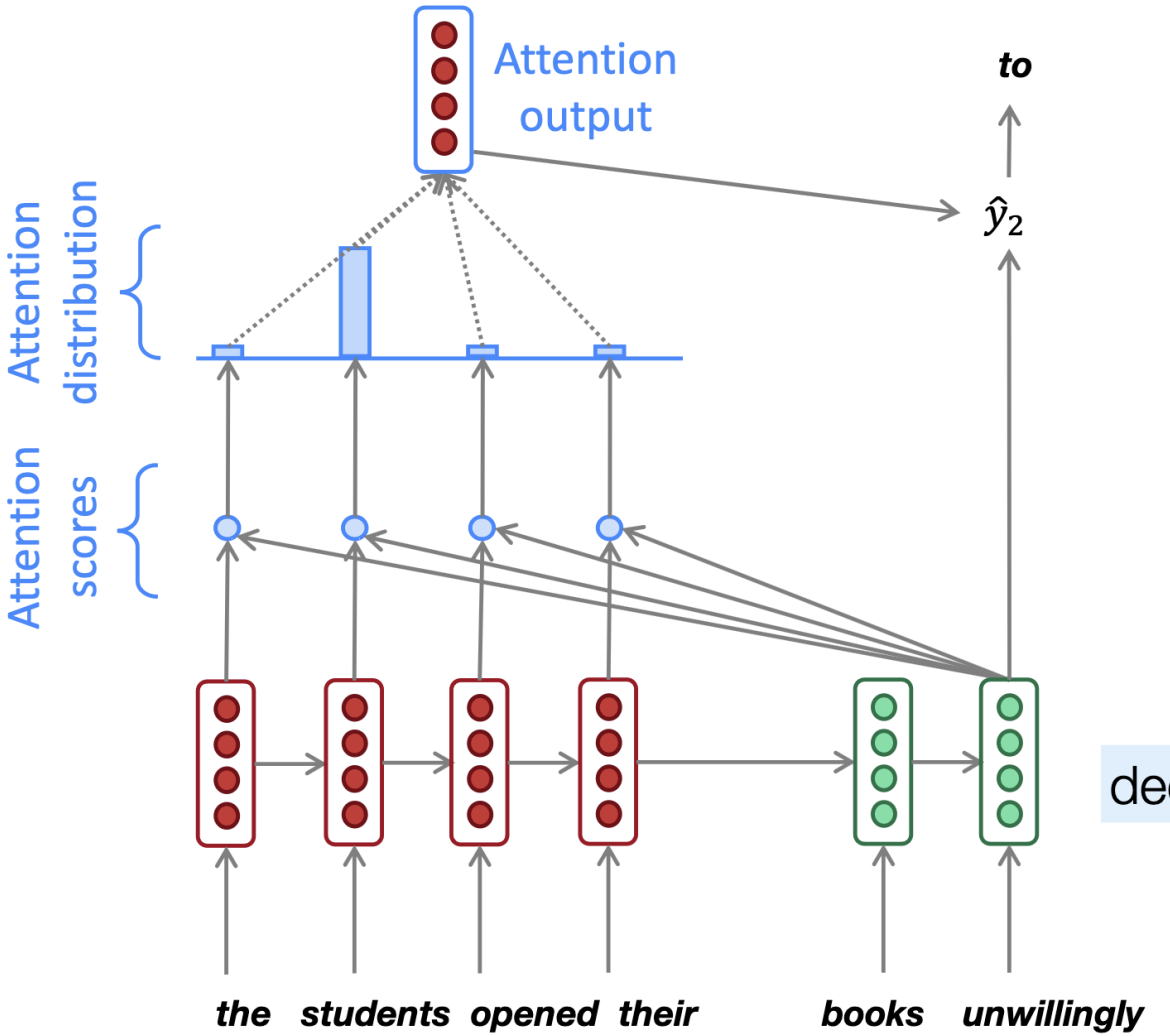
Intuitively, the resulting attention output contains information from hidden states that received high attention scores

Attention



Concatenate (or otherwise compose) the attention output with the current hidden state, then pass through a softmax layer to predict the next word

Attention



decoder, second time step

Attention

- Solves the bottleneck problem: allows decoder to look directly at source.
- Helps with vanishing gradient problem: provides shortcut to faraway states.
- Provides some interpretability: can inspect attention distribution to see what decoder was focusing on.

Many variants of attention

- **Original**: $a(\mathbf{q},\mathbf{k}) = w_2^T tanh(W_1[\mathbf{q};\mathbf{k}])$
- **Bilinear product**: $a(\mathbf{q},\mathbf{k}) = \mathbf{q}^T W \mathbf{k}$ Luong et al., 2015
- **Dot product**: $a(\mathbf{q},\mathbf{k}) = \mathbf{q}^T \mathbf{k}$ Luong et al., 2015
- **Scaled dot product**: $a(\mathbf{q},\mathbf{k}) = \mathbf{q}^T \mathbf{k} / \sqrt{|\mathbf{k}|}$ Vaswani et al., 2017

Self-Attention

- Typical attention is coupled with RNN.
- "Attention is all you need" paper in 2017 (`https://arxiv.org/abs/1706.03762`) - no need for RNNs, just use attention for the encoding, leading to birth of transformers.
- Used self-attention models inspired by Cheng et al. `https://arxiv.org/pdf/1601.06733.pdf`