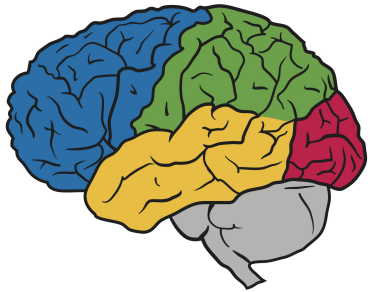
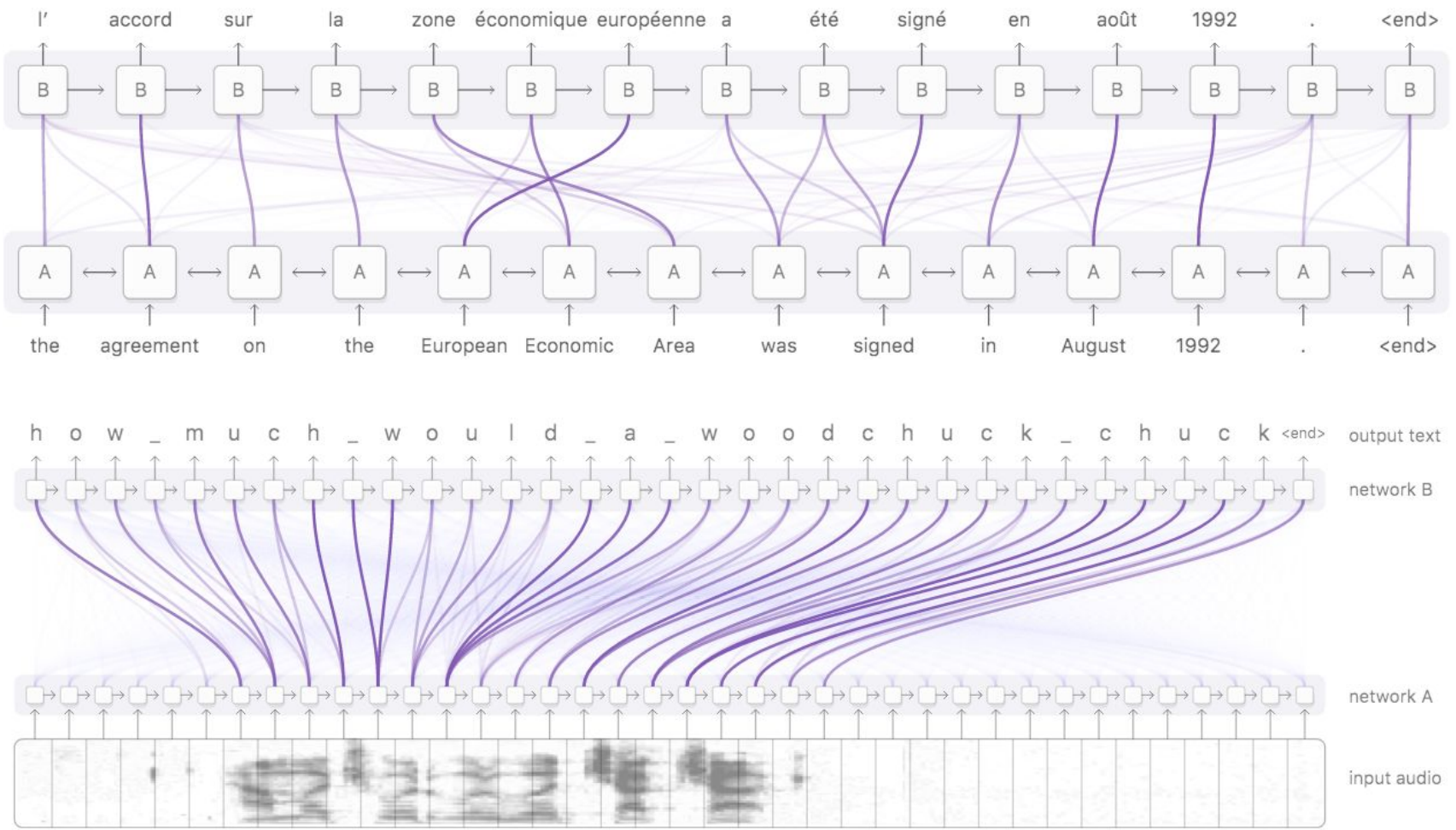
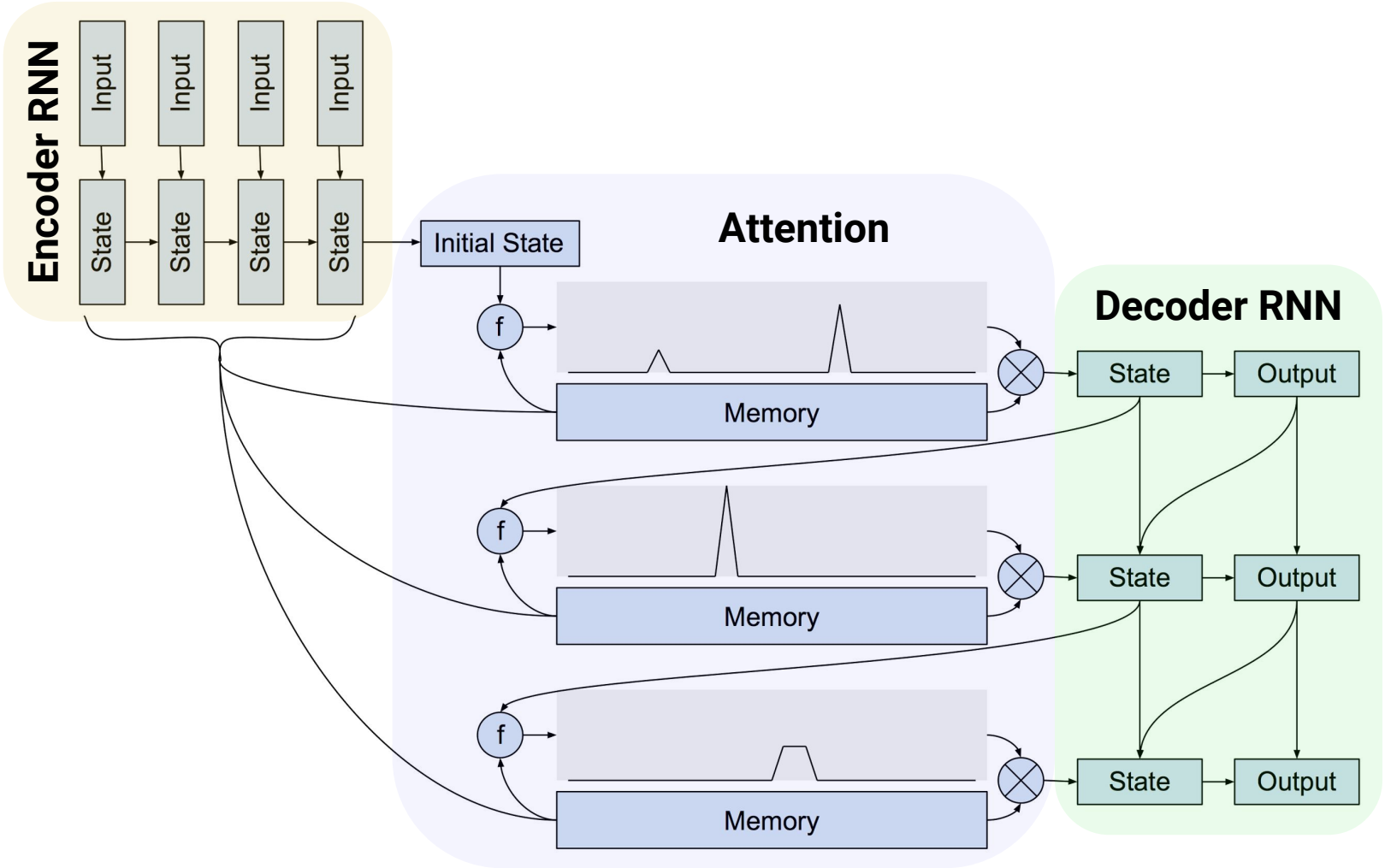
Doing Strange Things with Attention

Colin Raffel

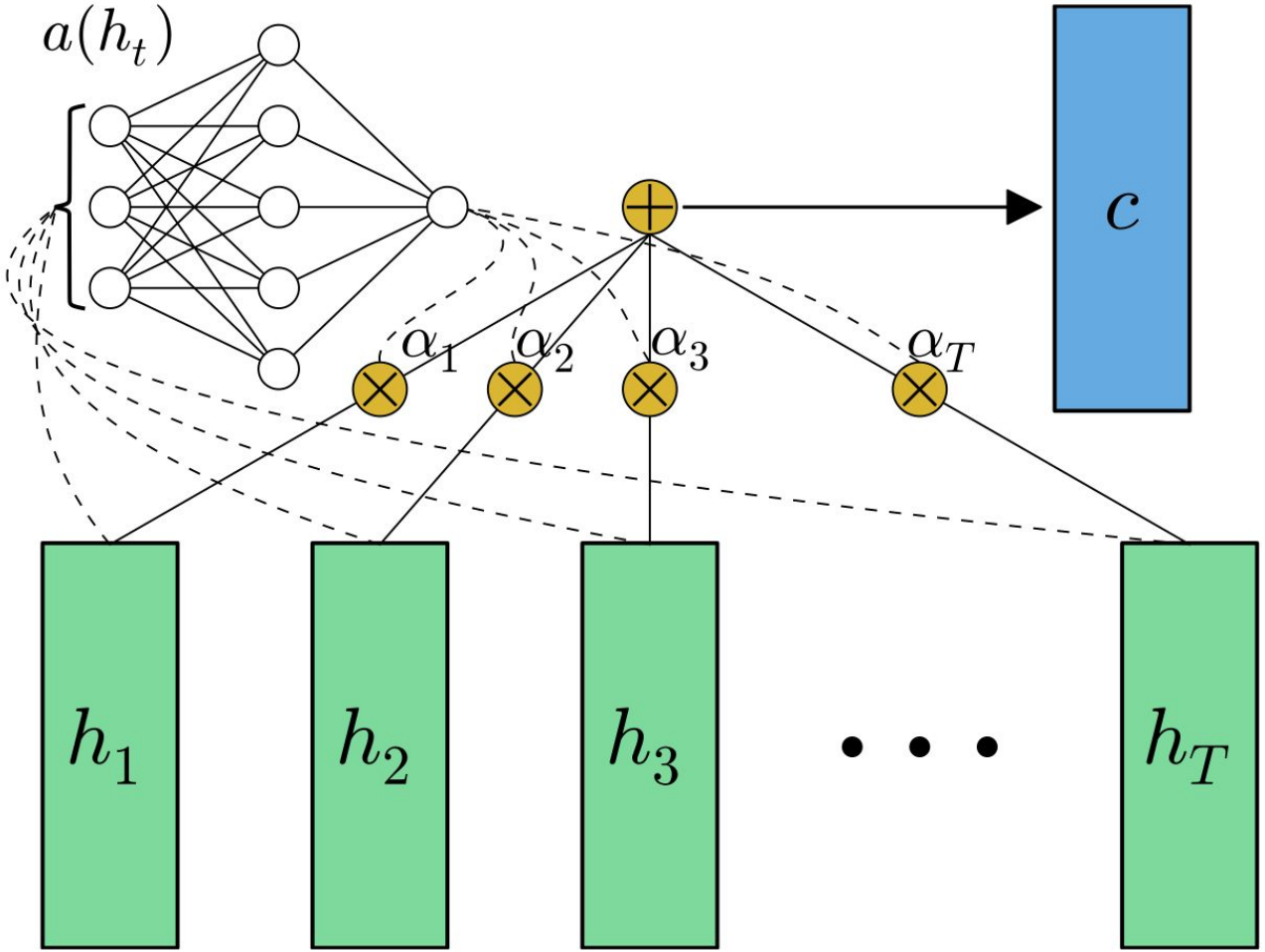




Figures from Olah & Carter, *"Attention and Augmented Recurrent Neural Networks"*



Bahdanau, Cho & Bengio, *"Neural Machine Translation by Jointly Learning to Align and Translate"*



Raffel & Ellis, "*Feed-Forward Networks with Attention Can Solve Some Long-Term Memory Problems*"

