Online and Linear-Time Attention by Enforcing Monotonic Alignments

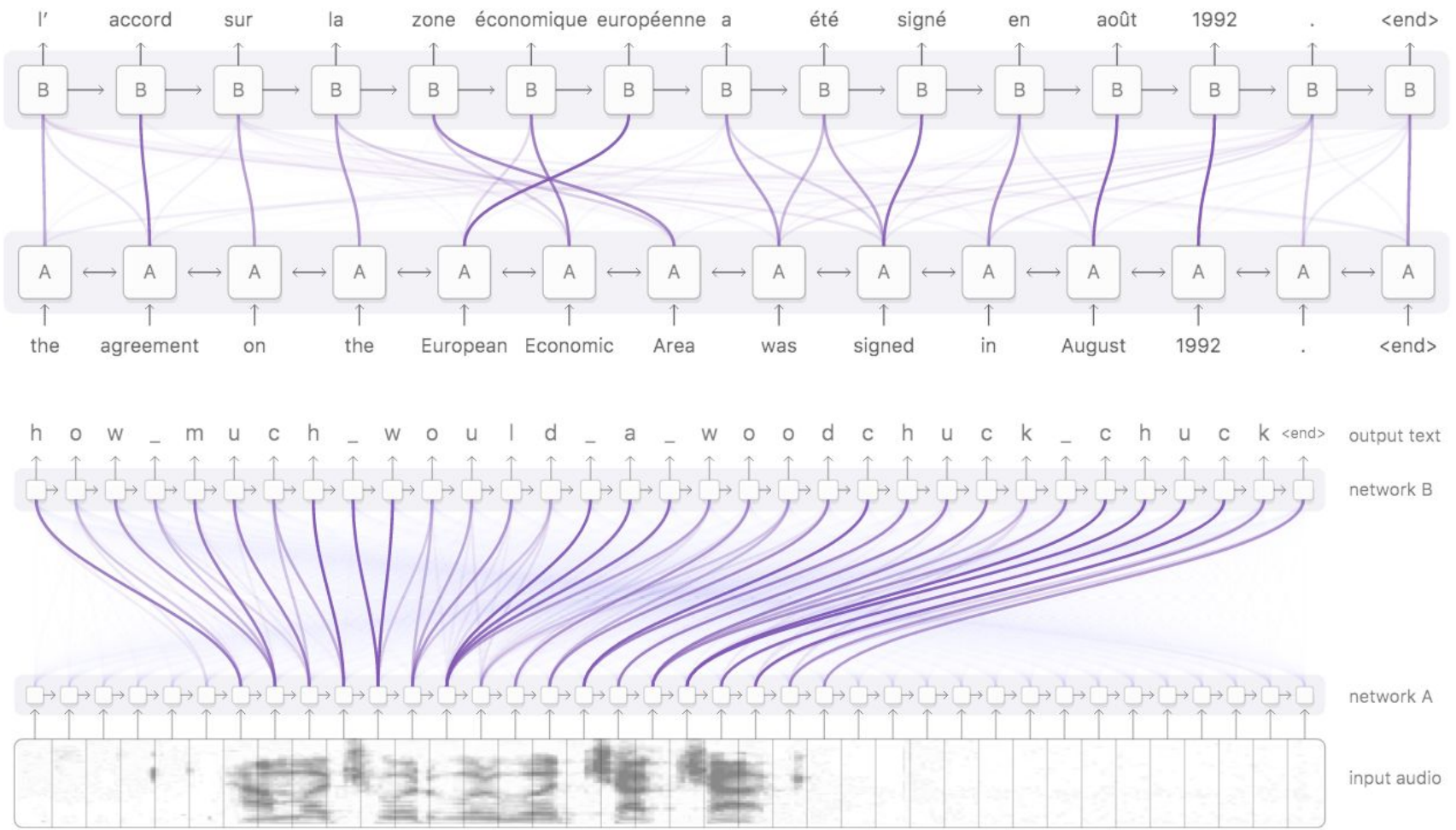
Colin Raffel


Thang Luong

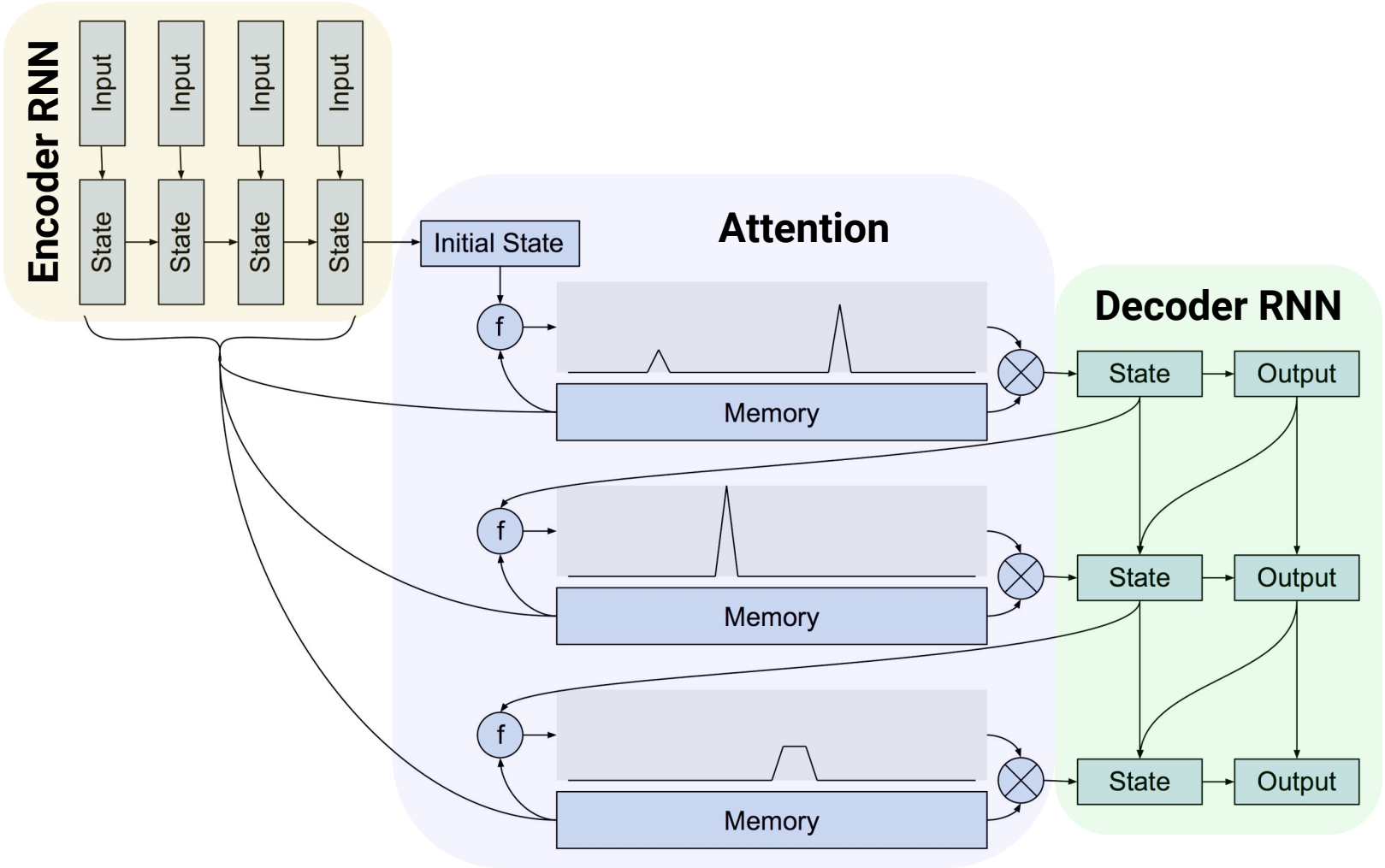

Peter J. Liu


