Attention!

...and it's use in Deep Learning

Anders Huss

Stockholm AI - Reading Group
2017-08-23

anders@watty.io

What about it?

Self driving cars

anders@watty.io

What about it?

Machine Translation

“narrow down the focus to the most relevant parts”

“recentrez l'attention sur les éléments les plus pertinents”

anders@watty.io

What about it?

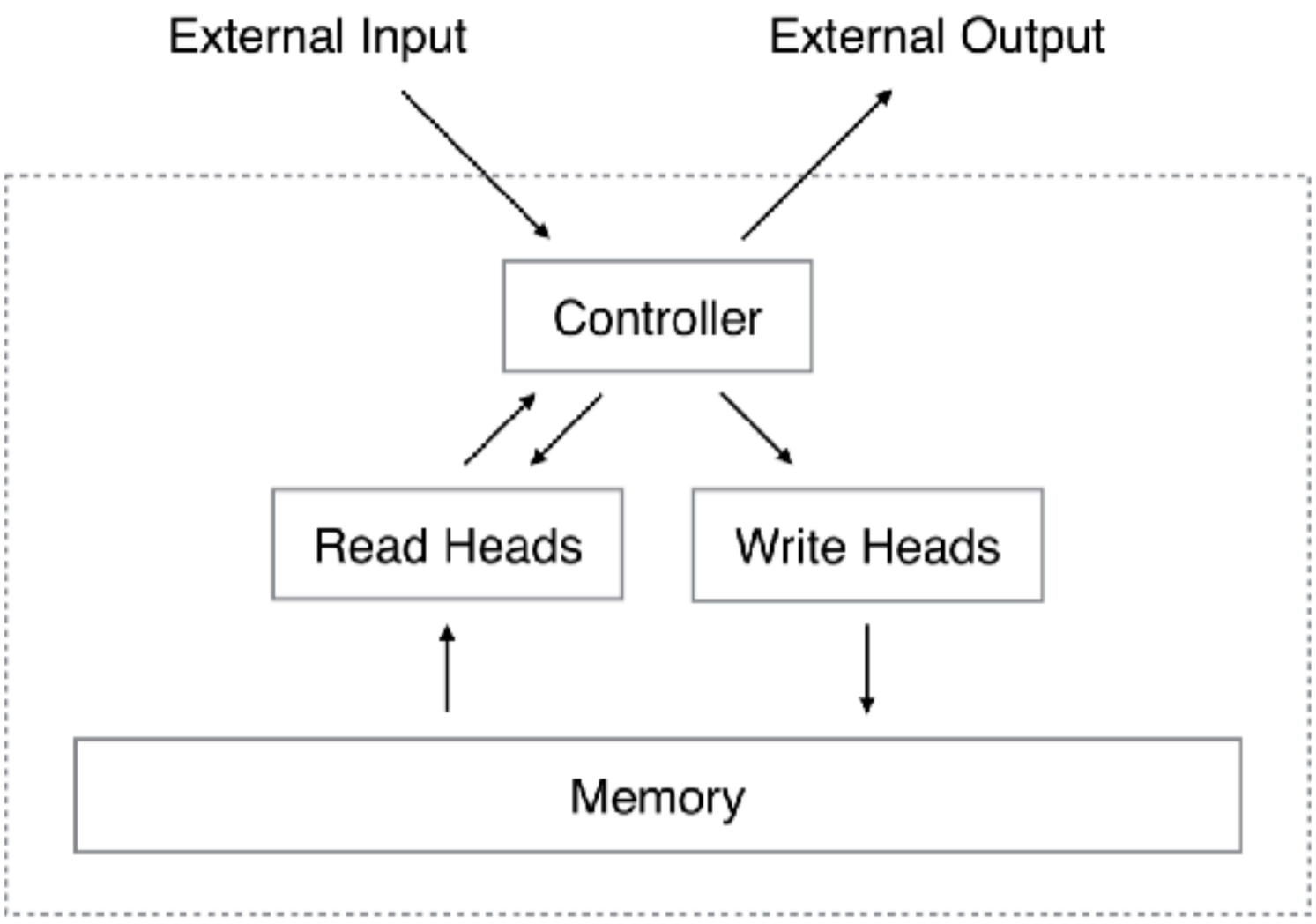
Solving Problems (Logical Reasoning)

“many times complexity
is a collection of
connected simplicities”

anders@watty.io

What about it?

Solving Problems (Logical Reasoning)



anders@watty.io

Agenda

- Problem Formulation
- Spatial Attention
  - Example/Benchmark
  - Applications
- Beyond Spatial Attention
  - Differentiable Neural Computers
- Discussion and Questions

anders@watty.io

Iterative Application of Attention (RNN)

Regular RNN/LSTM/GRU…

$$y^t, s^t = f(x^t, s^{t-1}) \qquad y^{1:T}, s^{1:T} = RNN_f(x^{1:T}, s^0)$$

With Attention ($\mathbf{Z}$)

$$y^t, s^t = g(x^t, s^{t-1}, \mathbf{Z}) \qquad y^{1:T}, s^{1:T} = RNN_g(x^{1:T}, s^0, \mathbf{Z})$$

E.g. by *altering input* to the regular RNN/GRU/LSTM…

$$y^t, s^t = f(x'^t, s^{t-1}), \quad x'^t = a(x^t, s^{t-1}, \mathbf{Z})$$

anders@watty.io

Attention Mechanisms

At each time step *t* we predict a **weighting** over input **Z** and map it to an attention encoding z^t