Input										Target	
0.5	-0.7	0.3	0.1	-0.2	. . .	-0.5	0.9	. . .	0.8	0.2	+	×
-1	0	1	0	0	. . .	0	1	. . .	0	-1	0.8	0.27

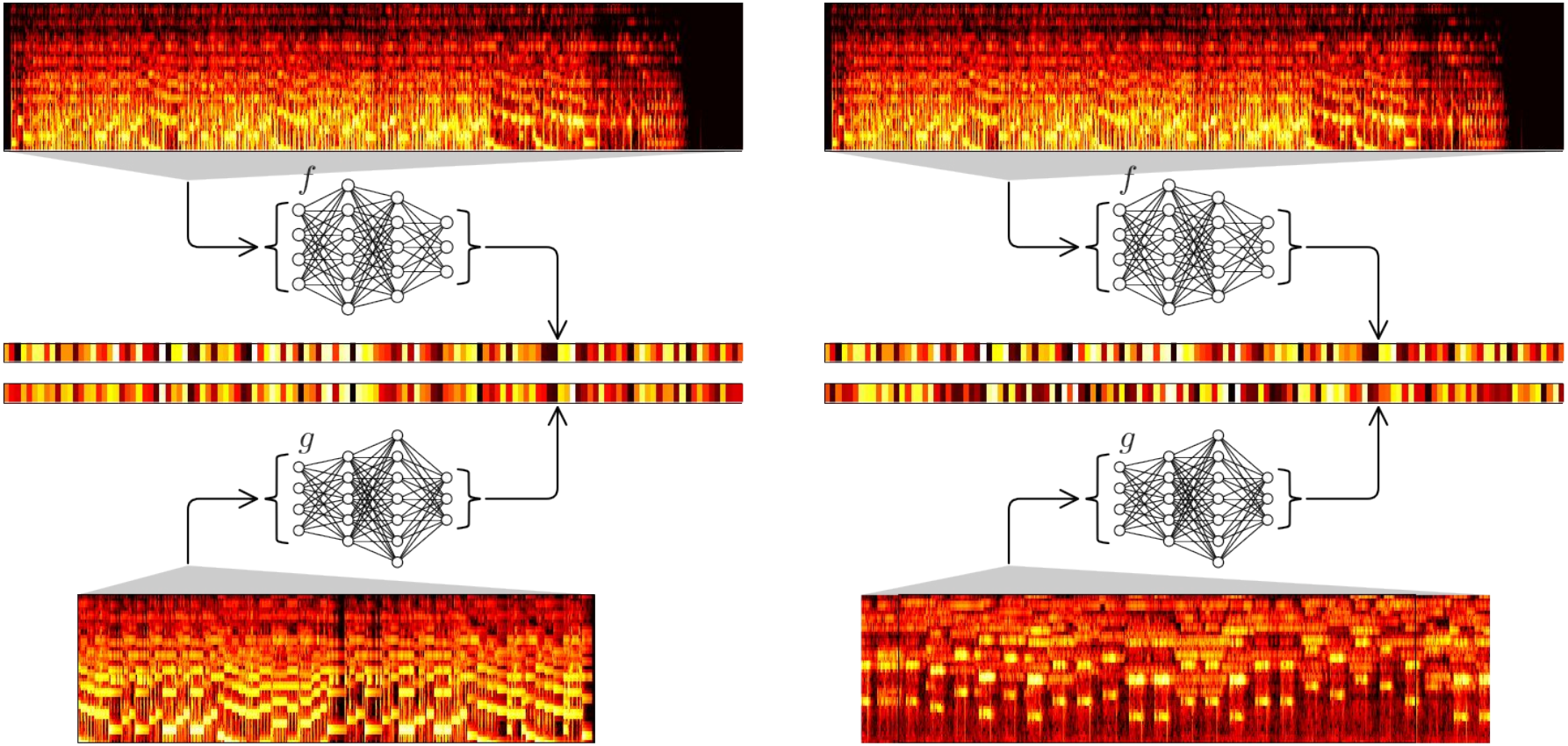
Task	Addition						
Length	50	100	500	1000	5000	10000	50-10000
Attention	1	1	1	1	2	3	99.9%
Unweighted	1	1	1	2	8	17	77.4%
Task	Multiplication						
Length	50	100	500	1000	5000	10000	50-10000
Attention	1	2	4	2	15	6	99.4%
Unweighted	2	2	8	33	89.8%	80.8%	55.5%

Raffel & Ellis, "*Feed-Forward Networks with Attention Can Solve Some Long-Term Memory Problems*"

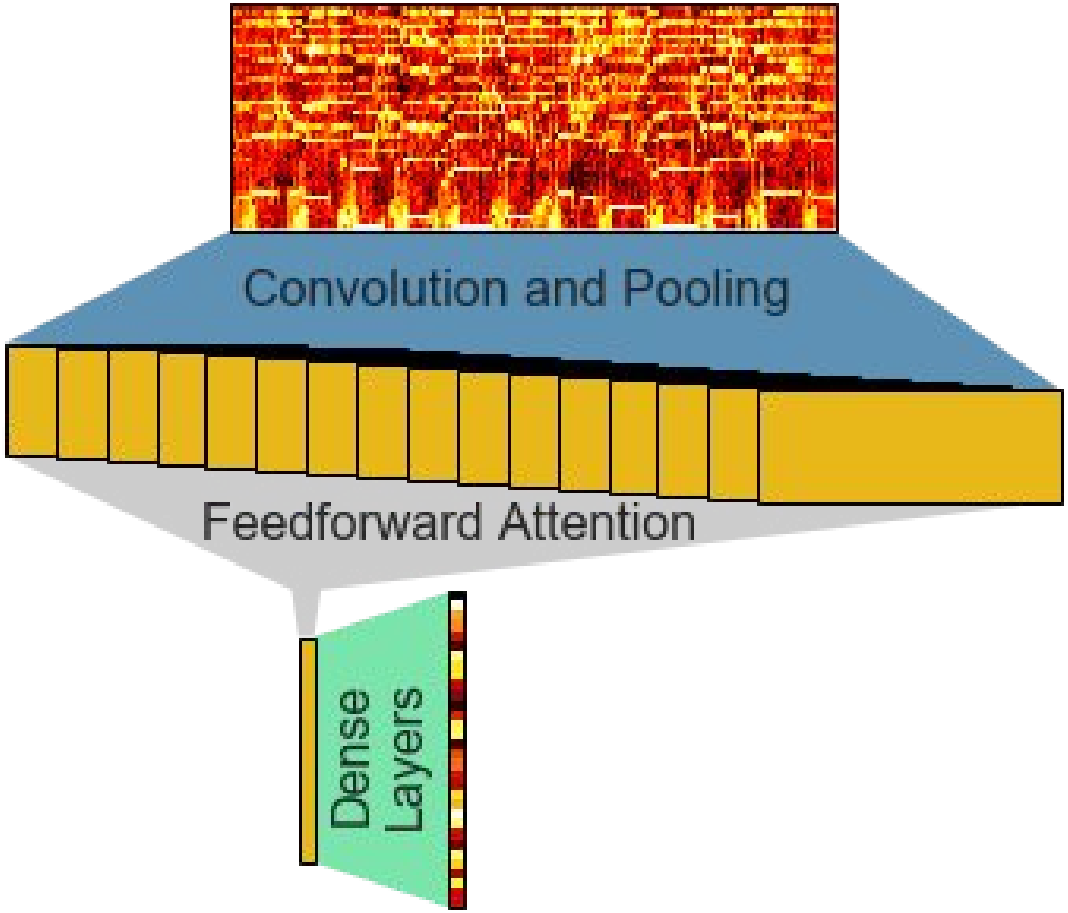
Input										Target	
0.5	-0.7	0.3	0.1	-0.2	. . .	-0.5	0.9	. . .	0.8	0.2	+	×
-1	0	1	0	0	. . .	0	1	. . .	0	-1	0.8	0.27

Task	Addition						
Length	50	100	500	1000	5000	10000	50-10000
Attention	1	1	1	1	2	3	99.9%
Unweighted	1	1	1	2	8	17	77.4%
Task	Multiplication						
Length	50	100	500	1000	5000	10000	50-10000
Attention	1	2	4	2	15	6	99.4%
Unweighted	2	2	8	33	89.8%	80.8%	55.5%

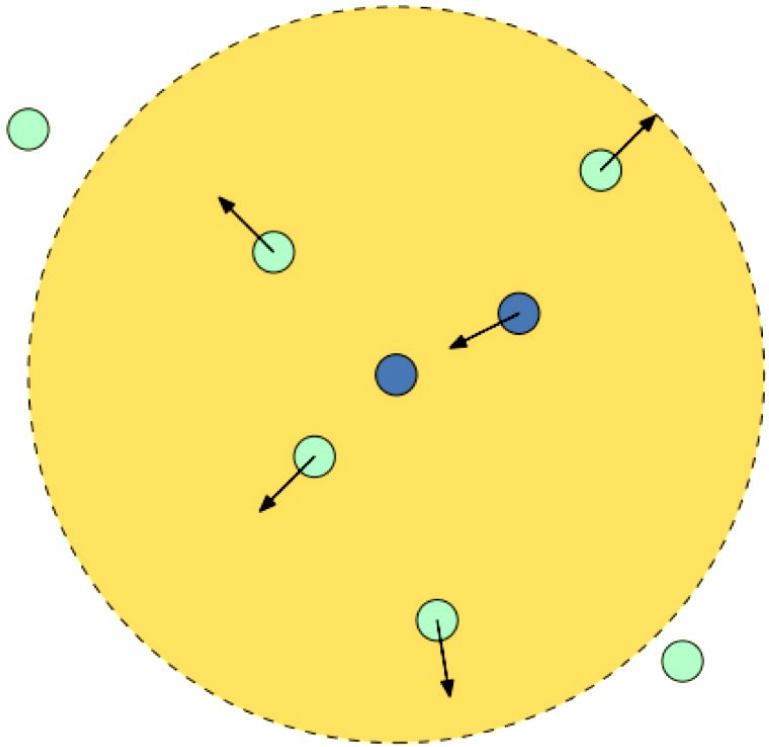
Raffel & Ellis, "*Feed-Forward Networks with Attention Can Solve Some Long-Term Memory Problems*"



Raffel & Ellis, "*Pruning Subsequence Search with Attention-Based Embedding*"

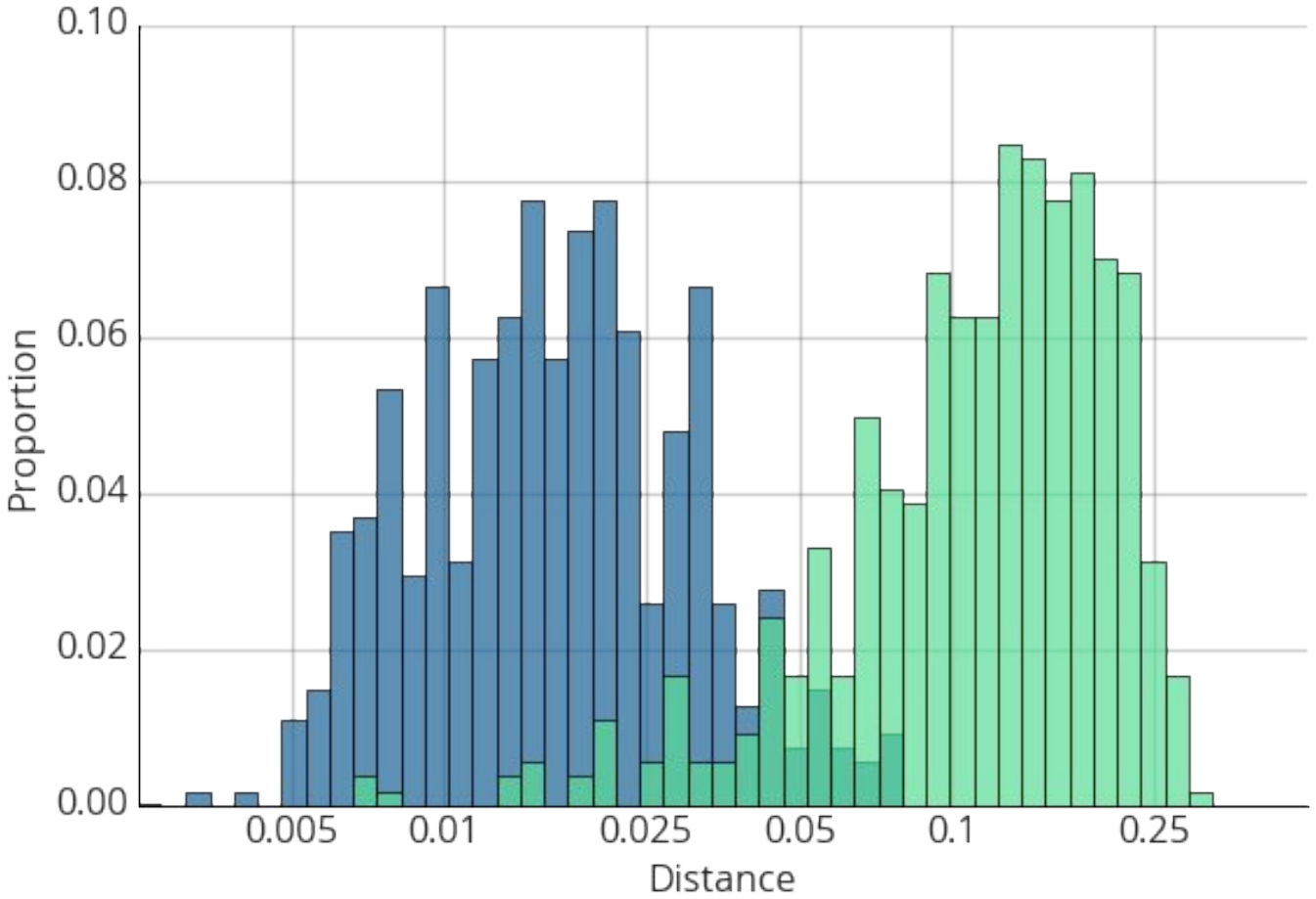


Raffel & Ellis, "*Pruning Subsequence Search with Attention-Based Embedding*"