Ron J. Weiss

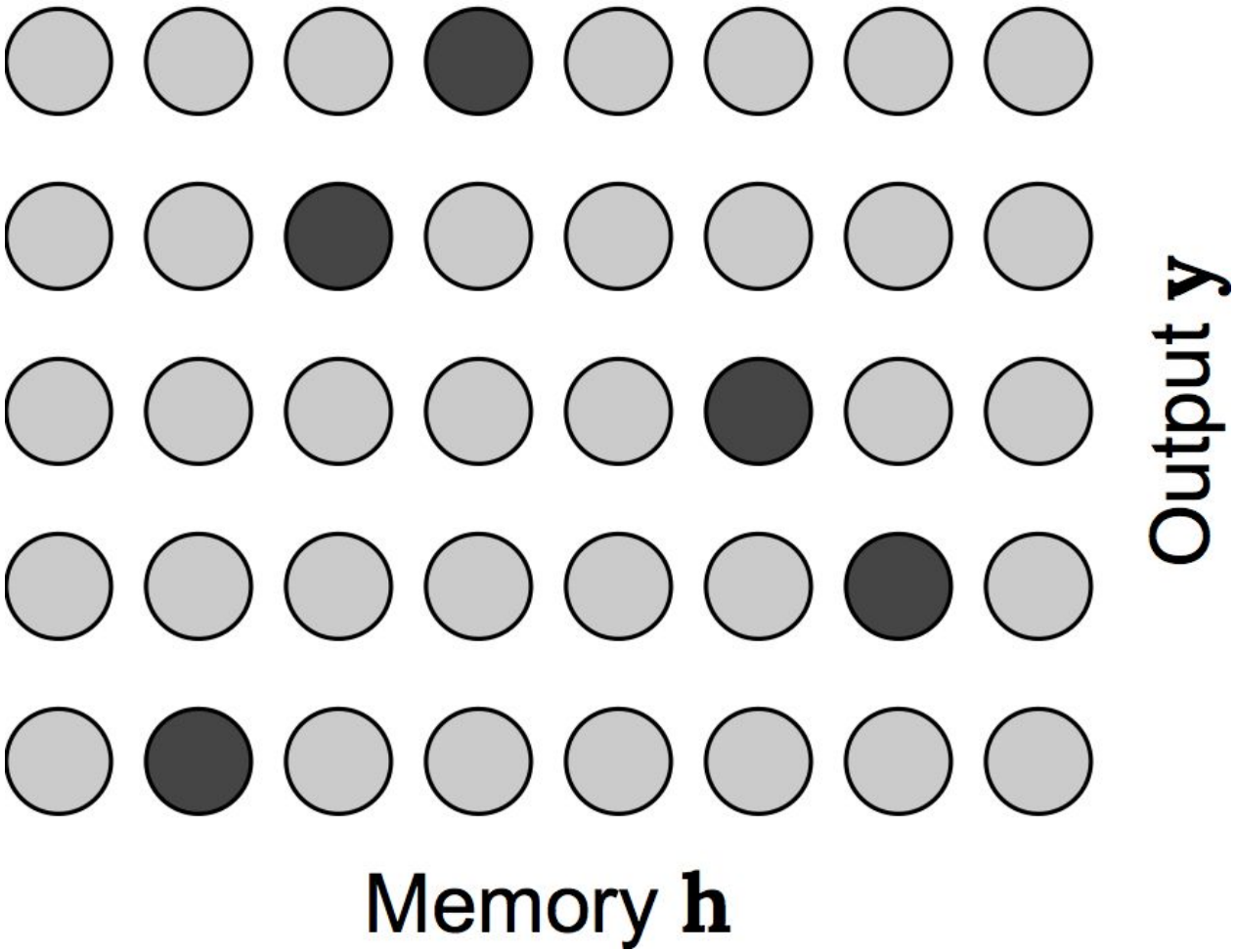

Douglas Eck

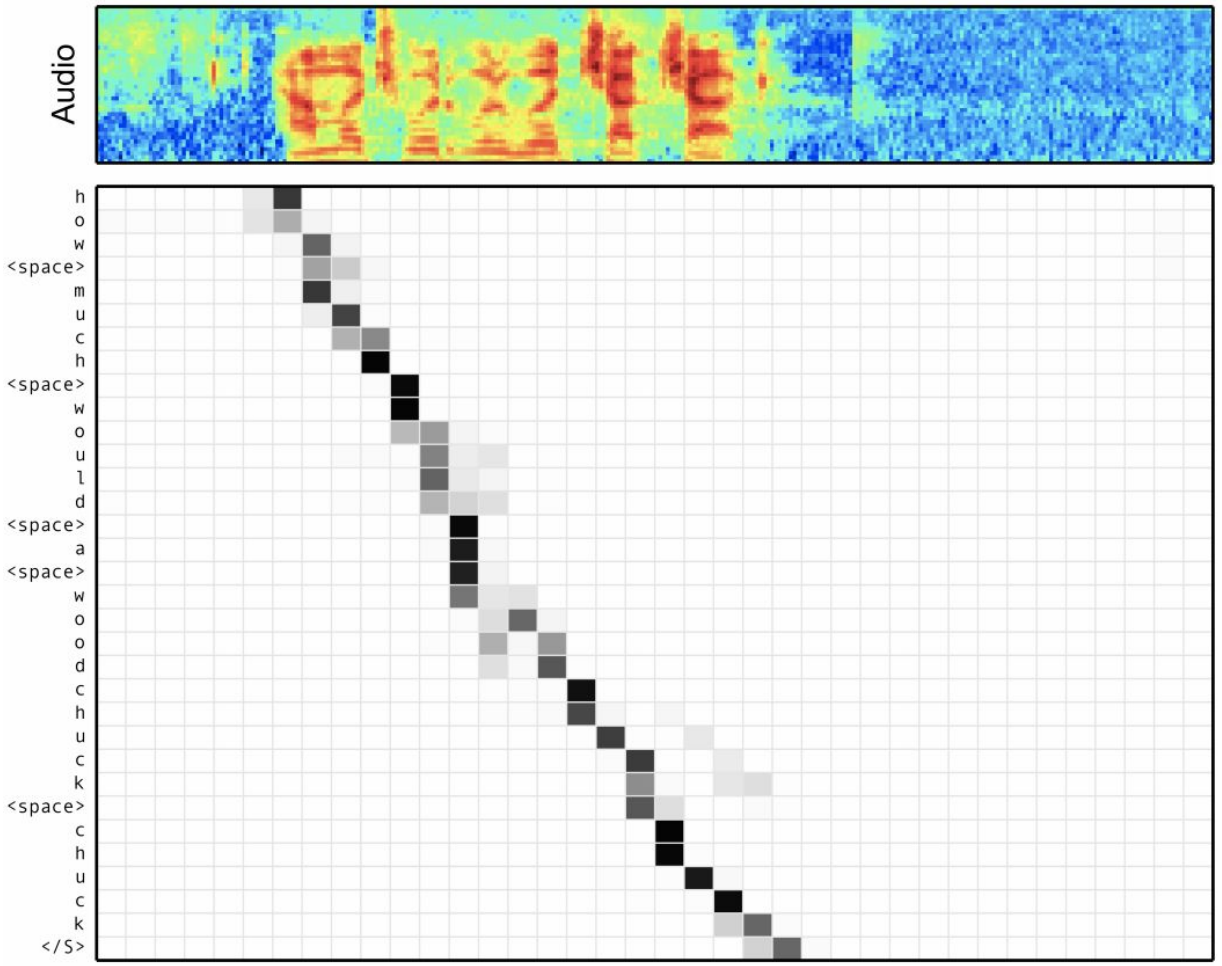


Figures from Olah & Carter, *"Attention and Augmented Recurrent Neural Networks"*