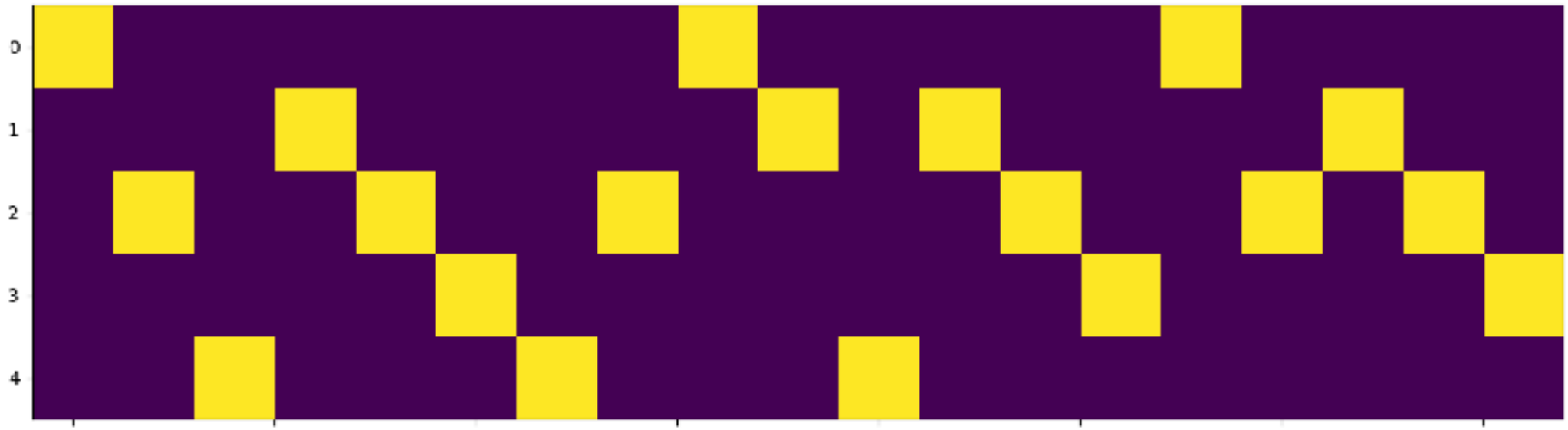
$$y^t, s^t = f(x'^t, s^{t-1})$$

$$x'^t = a(x^t, s^{t-1}, \mathbf{Z}) = [x^t, z^t(x^t, s^{t-1}, \mathbf{Z})]$$

There are many options for how to do this…

anders@watty.io

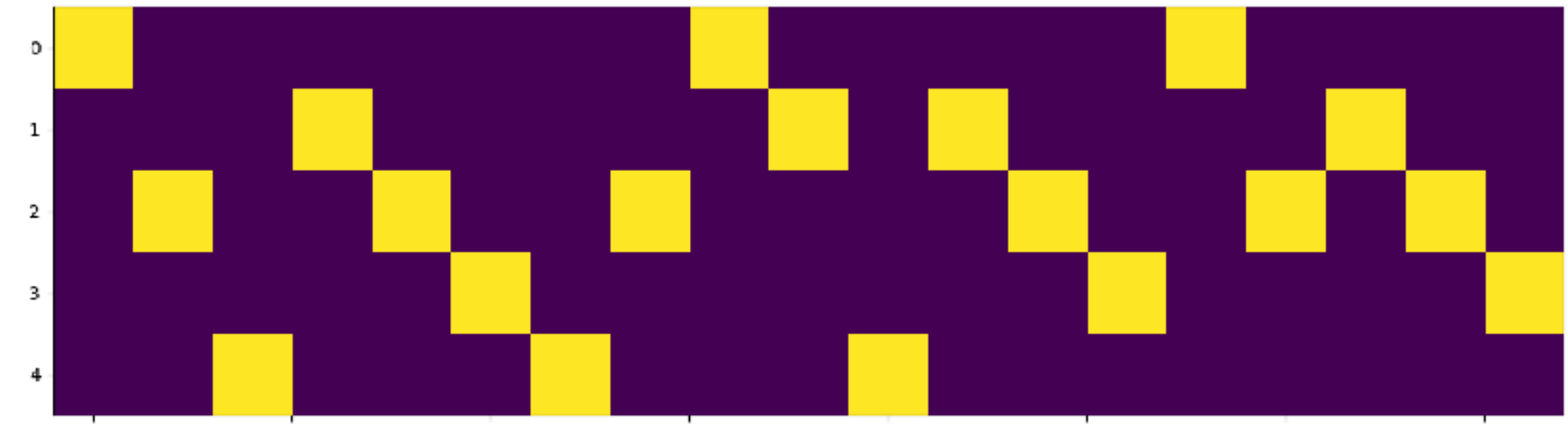
Spatial Attention: 1D

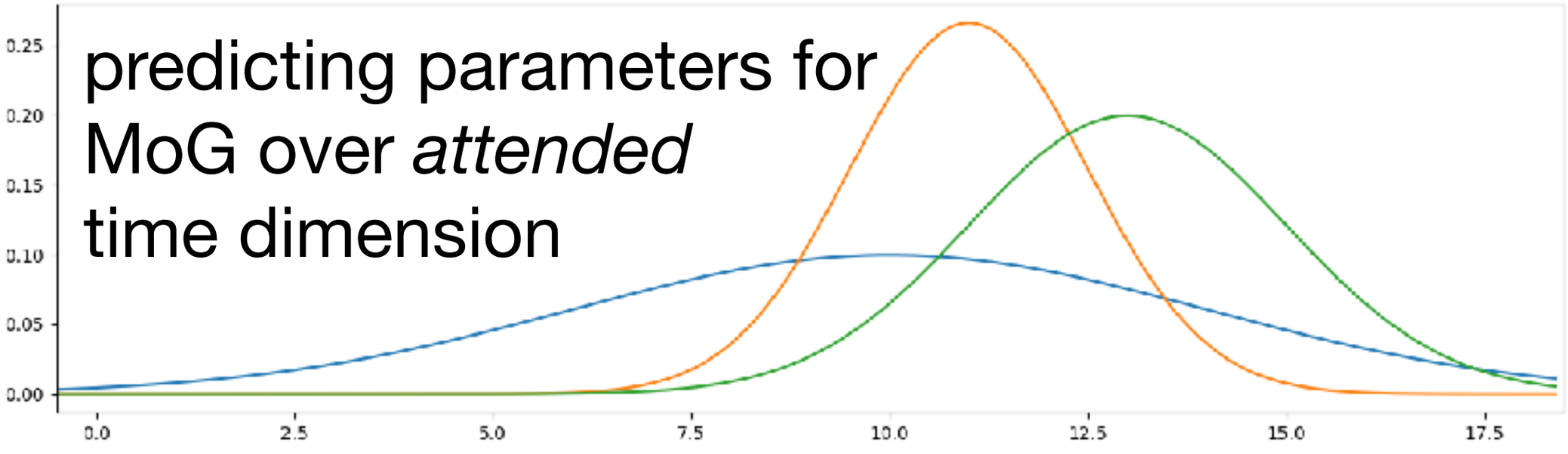


At each time step *t* we predict a **weighting** over input sequence **Z** and map it to an attention encoding z^t