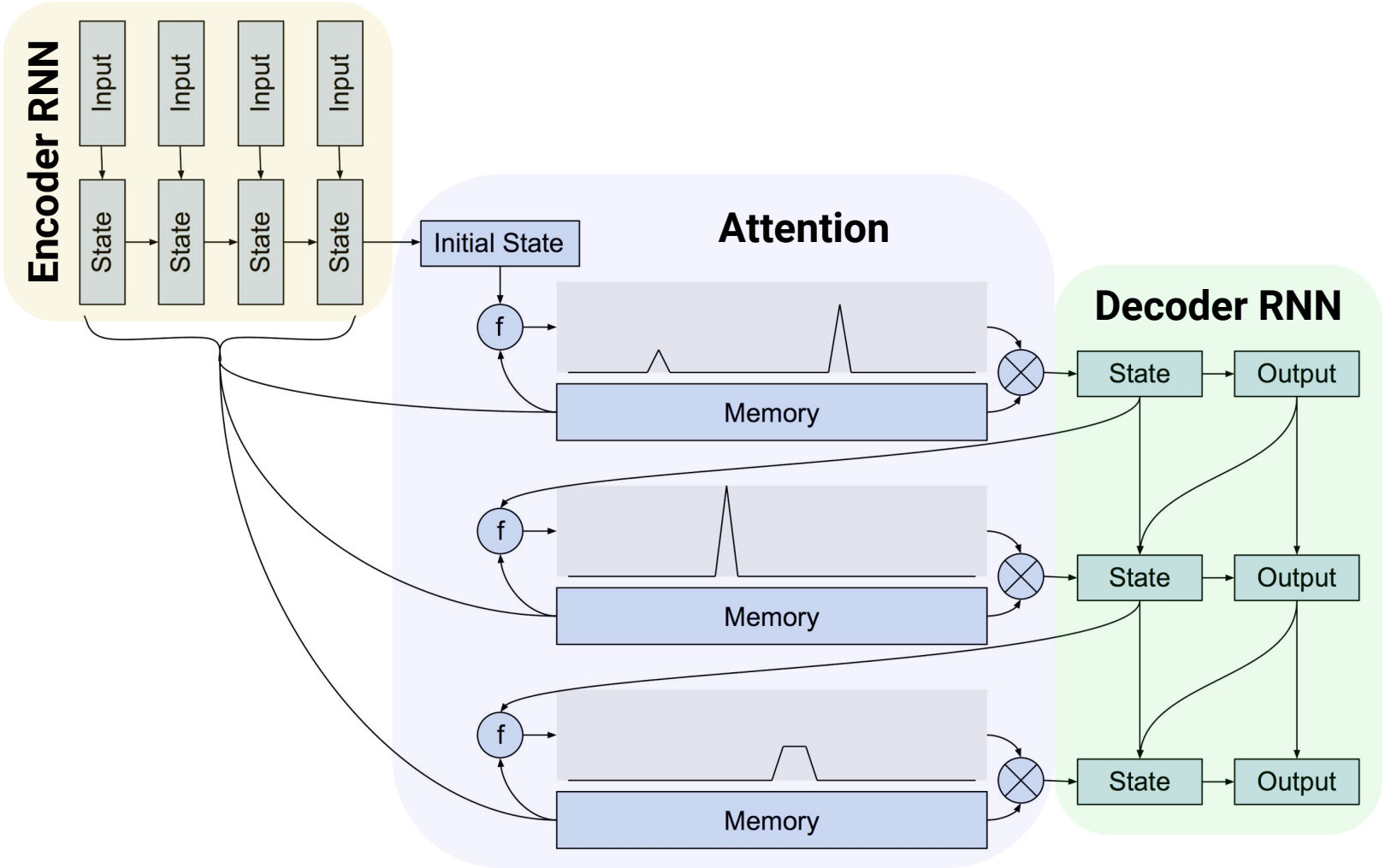


$$\mathcal{L} = \frac{1}{|\mathcal{P}|} \sum_{(x,y)\in\mathcal{P}} \|f(x) - g(y)\|_2^2 + \frac{\alpha}{|\mathcal{N}|} \sum_{(x,y)\in\mathcal{N}} \max(0, m - \|f(x) - g(y)\|_2)^2$$

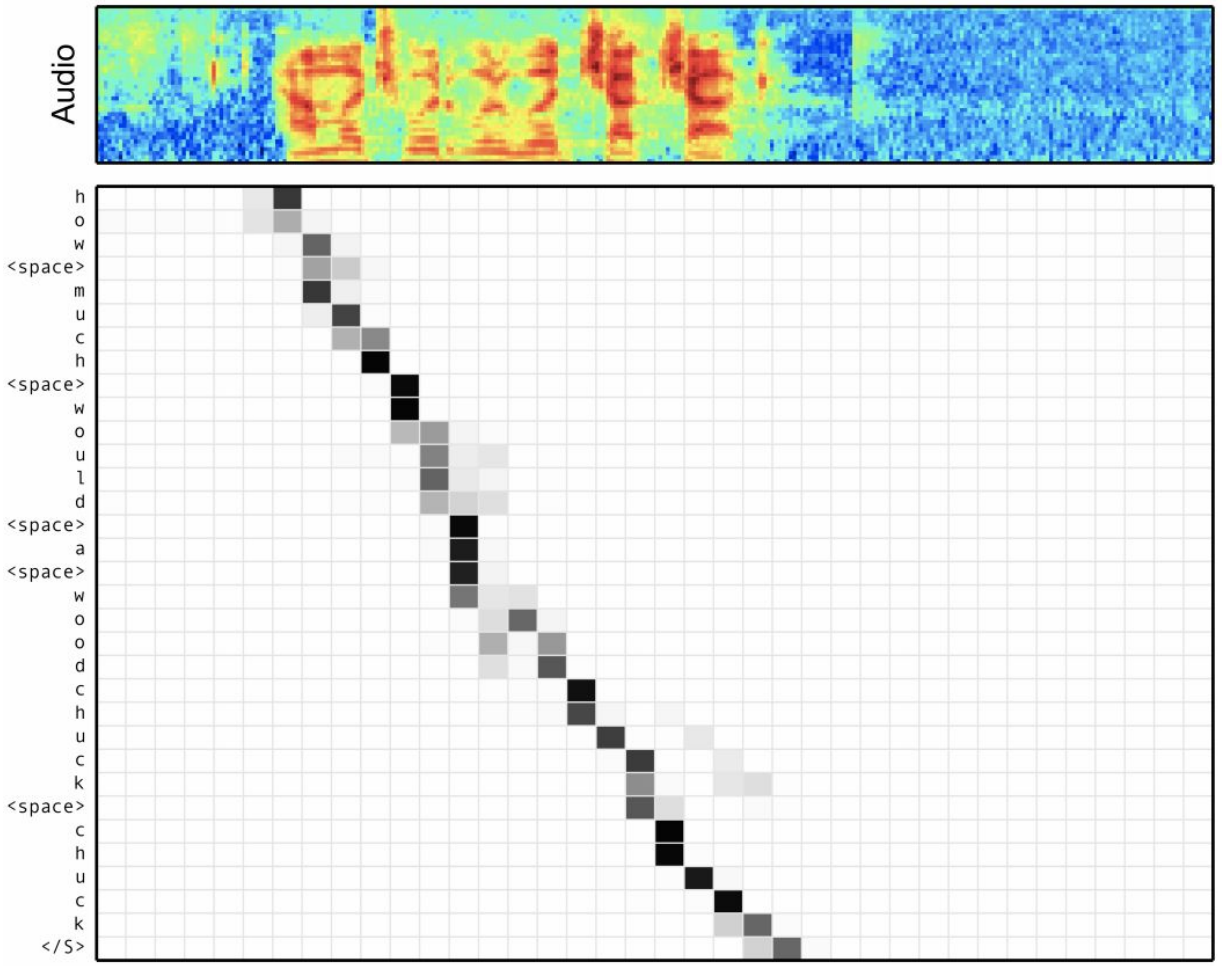
Raffel & Ellis, "*Pruning Subsequence Search with Attention-Based Embedding*"



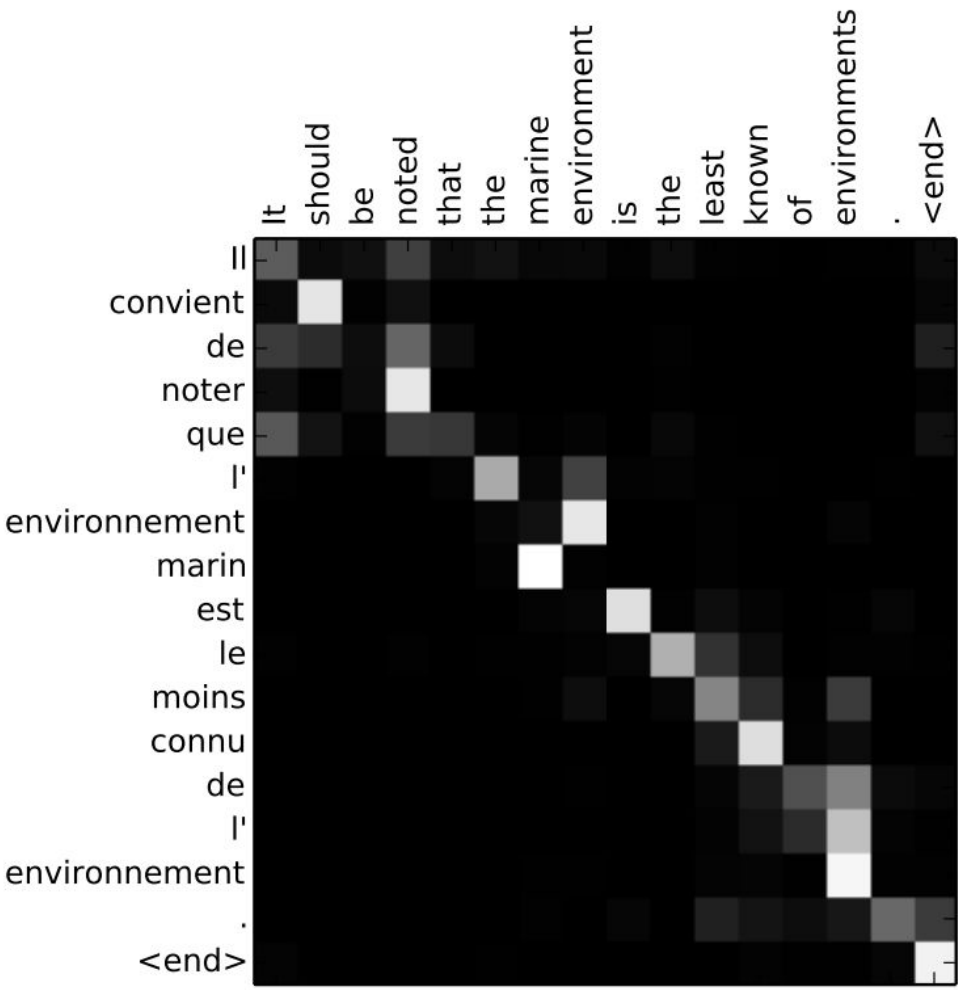
Raffel & Ellis, "*Pruning Subsequence Search with Attention-Based Embedding*"