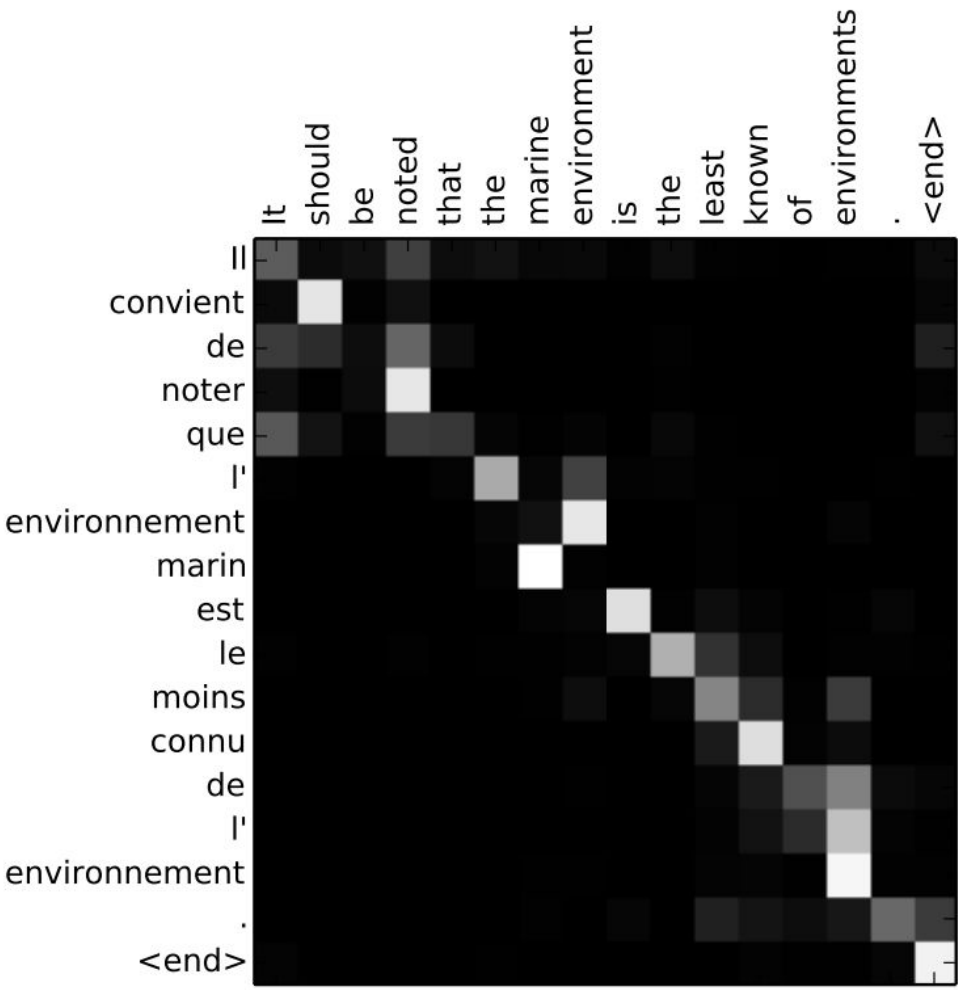


Bahdanau, Cho & Bengio, *"Neural Machine Translation by Jointly Learning to Align and Translate"*