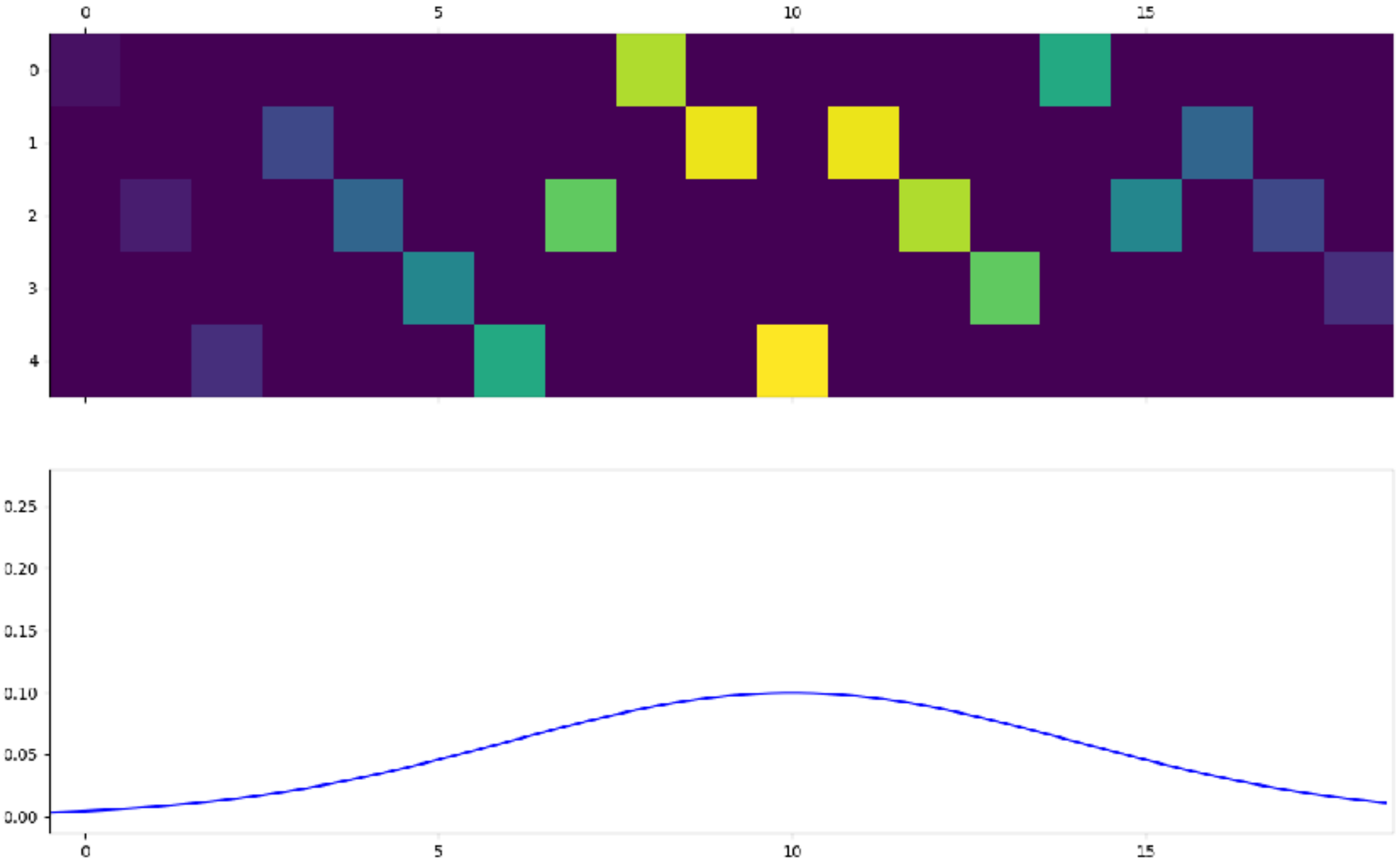
anders@watty.io

Spatial Attention: 1D

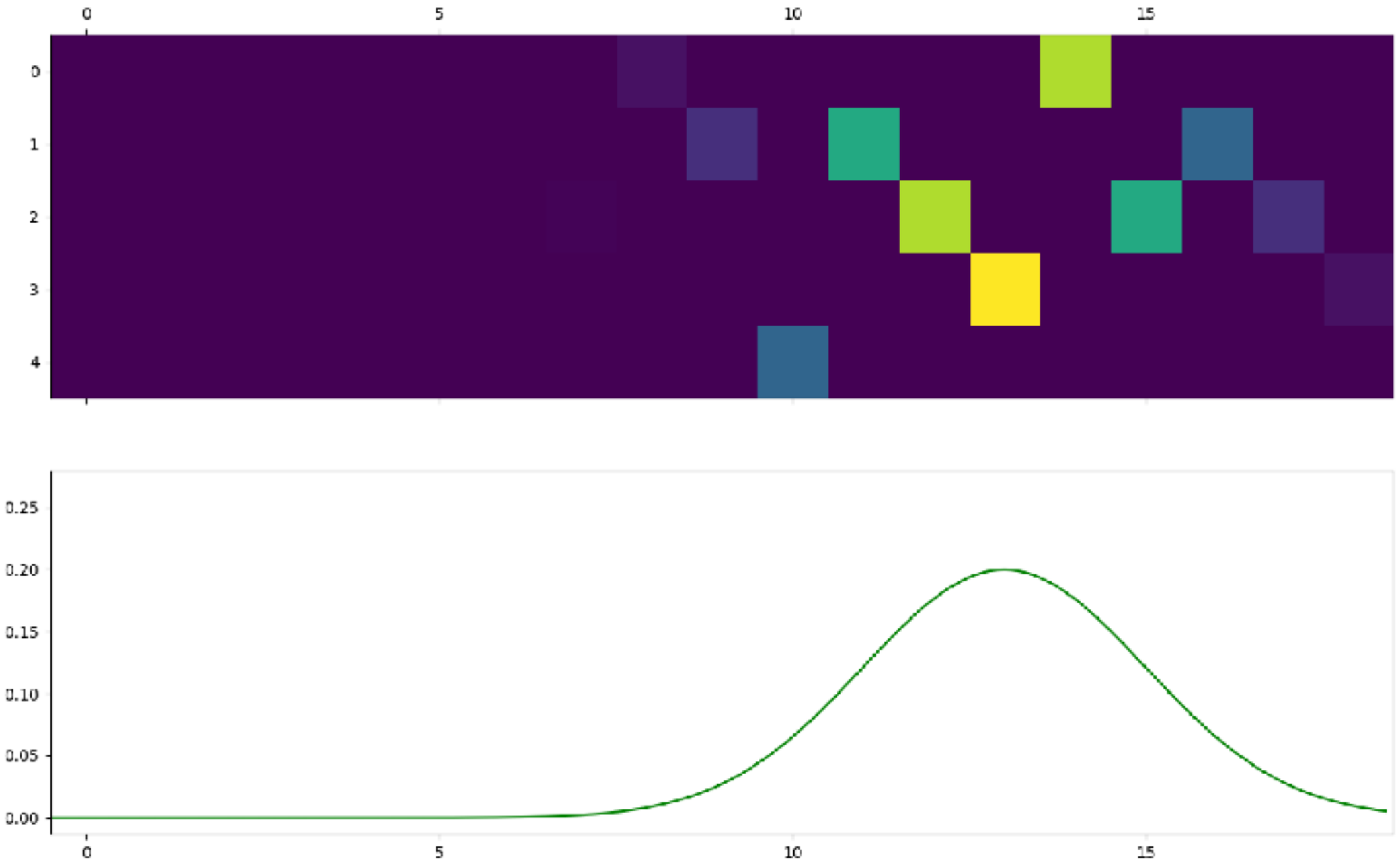




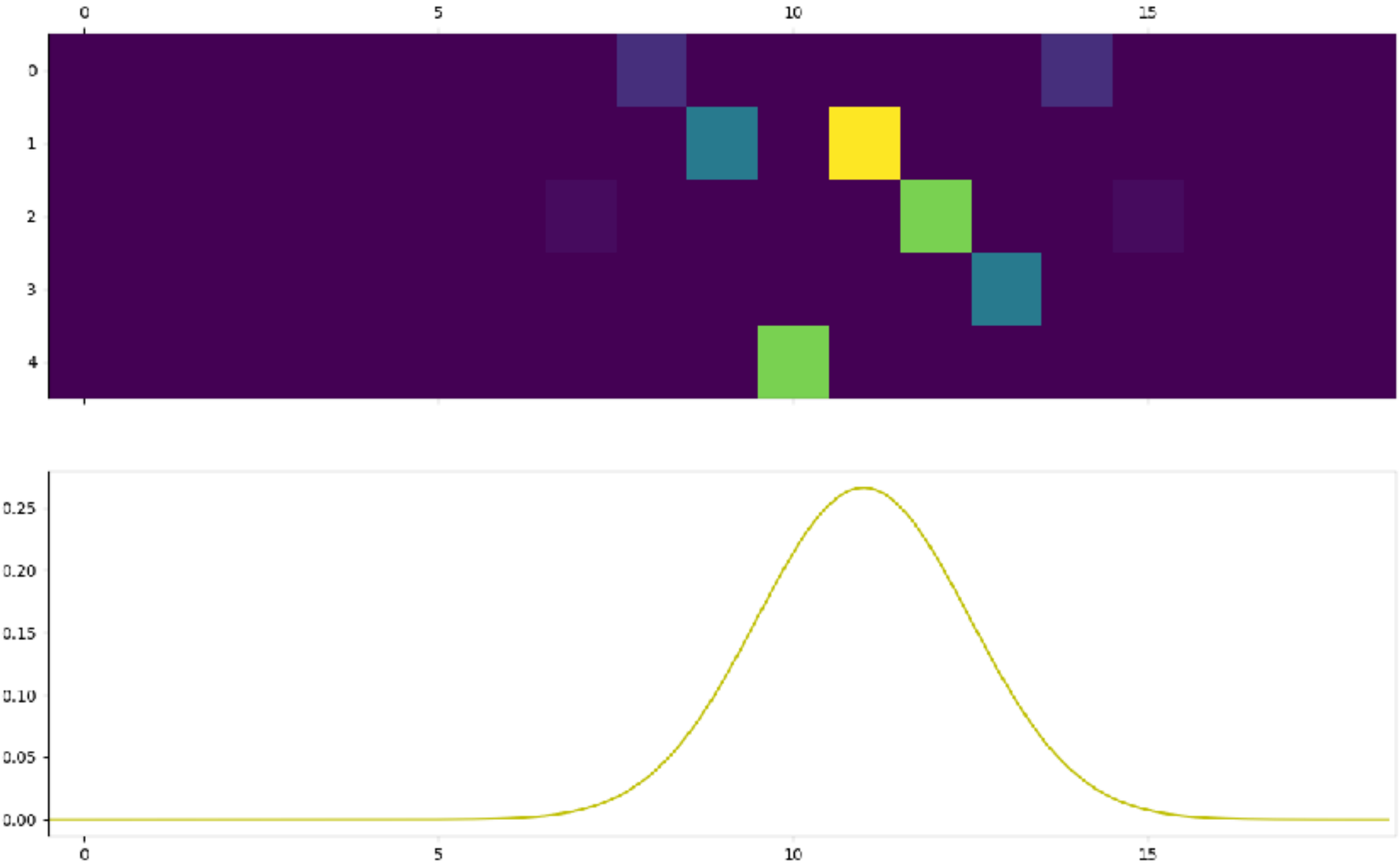