Bahdanau, Cho & Bengio, *"Neural Machine Translation by Jointly Learning to Align and Translate"*

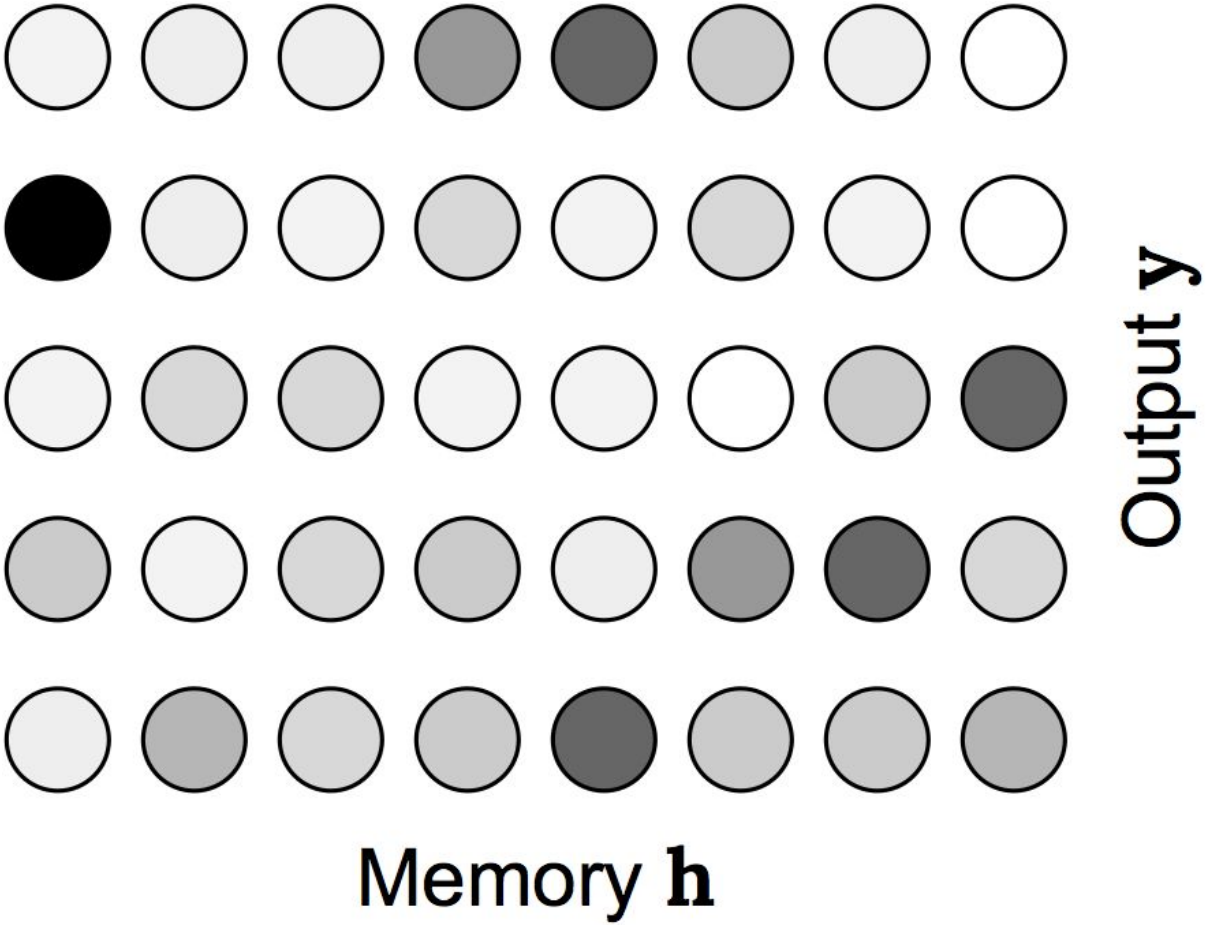


Chan, Jaitly, Le & Vinyals, *"Listen, Attend and Spell"*

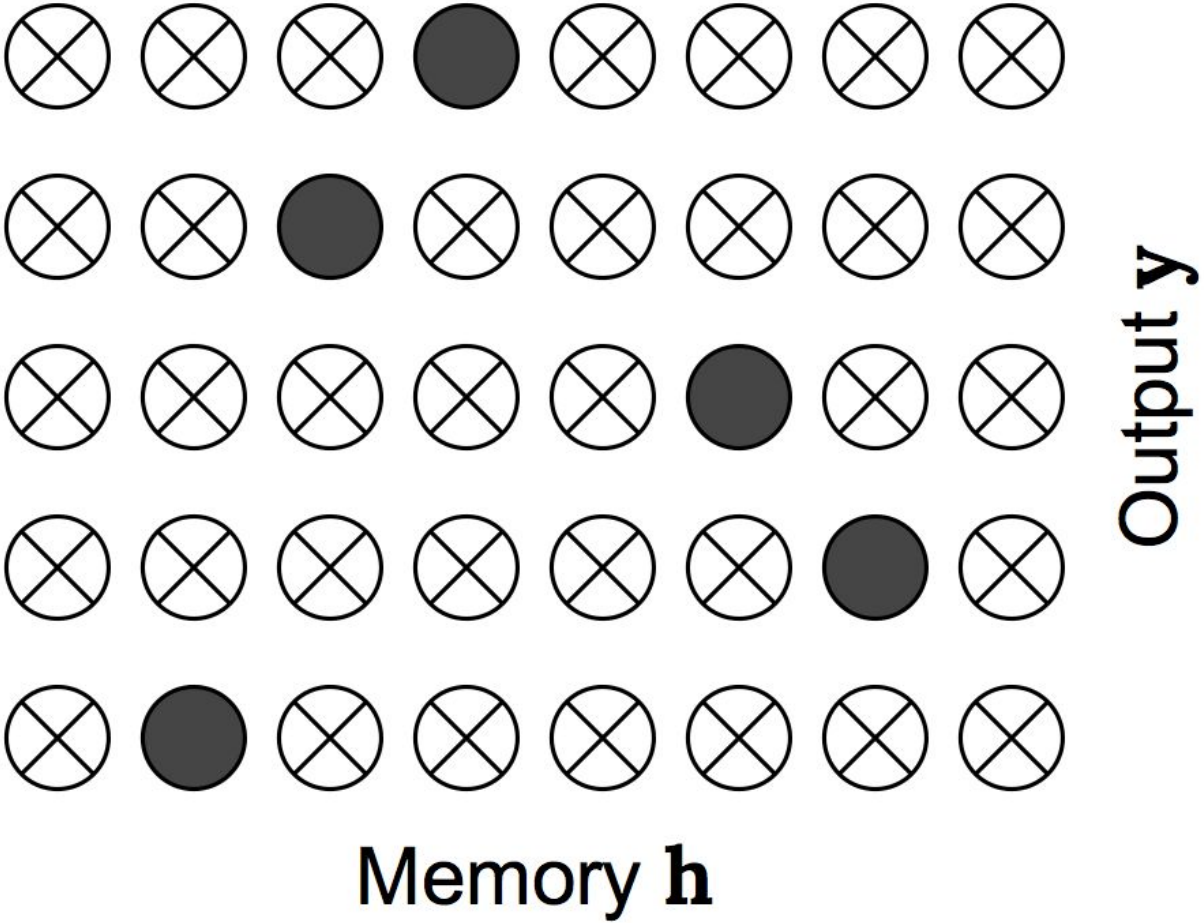


Bahdanau, Cho & Bengio, *"Neural Machine Translation by Jointly Learning to Align and Translate"*

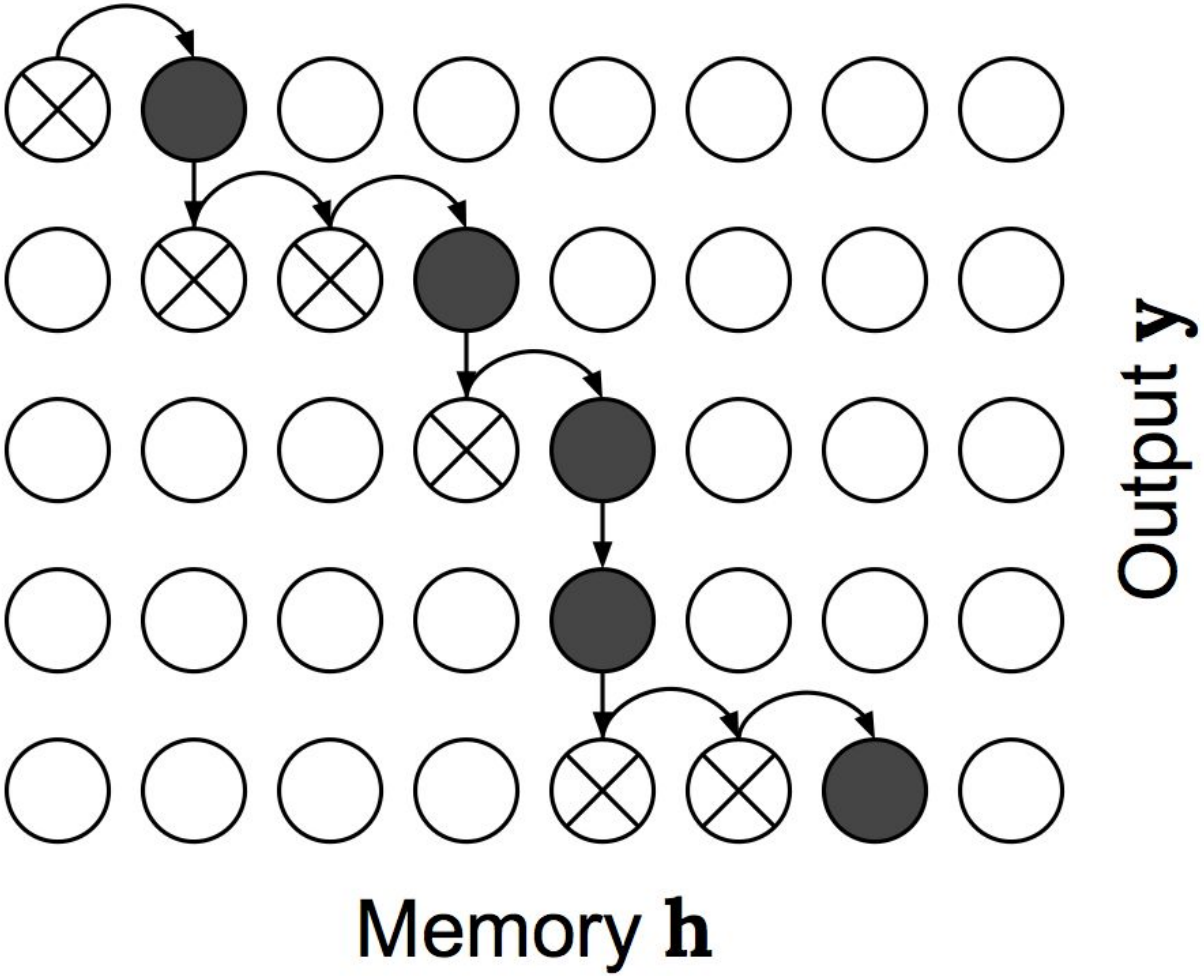
Xu, Ba, Kiros, Cho & Courville, *"Show, attend and tell: Neural image caption generation with visual attention"*



Raffel, Luong, Liu, Weiss & Eck, "*Online and Linear-Time Attention by Enforcing Monotonic Alignments*"



Raffel, Luong, Liu, Weiss & Eck, "*Online and Linear-Time Attention by Enforcing Monotonic Alignments*"