Output y
Memory h

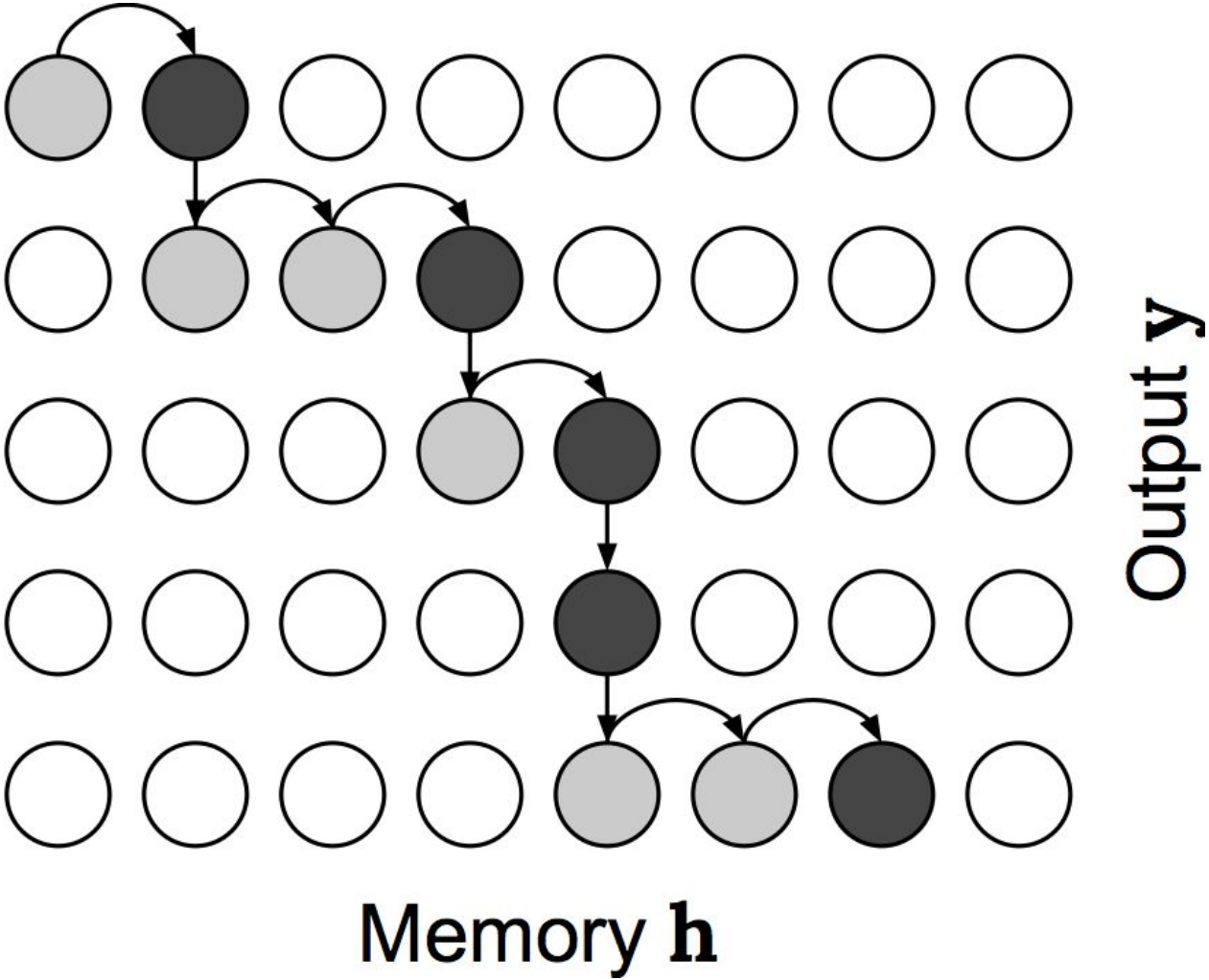


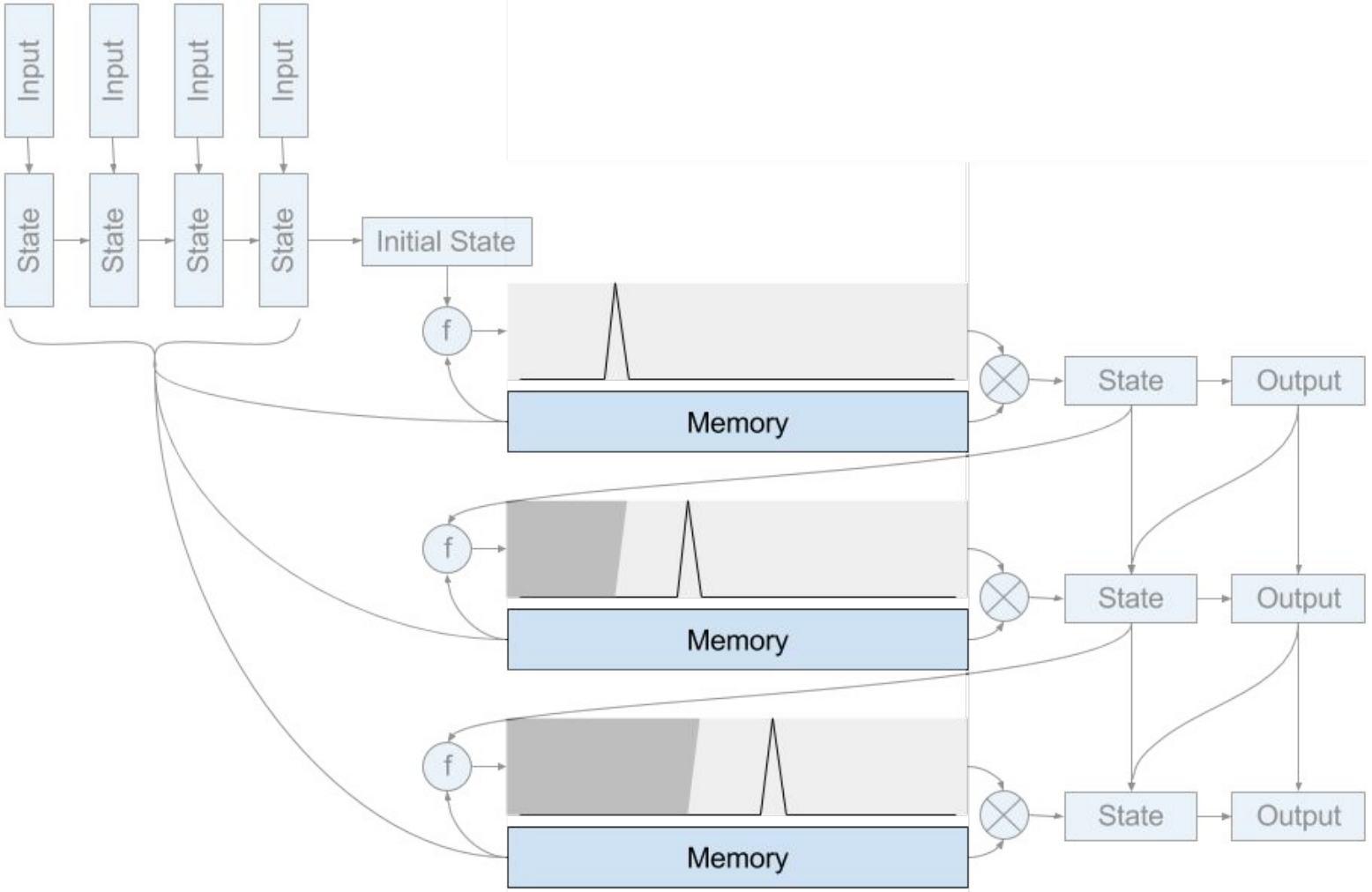
Chan, Jaitly, Le & Vinyals, *"Listen, Attend and Spell"*

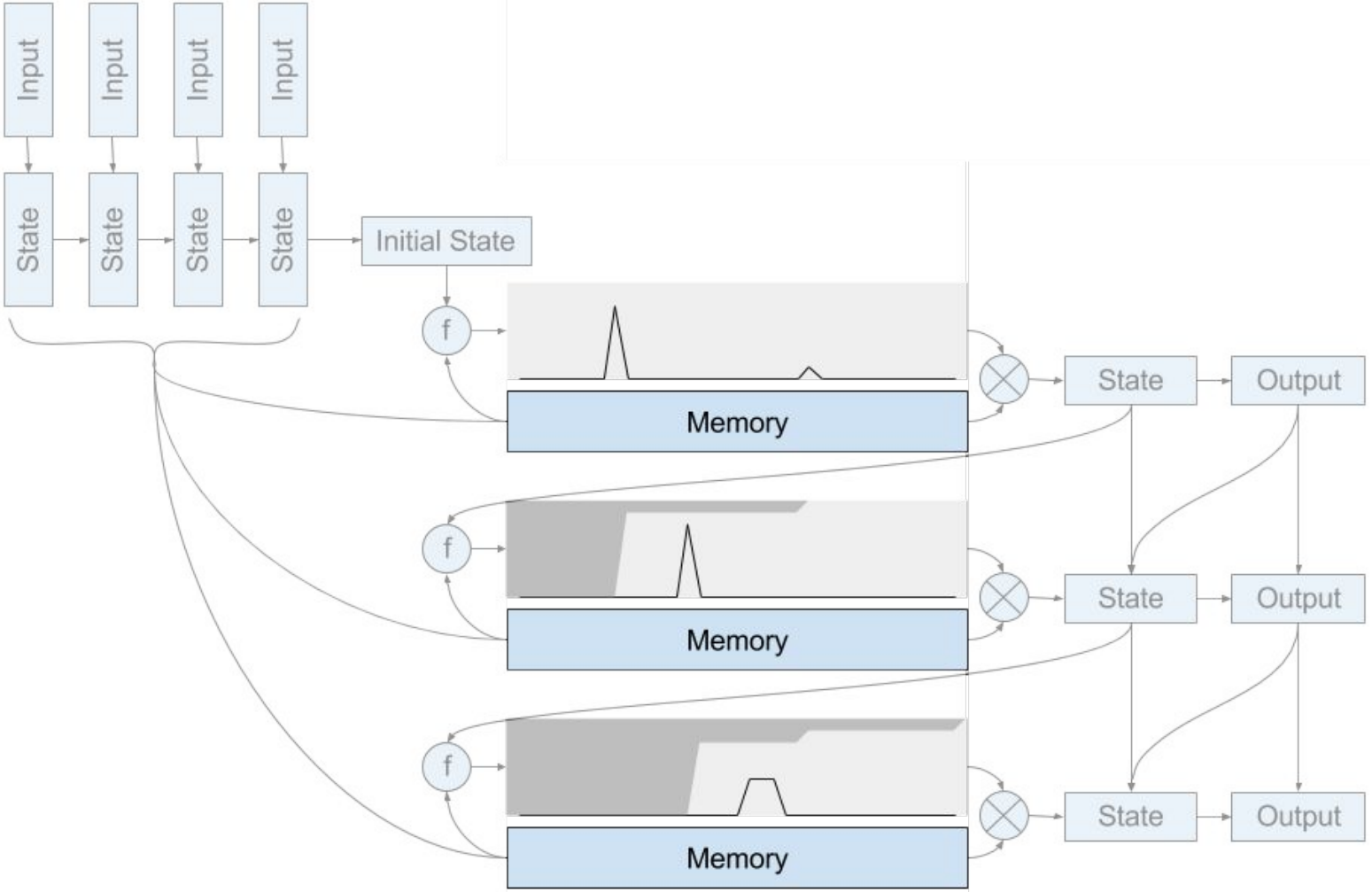


Bahdanau, Cho & Bengio, *"Neural Machine Translation by Jointly Learning to Align and Translate"*