Spatial Attention: 1D



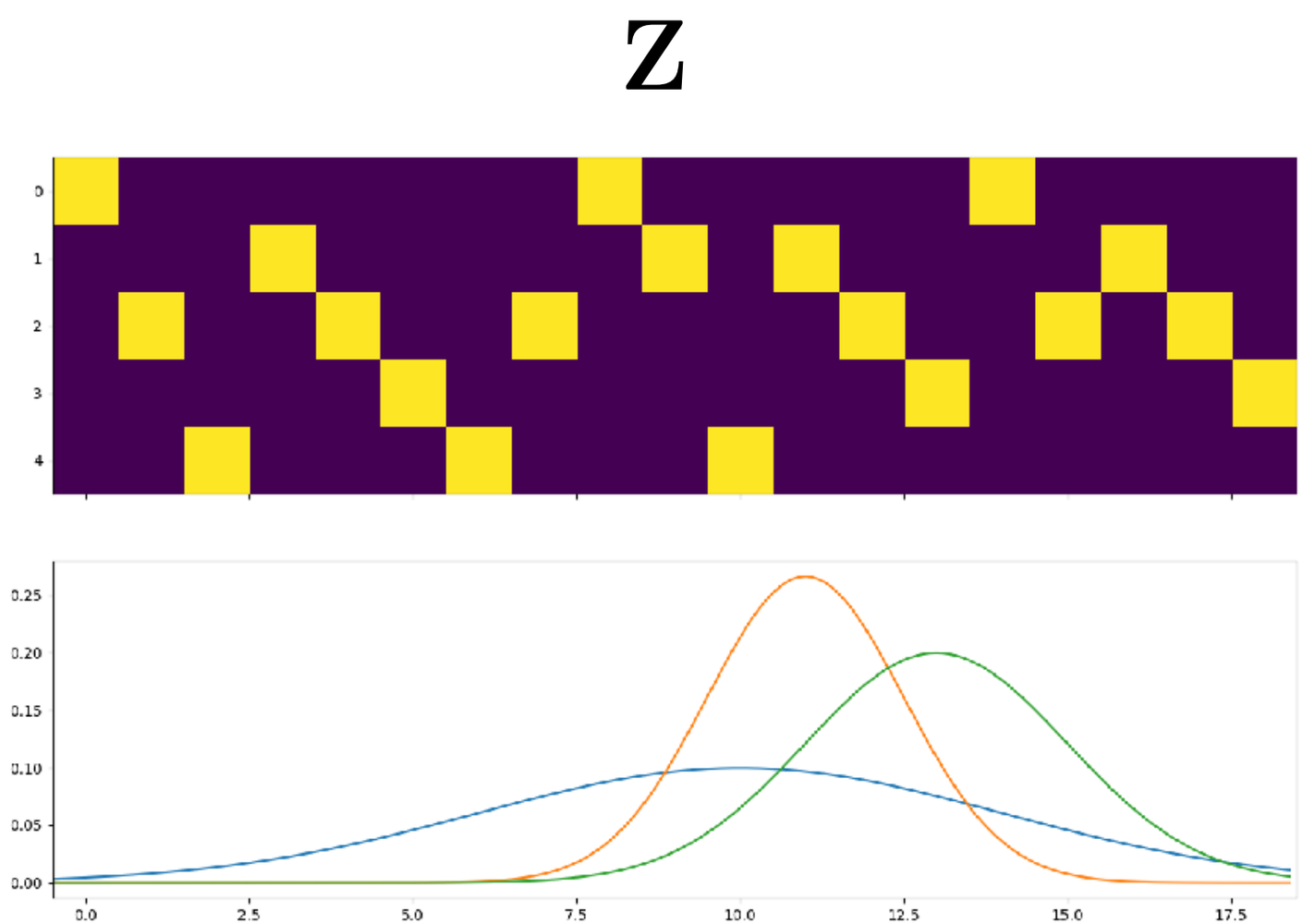
Spatial Attention: 1D



Spatial Attention: 1D

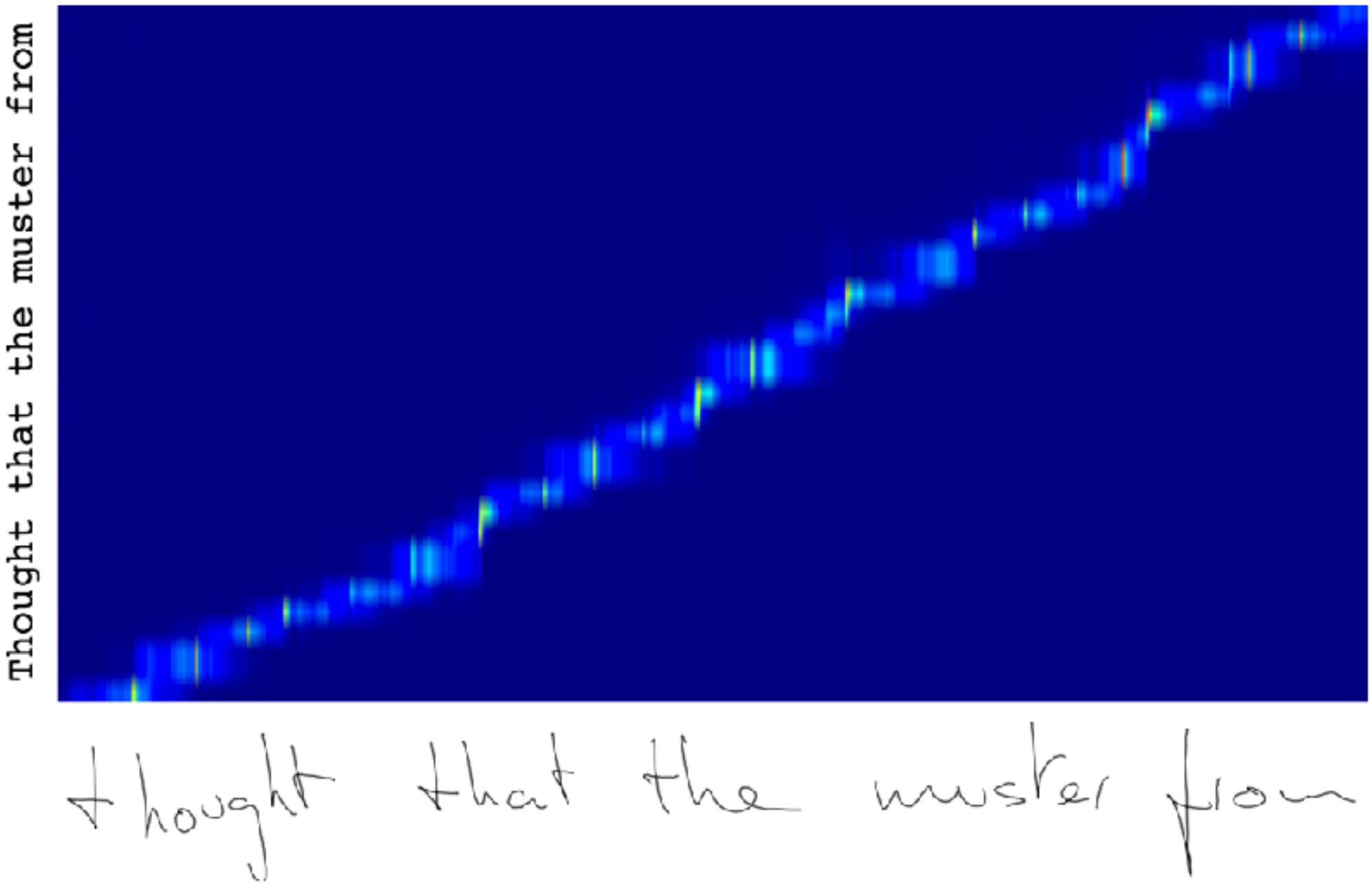


Spatial Attention: 1D