Raffel, Luong, Liu, Weiss & Eck, "*Online and Linear-Time Attention by Enforcing Monotonic Alignments*"

$$\text{energy}_{i,j} = \text{EnergyFunction}(\text{state}_{i-1}, \text{memory}_j)$$

$$\text{attention}_{i,j} = \exp(\text{energy}_{i,j}) \Big/ \sum_{k=1}^{T} \exp(\text{energy}_{i,k})$$

$$\text{energy}_{i,j} = \text{EnergyFunction}(\text{state}_{i-1}, \text{memory}_j)$$

$$\text{select}_{i,j} = \sigma(\text{energy}_{i,j})$$

$$\text{attention}_{i,j} = \text{select}_{i,j} \sum_{k=1}^{j} \left(\text{attention}_{i-1,k} \prod_{l=k}^{j-1} (1 - \text{select}_{i,l}) \right)$$

Raffel, Luong, Liu, Weiss & Eck, "*Online and Linear-Time Attention by Enforcing Monotonic Alignments*"

$$\text{energy}_{i,j} = \text{EnergyFunction}(\text{state}_{i-1}, \text{memory}_j)$$

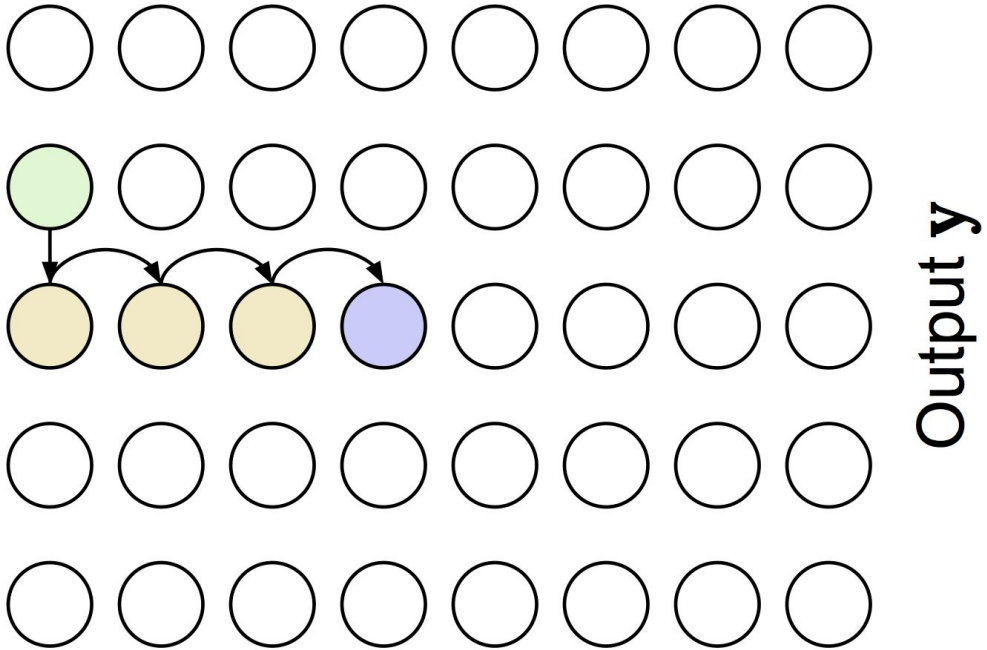
$$\text{attention}_{i,j} = \exp(\text{energy}_{i,j}) \Big/ \sum_{k=1}^{T} \exp(\text{energy}_{i,k})$$

$$\text{energy}_{i,j} = \text{EnergyFunction}(\text{state}_{i-1}, \text{memory}_j)$$

$$\text{select}_{i,j} = \sigma(\text{energy}_{i,j})$$

$$\text{attention}_{i,j} = \text{select}_{i,j} \sum_{k=1}^{j} \left(\text{attention}_{i-1,k} \prod_{l=k}^{j-1} (1 - \text{select}_{i,l}) \right)$$

Raffel, Luong, Liu, Weiss & Eck, "*Online and Linear-Time Attention by Enforcing Monotonic Alignments*"



$$\text{attention}_{i,j} = \text{select}_{i,j} \sum_{k=1}^{j} \left(\text{attention}_{i-1,k} \prod_{l=k}^{j-1} (1 - \text{select}_{i,l}) \right)$$

Raffel, Luong, Liu, Weiss & Eck, "*Online and Linear-Time Attention by Enforcing Monotonic Alignments*"

$$\text{attention}_{i,j} = \text{select}_{i,j} \sum_{k=1}^{j} \left(\text{attention}_{i-1,k} \prod_{l=k}^{j-1} (1 - \text{select}_{i,l}) \right)$$

$$\text{attention}_{i,j} = \text{select}_{i,j} \left((1 - \text{select}_{i,j-1}) \frac{\text{attention}_{i,j-1}}{\text{select}_{i,j-1}} + \text{attention}_{i-1,j} \right)$$

$$\text{attention}_i = \text{select}_i \, \texttt{cumprod}(1 - \text{select}_i) \, \texttt{cumsum} \left(\frac{\text{attention}_{i-1}}{\texttt{cumprod}(1 - \text{select}_i)} \right)$$

Raffel, Luong, Liu, Weiss & Eck, "*Online and Linear-Time Attention by Enforcing Monotonic Alignments*"

$$\text{attention}_{i,j} = \text{select}_{i,j} \sum_{k=1}^{j} \left(\text{attention}_{i-1,k} \prod_{l=k}^{j-1} (1 - \text{select}_{i,l}) \right)$$

$$\text{attention}_{i,j} = \text{select}_{i,j} \left((1 - \text{select}_{i,j-1}) \frac{\text{attention}_{i,j-1}}{\text{select}_{i,j-1}} + \text{attention}_{i-1,j} \right)$$

$$\text{attention}_i = \text{select}_i \; \texttt{cumprod}(1 - \text{select}_i) \; \texttt{cumsum} \left(\frac{\text{attention}_{i-1}}{\texttt{cumprod}(1 - \text{select}_i)} \right)$$

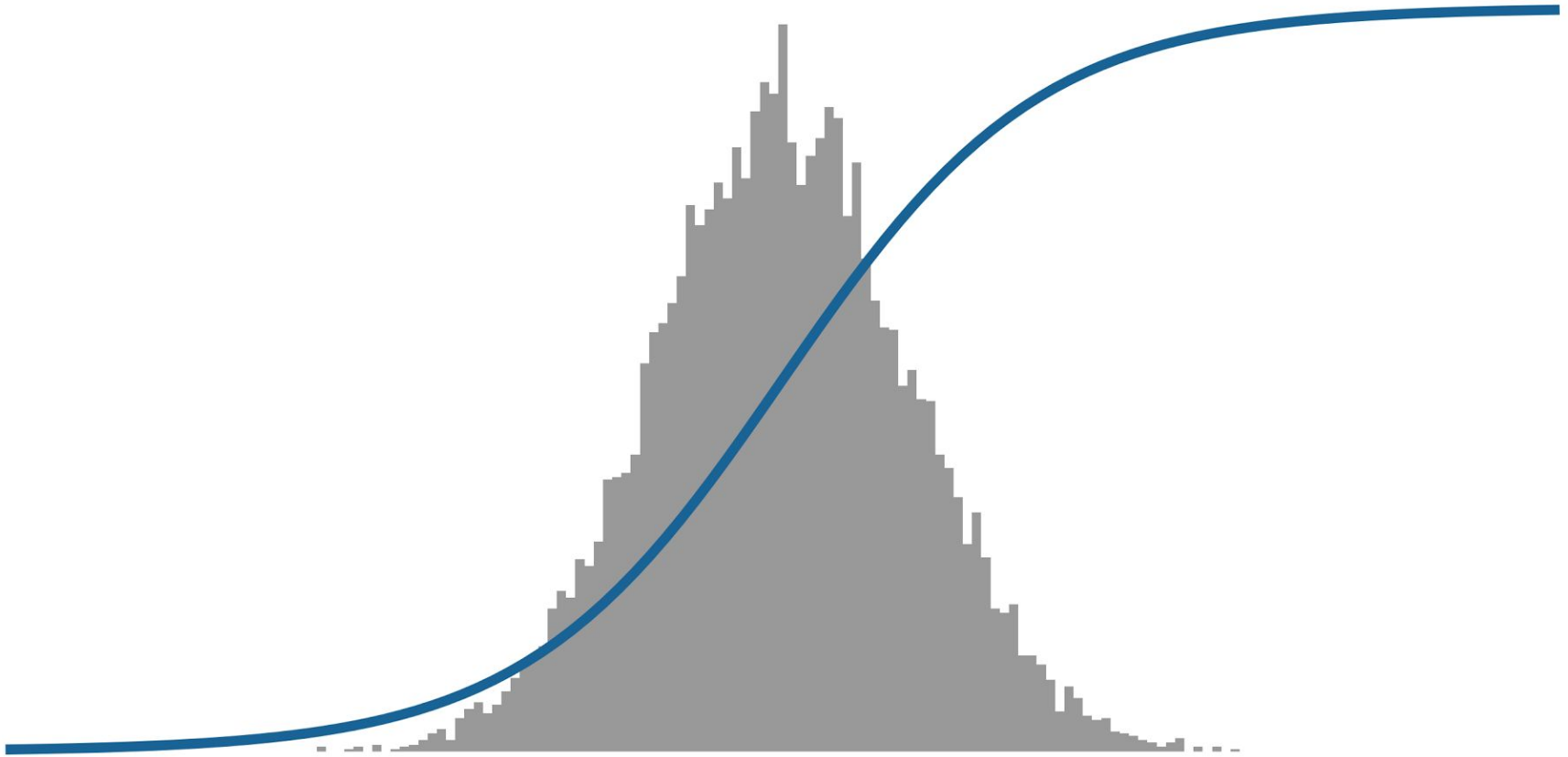
Raffel, Luong, Liu, Weiss & Eck, "*Online and Linear-Time Attention by Enforcing Monotonic Alignments*"

$$\text{attention}_{i,j} = \text{select}_{i,j} \sum_{k=1}^{j} \left(\text{attention}_{i-1,k} \prod_{l=k}^{j-1} (1 - \text{select}_{i,l}) \right)$$

$$\text{attention}_{i,j} = \text{select}_{i,j} \left((1 - \text{select}_{i,j-1}) \frac{\text{attention}_{i,j-1}}{\text{select}_{i,j-1}} + \text{attention}_{i-1,j} \right)$$

$$\text{attention}_i = \text{select}_i \; \mathbf{cumprod}(1 - \text{select}_i) \; \mathbf{cumsum} \left(\frac{\text{attention}_{i-1}}{\mathbf{cumprod}(1 - \text{select}_i)} \right)$$