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


Output $\mathbf{y}$
Memory $\mathbf{h}$


Input
Input
Input
Input
State
State
State
State
Initial State
f
Memory
State
Output
f
Memory
State
Output
f
Memory
State
Output


Input
Input
Input
Input
State
State
State
State
Initial State
f
f
f
Memory
Memory
Memory
State
Output
State
Output
State
Output

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

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

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

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

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

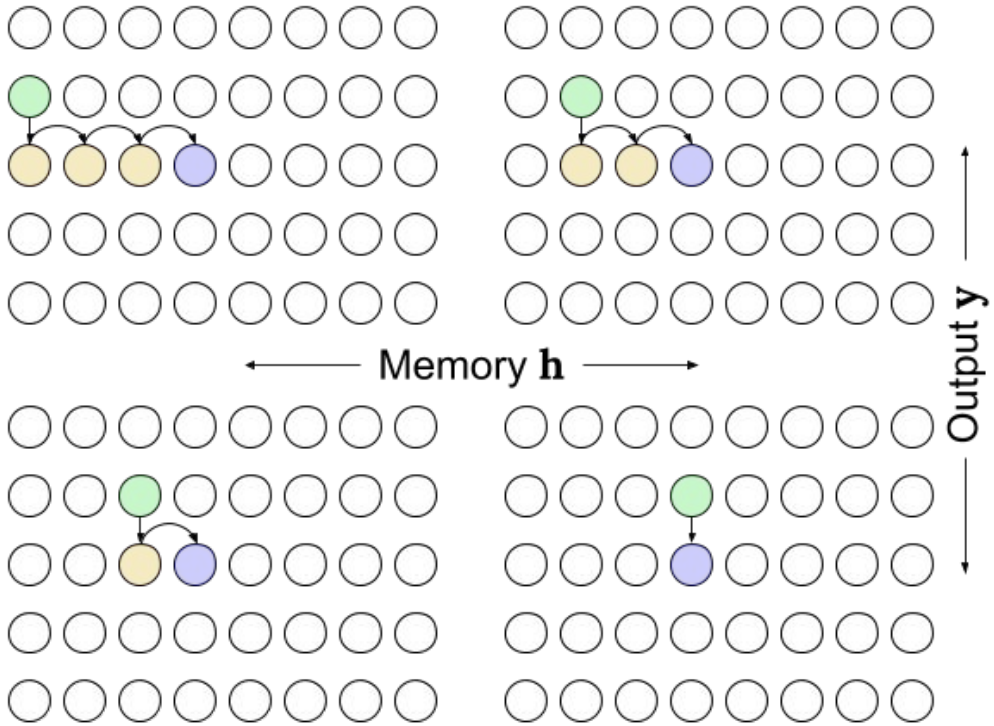
$$\text{energy}_{i,j} = \text{EnergyFunction}(\text{state}_{i-1}, \text{memory}_j)$$

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

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

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

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



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

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

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

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

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

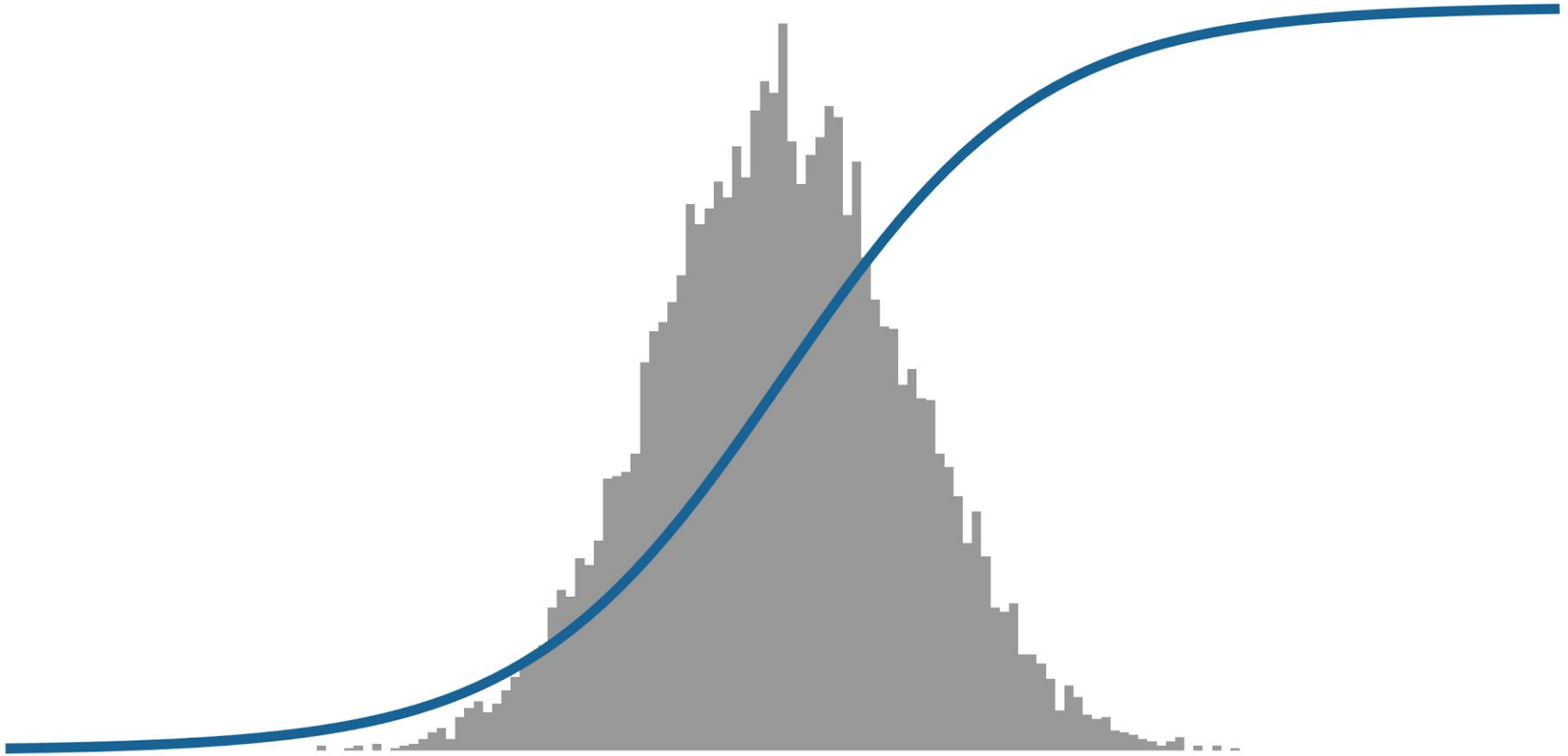
$$\text{attention}_{i,j} = \text{select}_{i,j} \left((1 - \text{select}_{i,j-1}) \frac{\text{attention}_{i,j-1}}{\text{select}_{i,j-1}} + \text{attention}_{i-1,j} \right)$$