anders@watty.io

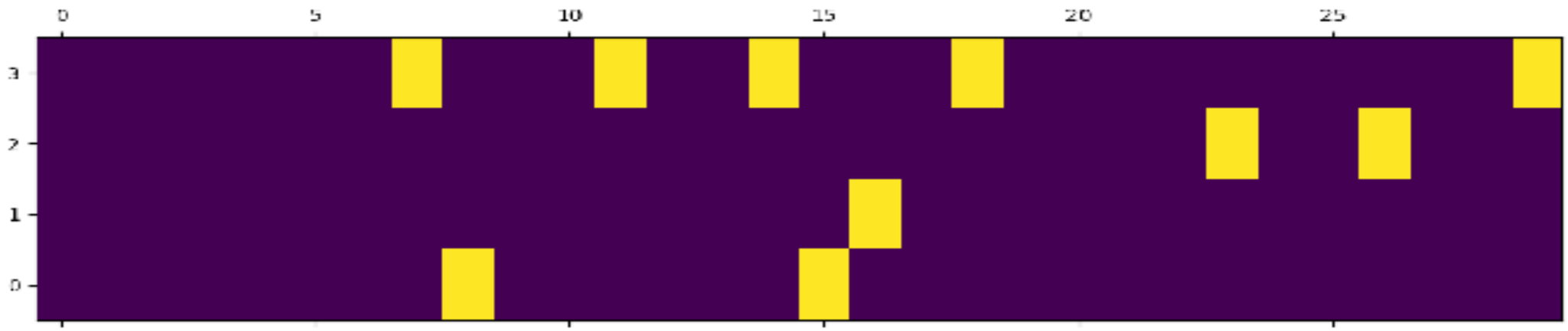
Application: Handwriting synthesis



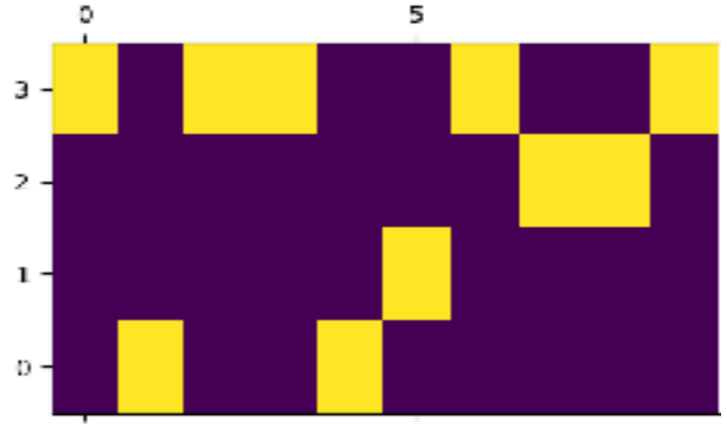
anders@watty.io