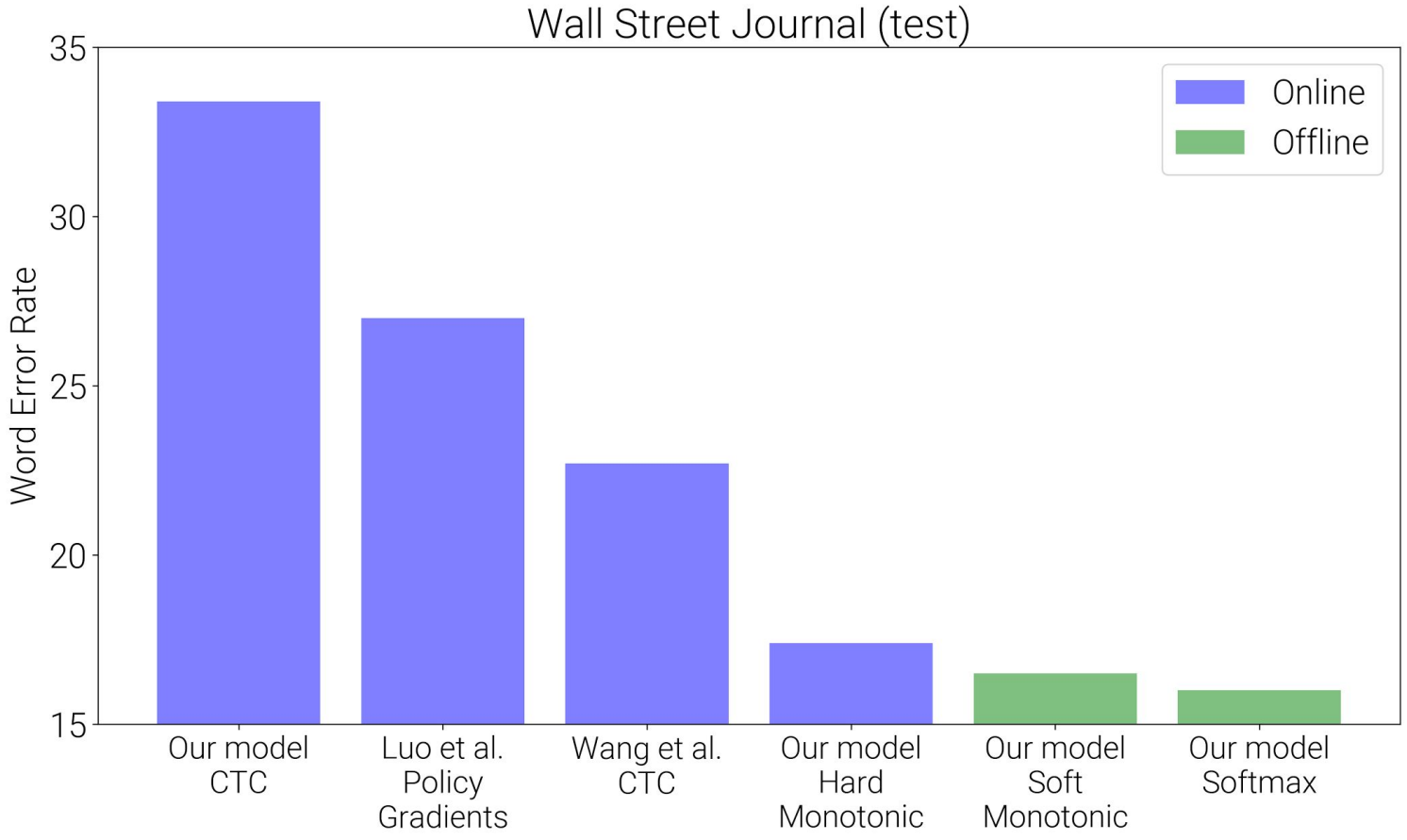
Raffel, Luong, Liu, Weiss & Eck, "*Online and Linear-Time Attention by Enforcing Monotonic Alignments*"



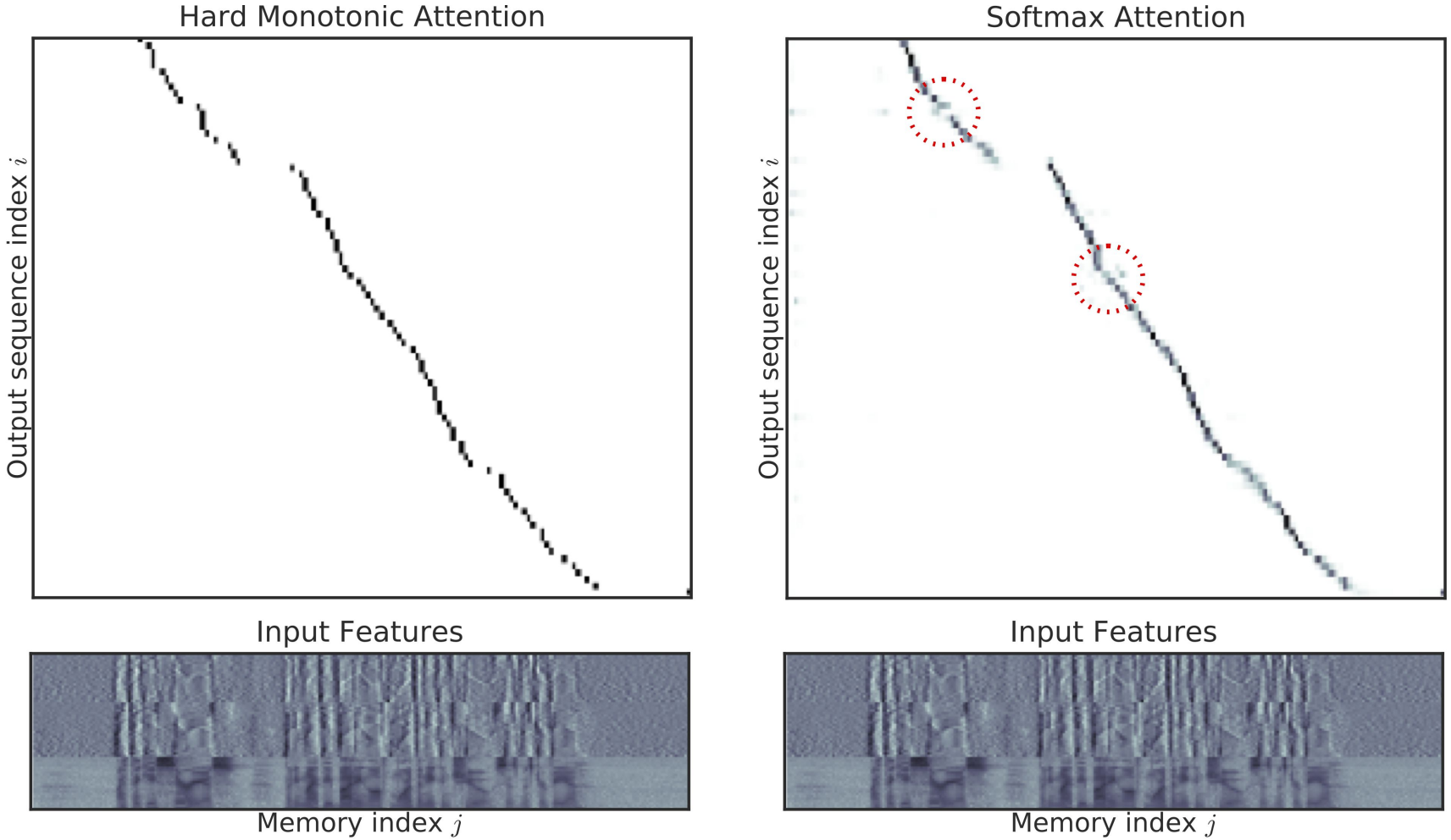
Frey, "*Continuous Sigmoidal Belief Networks Trained Using Slice Sampling*"

Salakhutdinov & Hinton, *"Semantic Hashing"*

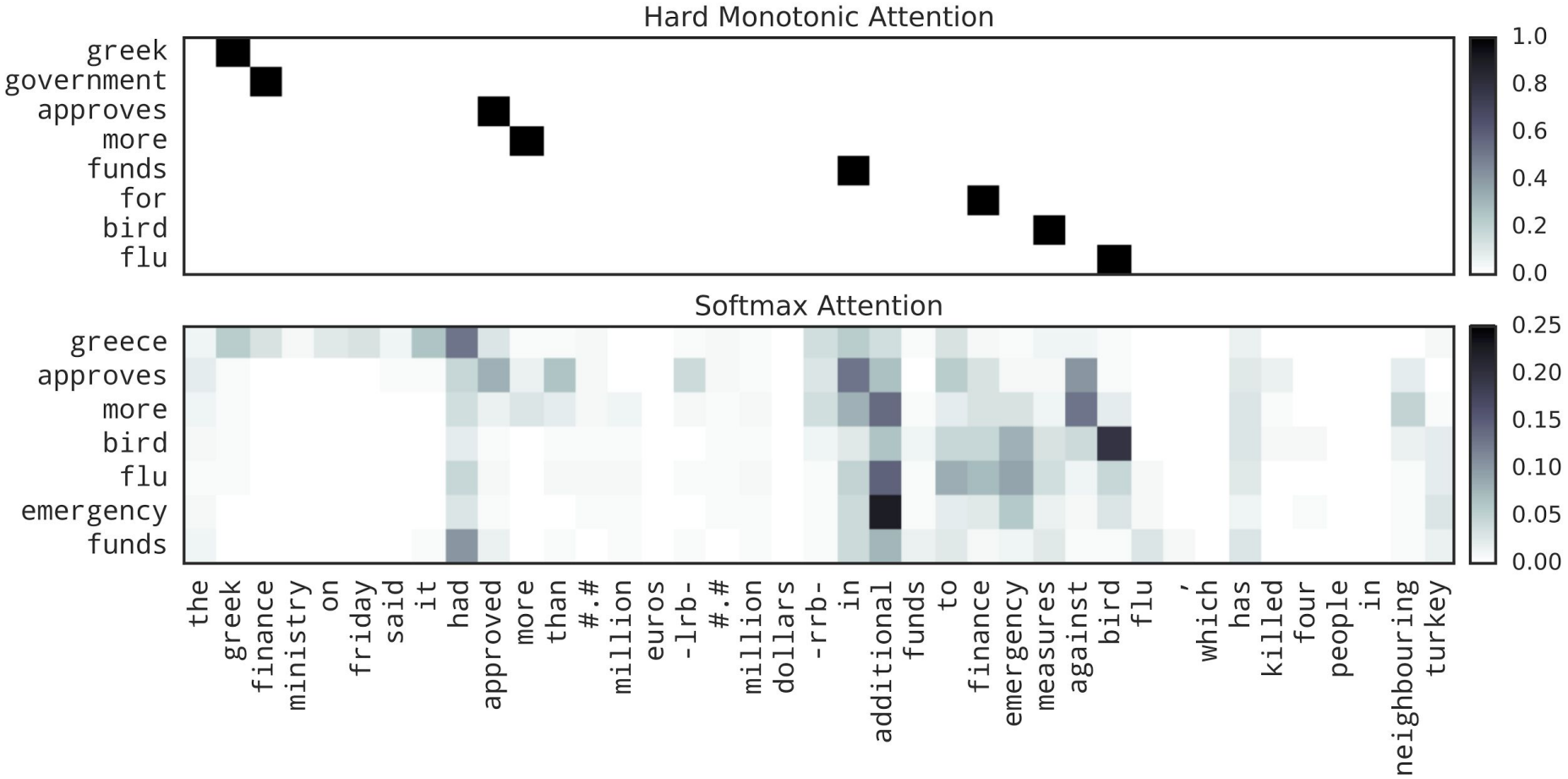
Foerster, Assael, de Freitas & Whiteson, *"Learning to Communicate with Deep Multi-Agent Reinforcement Learning"*



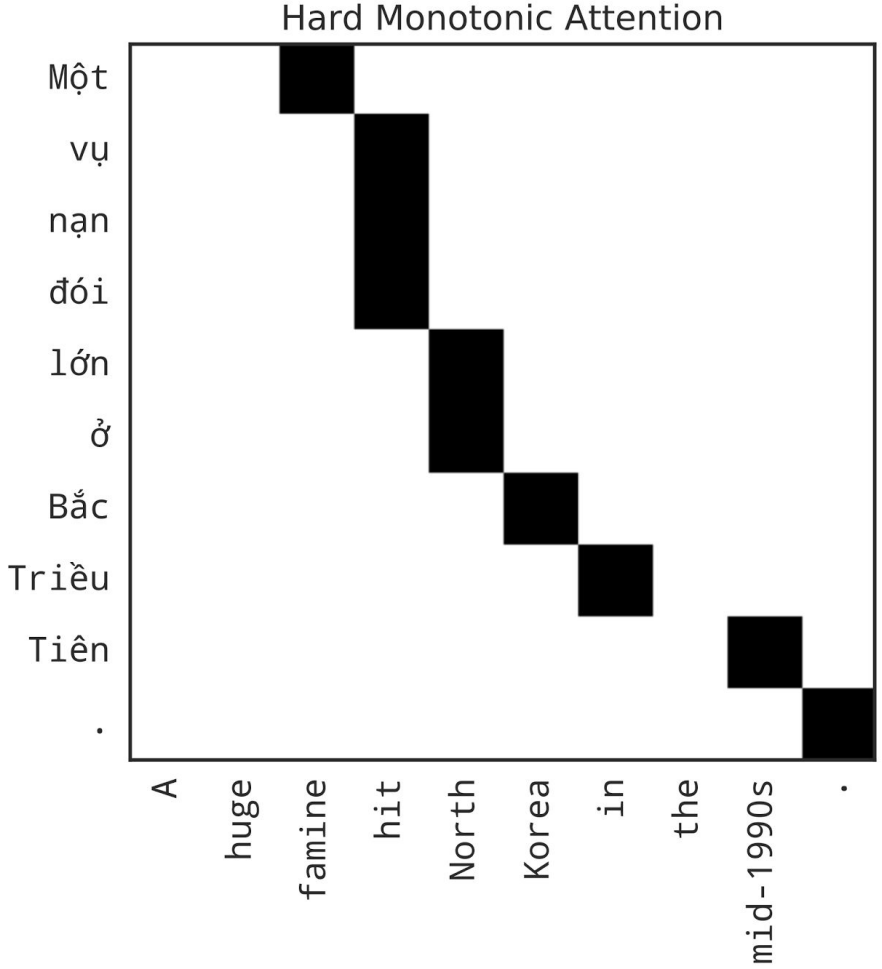
Raffel, Luong, Liu, Weiss & Eck, "*Online and Linear-Time Attention by Enforcing Monotonic Alignments*"