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

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

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

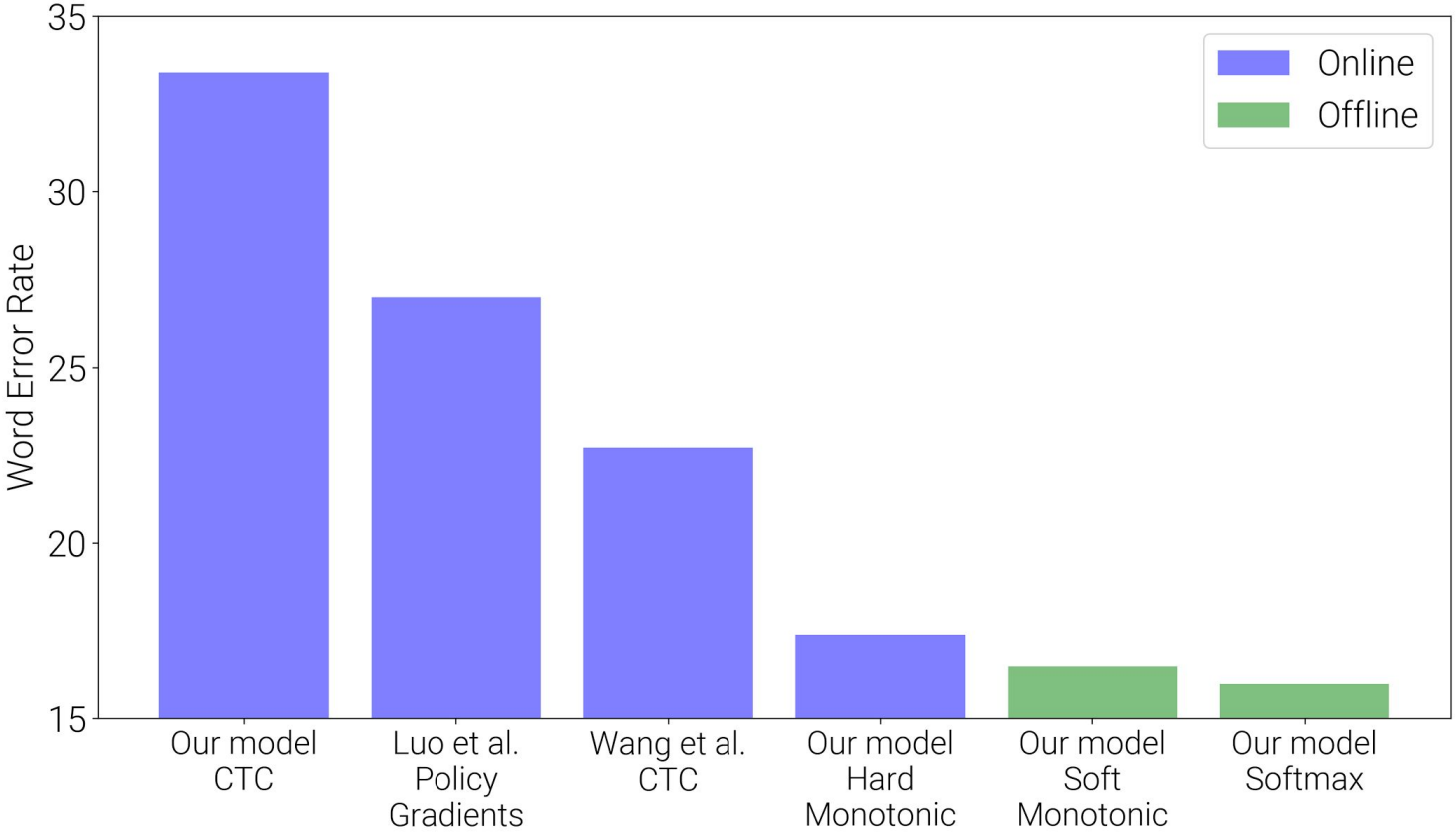
$$\text{attention}_i = \text{select}_i \; \texttt{cumprod}(1 - \text{select}_i) \; \texttt{cumsum} \left(\frac{\text{attention}_{i-1}}{\texttt{cumprod}(1 - \text{select}_i)} \right)$$