Canonical Example of Alignment/Selection Problem

Input/Attended

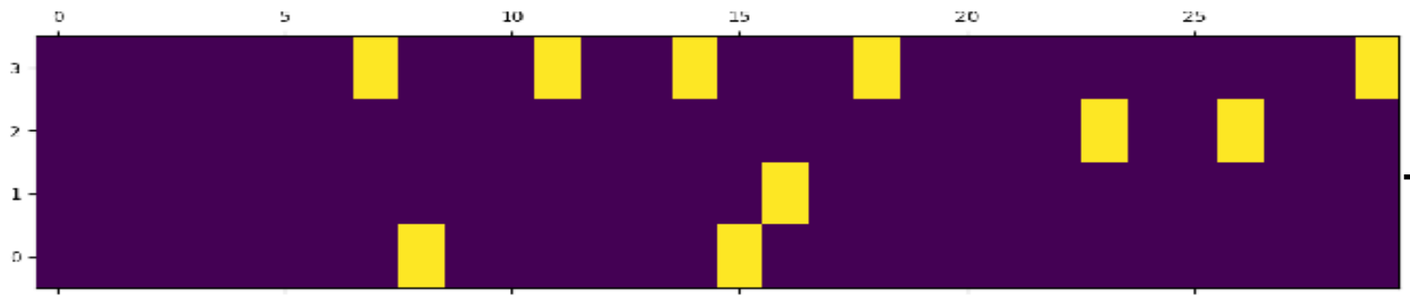


Target



anders@watty.io

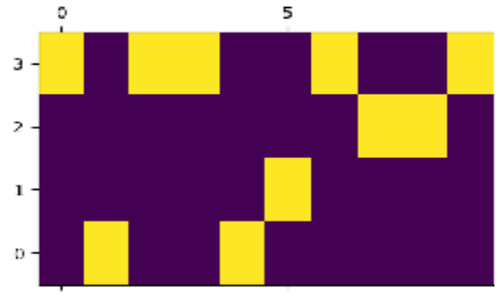