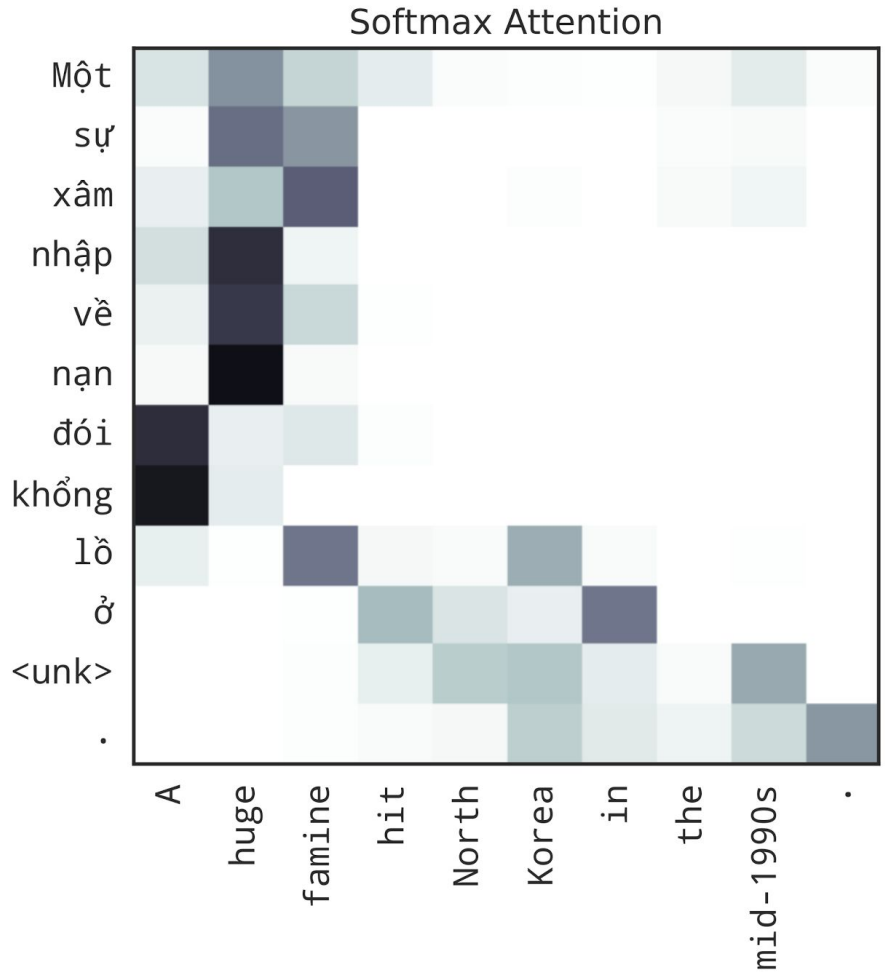


Raffel, Luong, Liu, Weiss & Eck, "*Online and Linear-Time Attention by Enforcing Monotonic Alignments*"

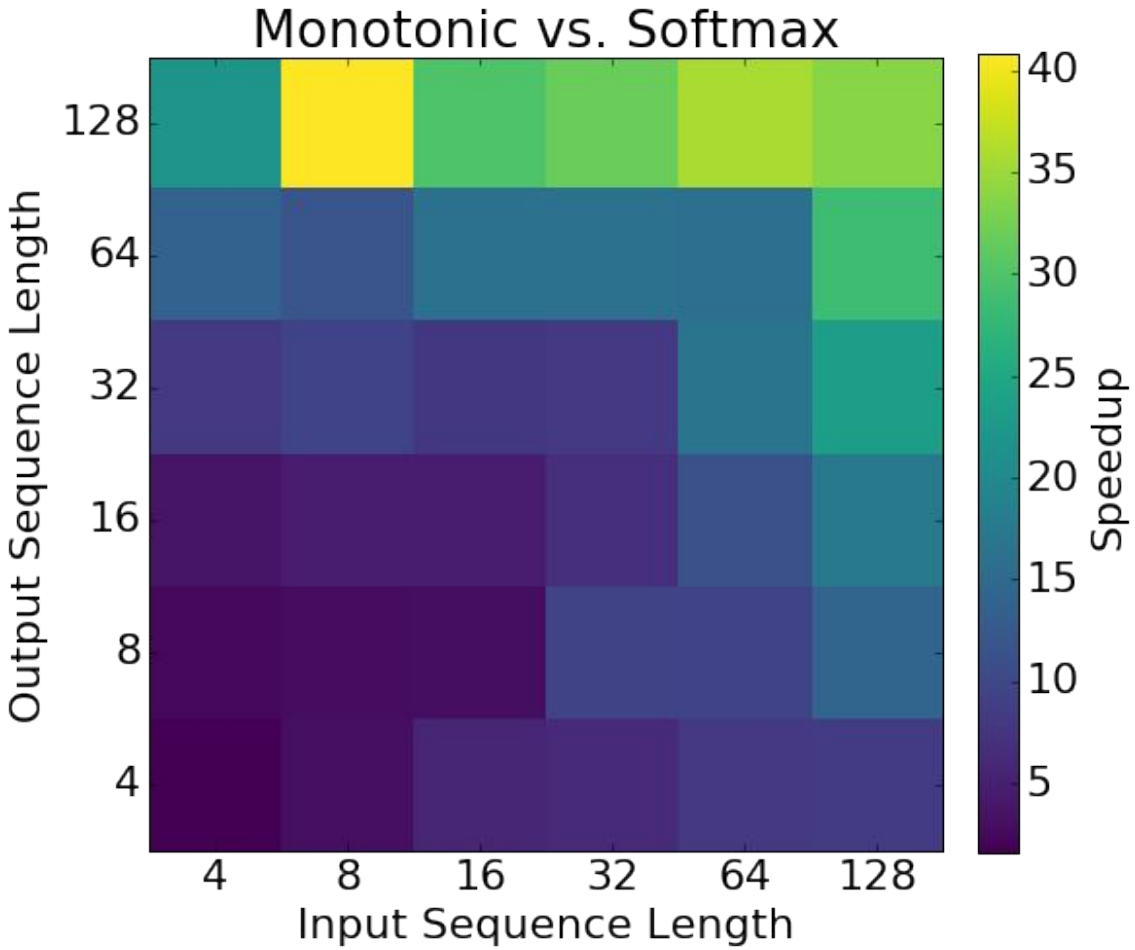


Raffel, Luong, Liu, Weiss & Eck, "*Online and Linear-Time Attention by Enforcing Monotonic Alignments*"

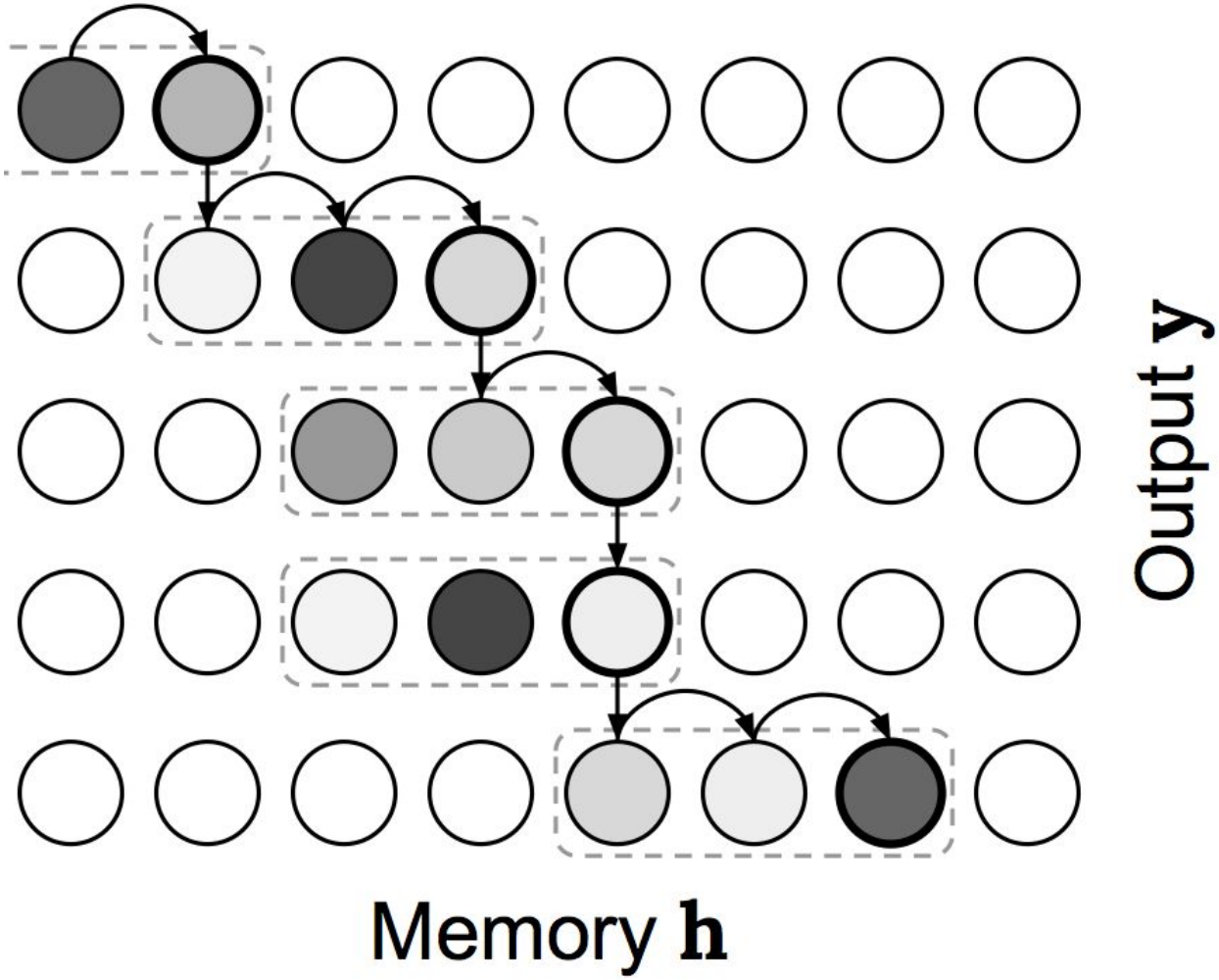




Raffel, Luong, Liu, Weiss & Eck, "*Online and Linear-Time Attention by Enforcing Monotonic Alignments*"



Raffel, Luong, Liu, Weiss & Eck, "*Online and Linear-Time Attention by Enforcing Monotonic Alignments*"