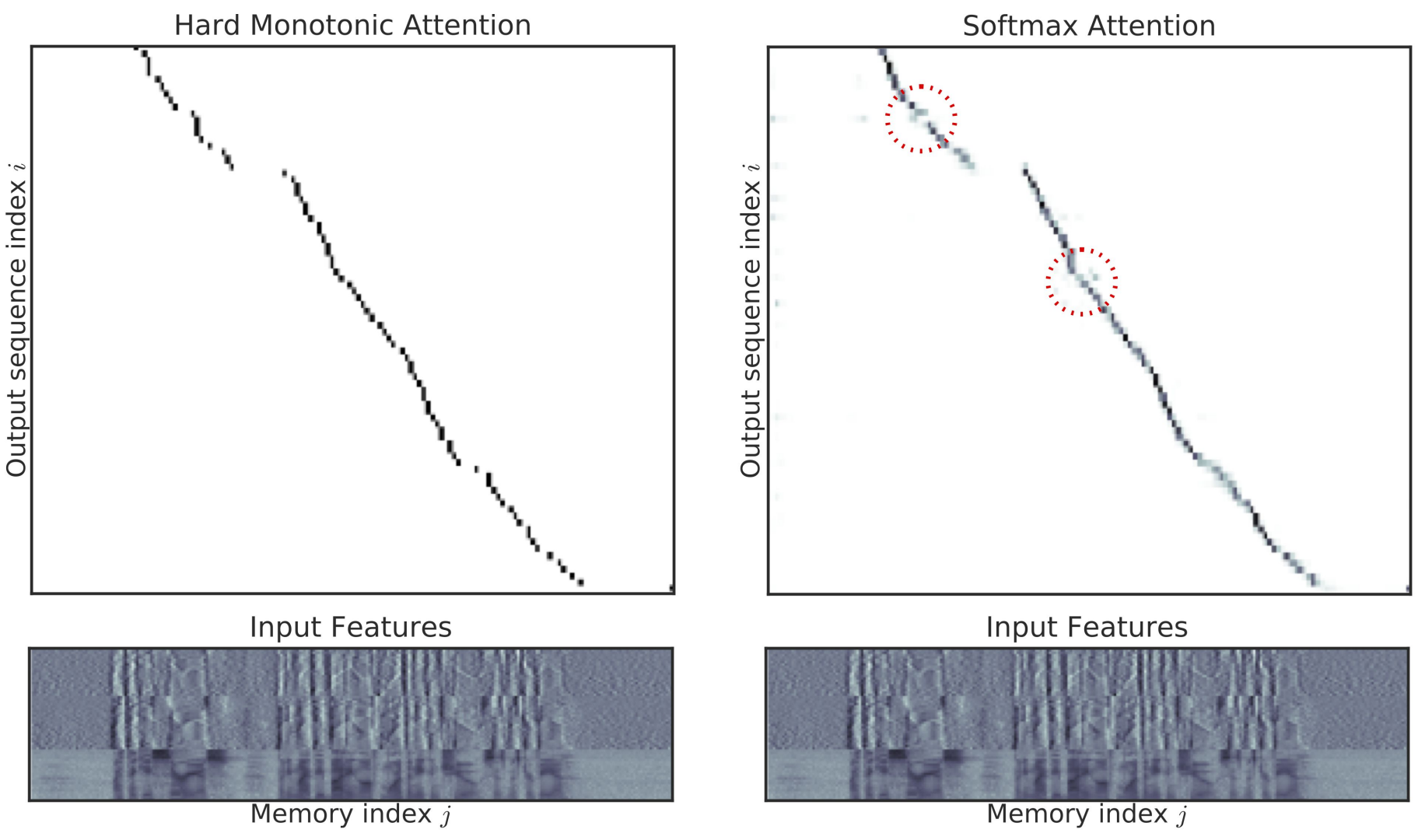


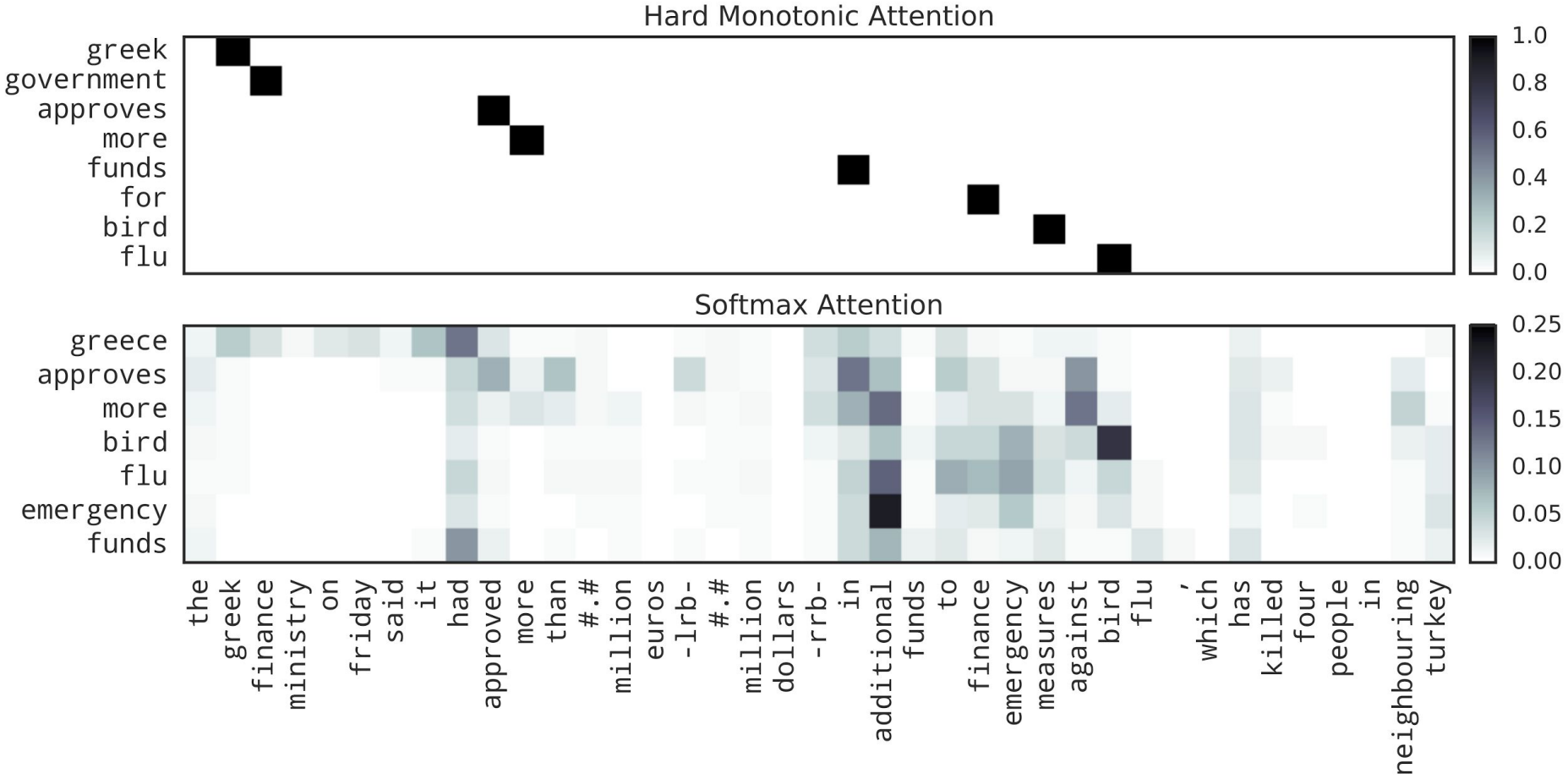
Frey, “*Continuous Sigmoidal Belief Networks Trained Using Slice Sampling*”

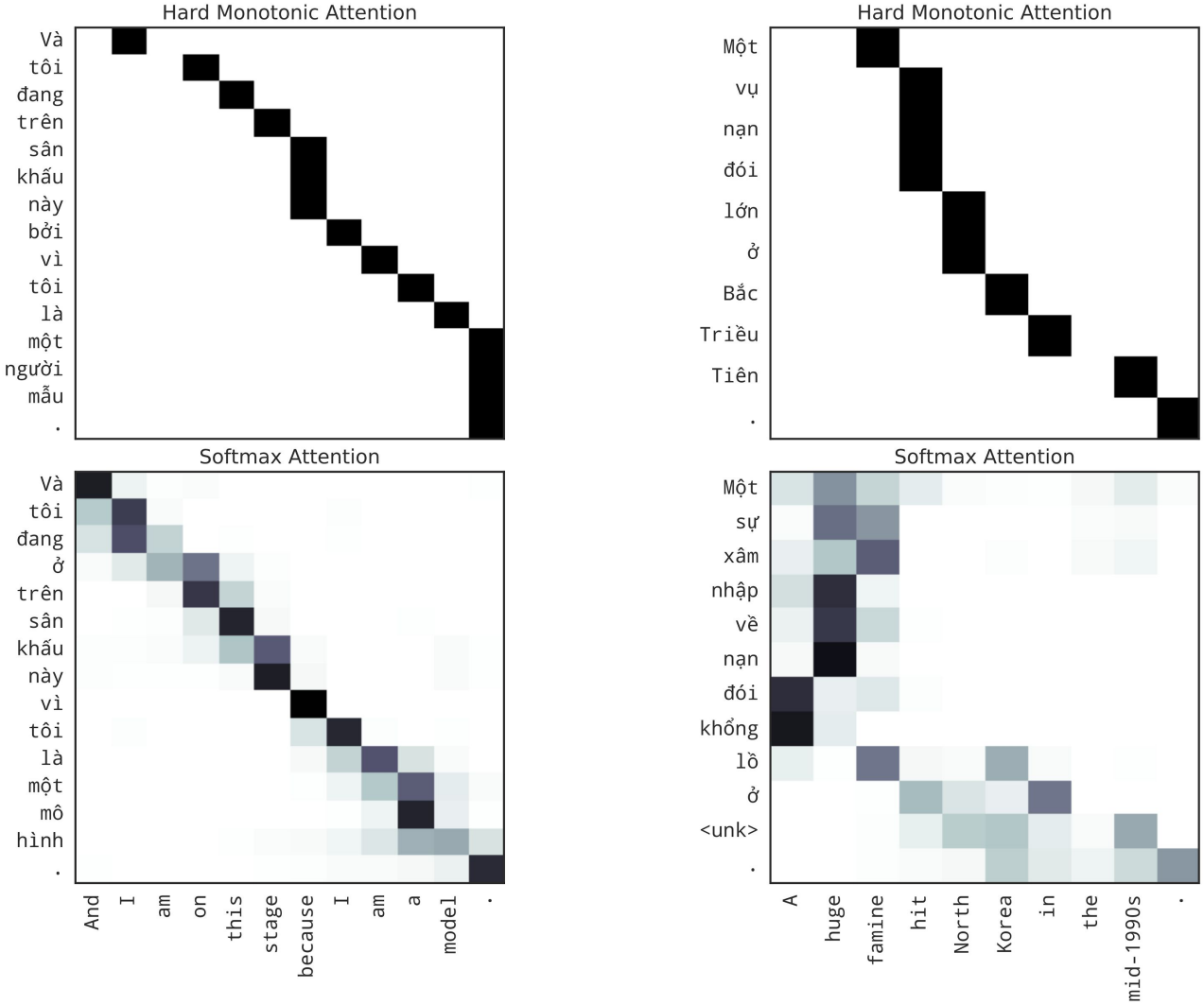
Salakhutdinov & Hinton, *“Semantic Hashing”*