Sequence-to-Sequence



LSTM-Encoder

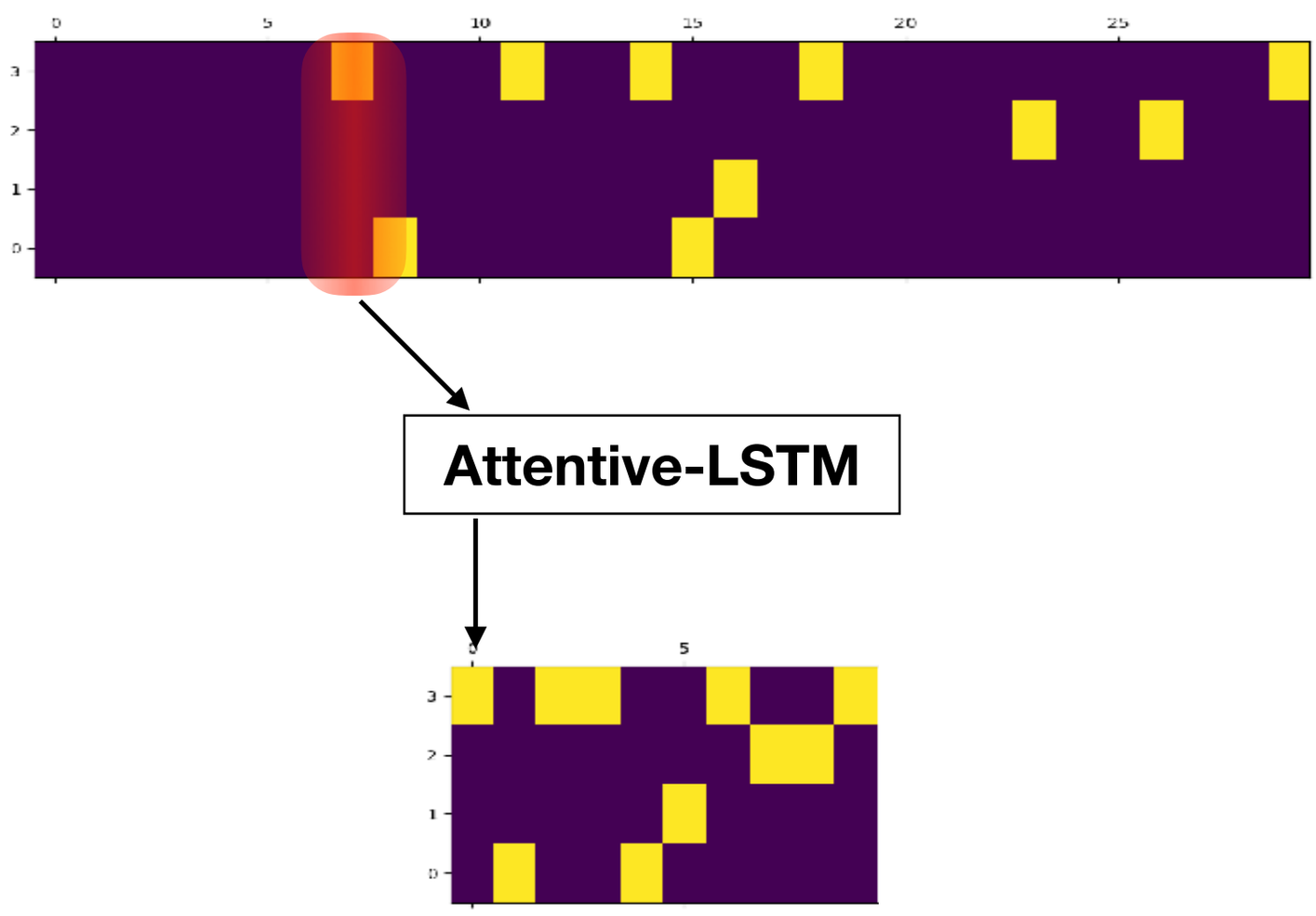
Encoding

LSTM-Decoder



anders@watty.io

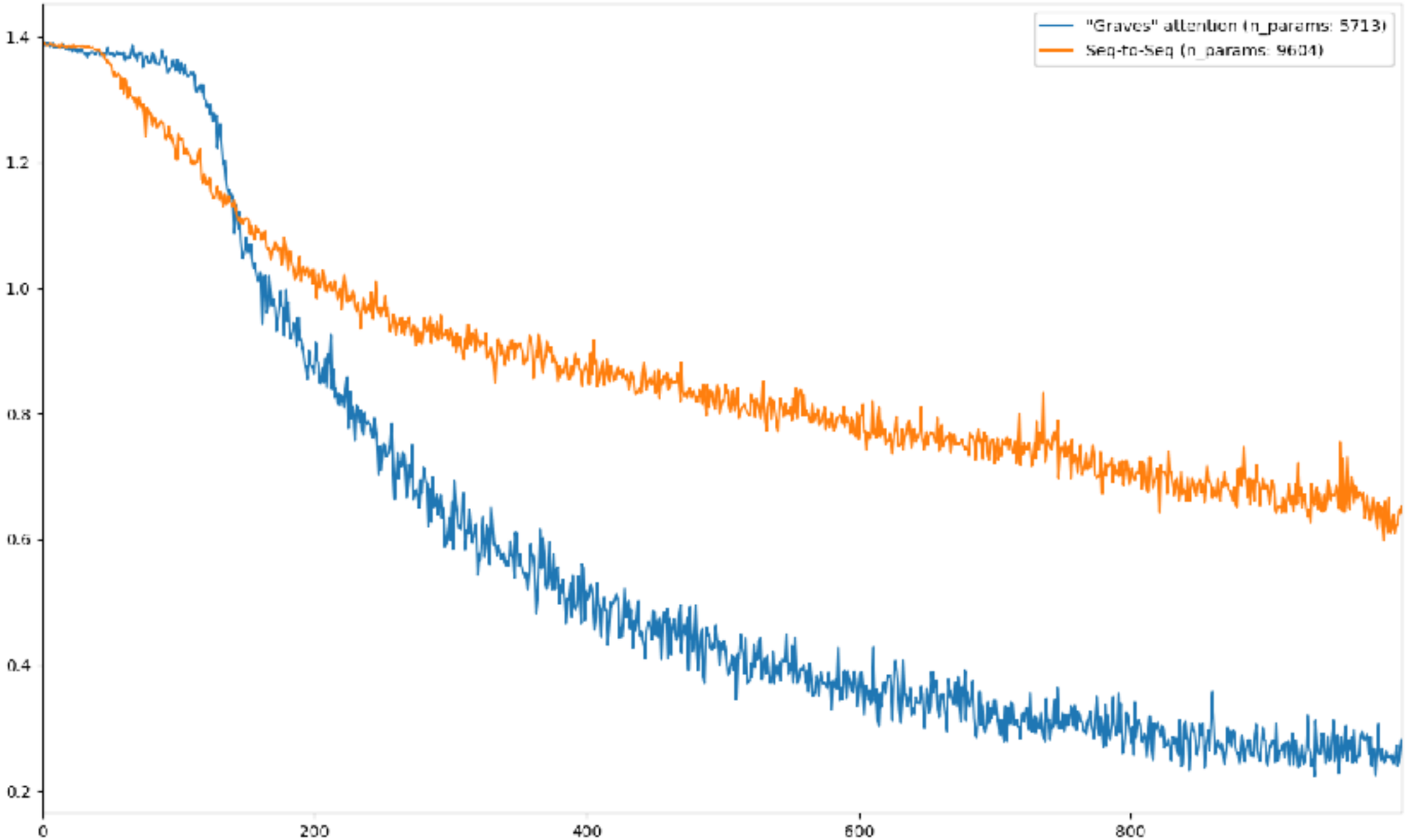