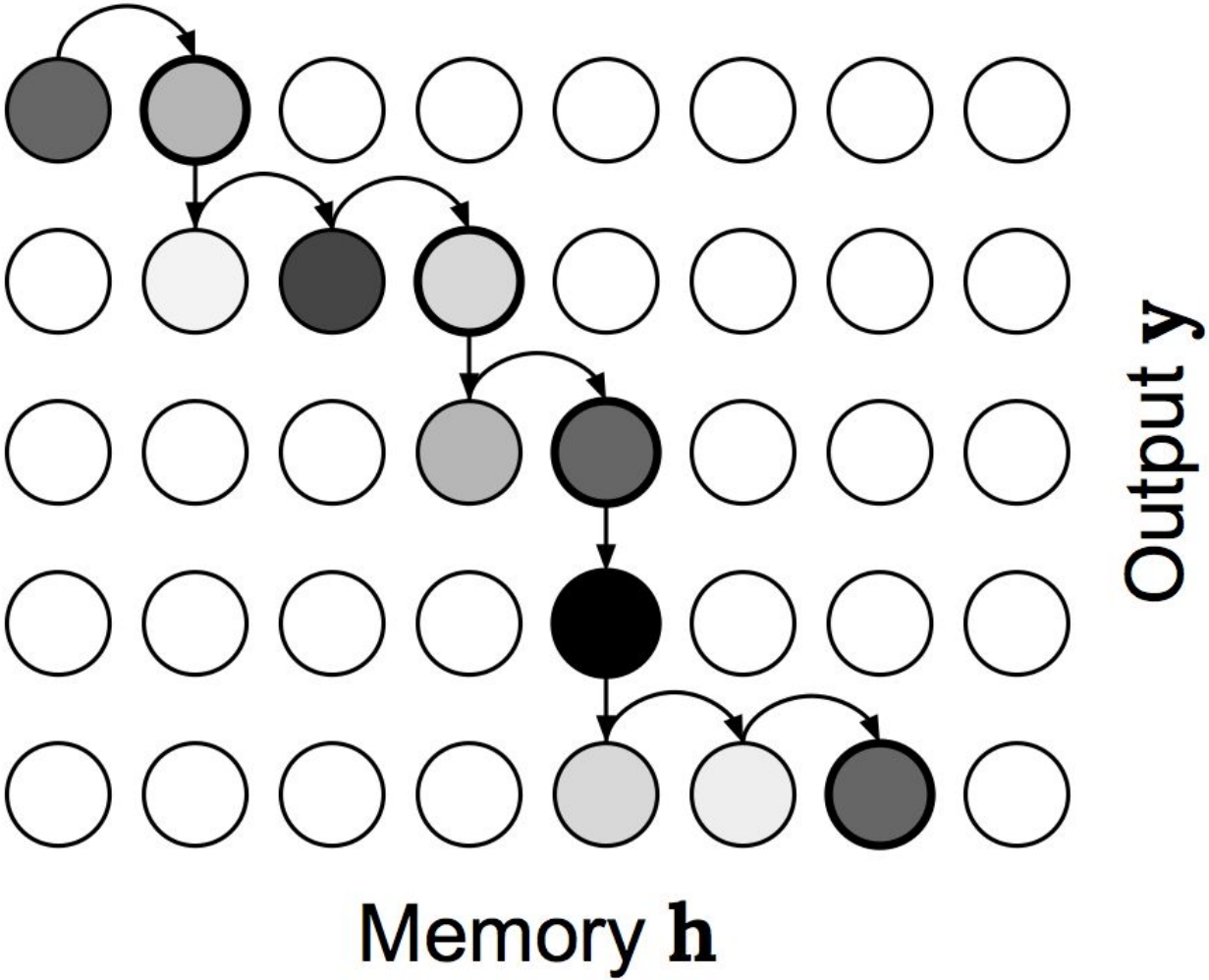

Output y
Memory h

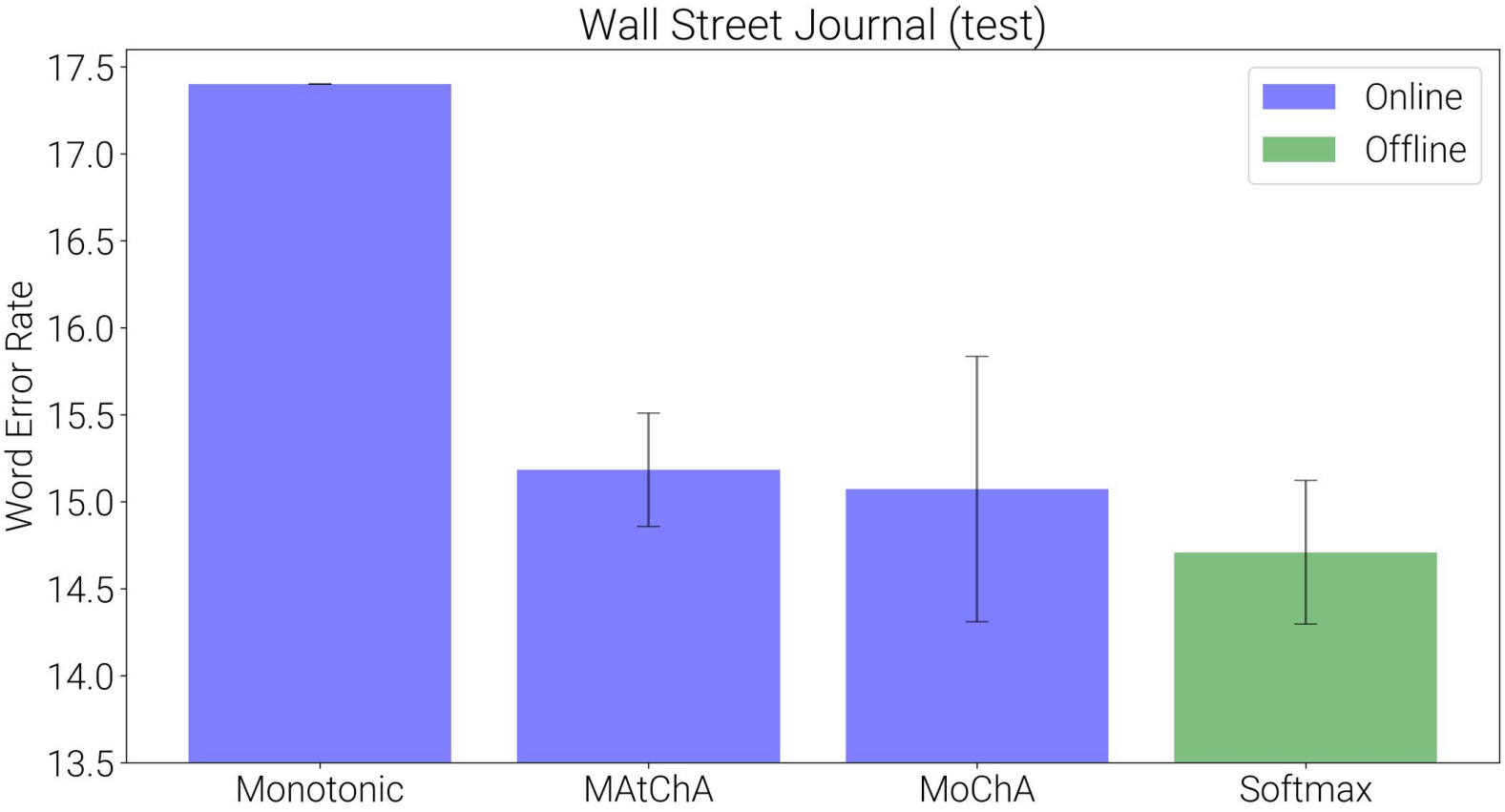
$$\text{select}_{i,j} = \sigma\left(\text{MonotonicEnergy}(\text{state}_{i-1}, \text{memory}_j)\right)$$

$$\text{monotonic}_{i,j} = \text{select}_{i,j} \sum_{k=1}^{j} \left(\text{monotonic}_{i-1,k} \prod_{l=k}^{j-1} (1 - \text{select}_{i,l}) \right)$$

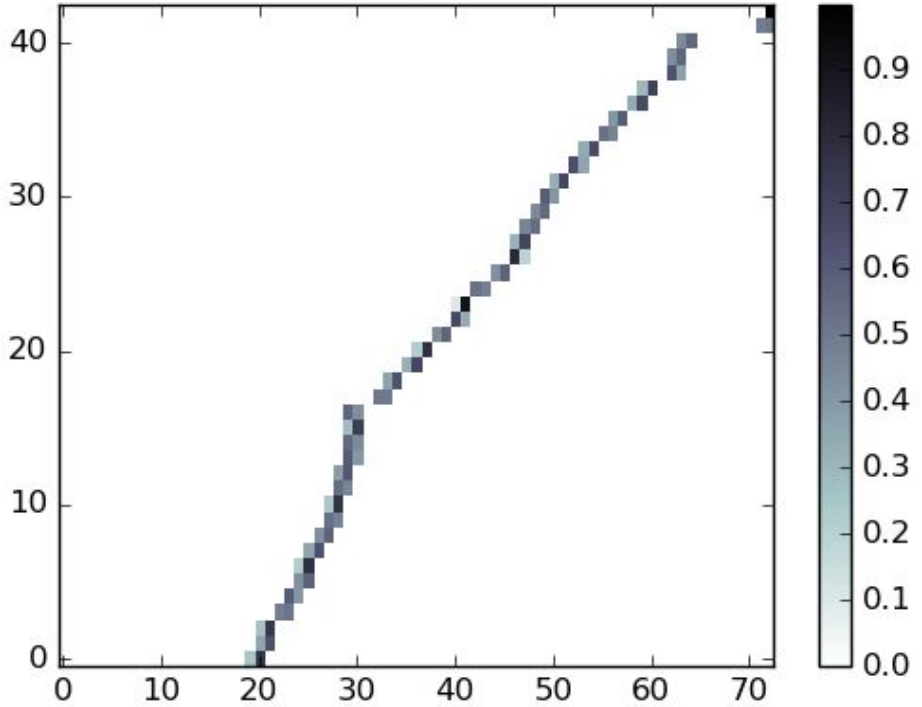
$$\text{logits}_{i,j} = \text{SoftmaxEnergy}(\text{state}_{i-1}, \text{memory}_j)$$

$$\text{attention}_{i,j} = \sum_{k=j}^{j+N} \text{monotonic}_{i,k} \, \text{softmax}(\text{logits}_{i,k-N:k})$$

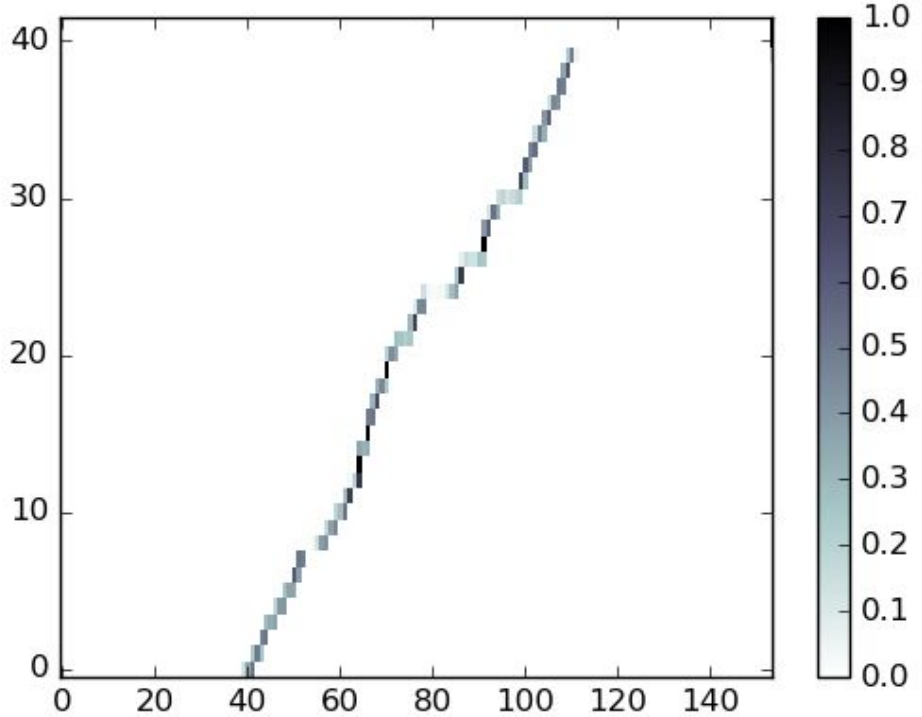

Output y
Memory h


Wall Street Journal (test)
Word Error Rate
17.5
17.0
16.5
16.0
15.5
15.0
14.5
14.0
13.5
Online
Offline
Monotonic
MAtChA
MoChA
Softmax

MoChA (chunk size=2)



MAtChA



Collaborators


Daniel P. W. Ellis


Chung-Cheng Chiu


Thang Luong


Peter J. Liu


Ron J. Weiss


Douglas Eck

References

Raffel & Ellis, "*Feed-Forward Networks with Attention Can Solve Some Long-Term Memory Problems*", ICLR 2016.
Raffel & Ellis, "*Pruning Subsequence Search with Attention-Based Embedding*", ICASSP 2016.
Raffel, Luong, Liu, Weiss & Eck, "*Online and Linear-Time Attention by Enforcing Monotonic Alignments*", ICML 2017.

Thanks!