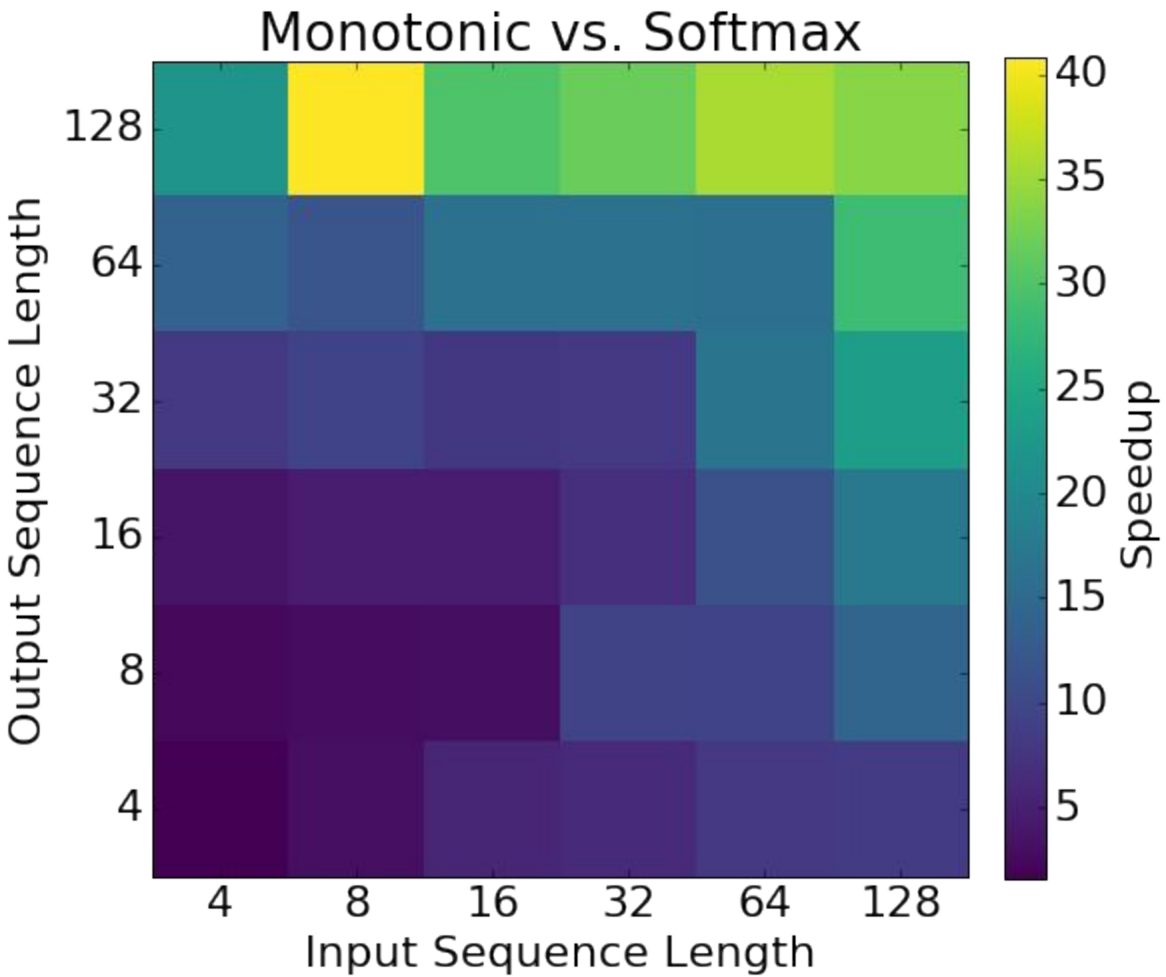
Foerster, Assael, de Freitas & Whiteson, *“Learning to Communicate with Deep Multi-Agent Reinforcement Learning”*


35
30
25
20
15
Word Error Rate
Online
Offline
Our model CTC
Luo et al. Policy Gradients
Wang et al. CTC
Our model Hard Monotonic
Our model Soft Monotonic
Our model Softmax


Hard Monotonic Attention
Output sequence index i
Input Features
Memory index j
Softmax Attention
Output sequence index i
Input Features
Memory index j


Hard Monotonic Attention
greek
government
approves
more
funds
for
bird
flu
1.0
0.8
0.6
0.4
0.2
0.0
Softmax Attention
greece
approves
more
bird
flu
emergency
funds
0.25
0.20
0.15
0.10
0.05
0.00
the
greek
finance
ministry
on
friday
said
it
had
approved
more
than
#.#
million
euros
-lrb-
#.#
million
dollars
-rrb-
in
additional
funds
to
finance
emergency
measures
against
bird
flu
,
which
has
killed
four
people
in
neighbouring
turkey


Hard Monotonic Attention
Và
tôi
đang
trên
sân
khấu
này
bởi
vì
tôi
là
một
người
mẫu
.
Hard Monotonic Attention
Một
vụ
nạn
đói
lớn
ở
Bắc
Triều
Tiên
.
Softmax Attention
Và
tôi
đang
ở
trên
sân
khấu
này
vì
tôi
là
một
mô
hình
.
And
I
am
on
this
stage
because
I
am
a
model
.
Softmax Attention
Một
sự
xâm
nhập
về
nạn
đói
khổng
lồ
ở
<unk>
.
A
huge
famine
hit
North
Korea
in
the
mid-1990s
.


Monotonic vs. Softmax
Output Sequence Length
128
64
32
16
8
4
Input Sequence Length
4
8
16
32
64
128
Speedup
40
35
30
25
20
15
10
5

Pointers

Implemented in **`tf.contrib.seq2seq`**

Additional code at `http://github.com/craffel/mad`

“Practitioner’s Guide” in Appendix G

Blog post: `http://colinraffel.com/blog`

These slides: `http://colinraffel.com/talks/icml2017online.pdf`

Poster #70 tonight

My email: `craffel@gmail.com`

Thanks!