Attention



anders@watty.io

Seq-to-Seq vs Attention

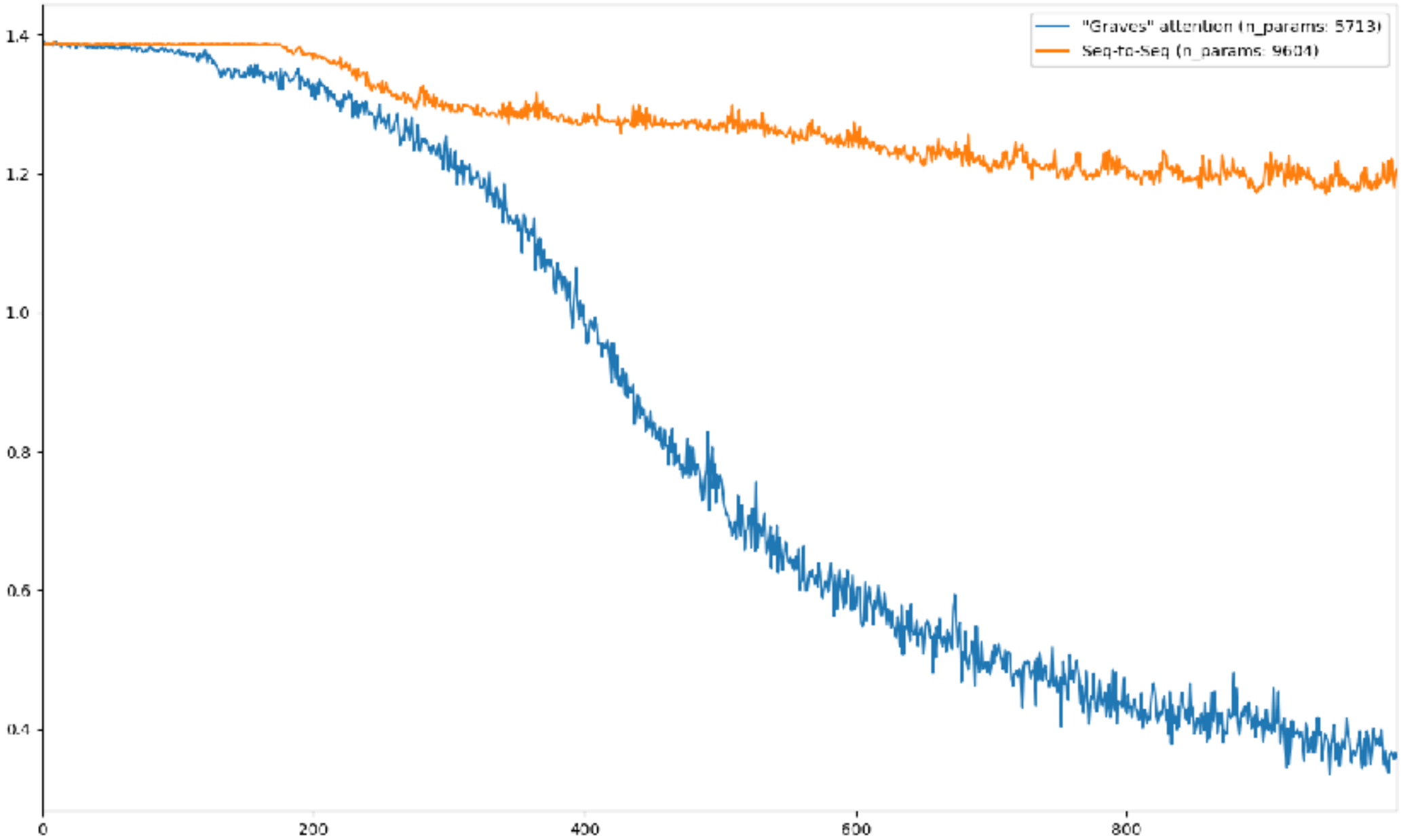
attended sequence length: 30, target sequence length: 10



anders@watty.io

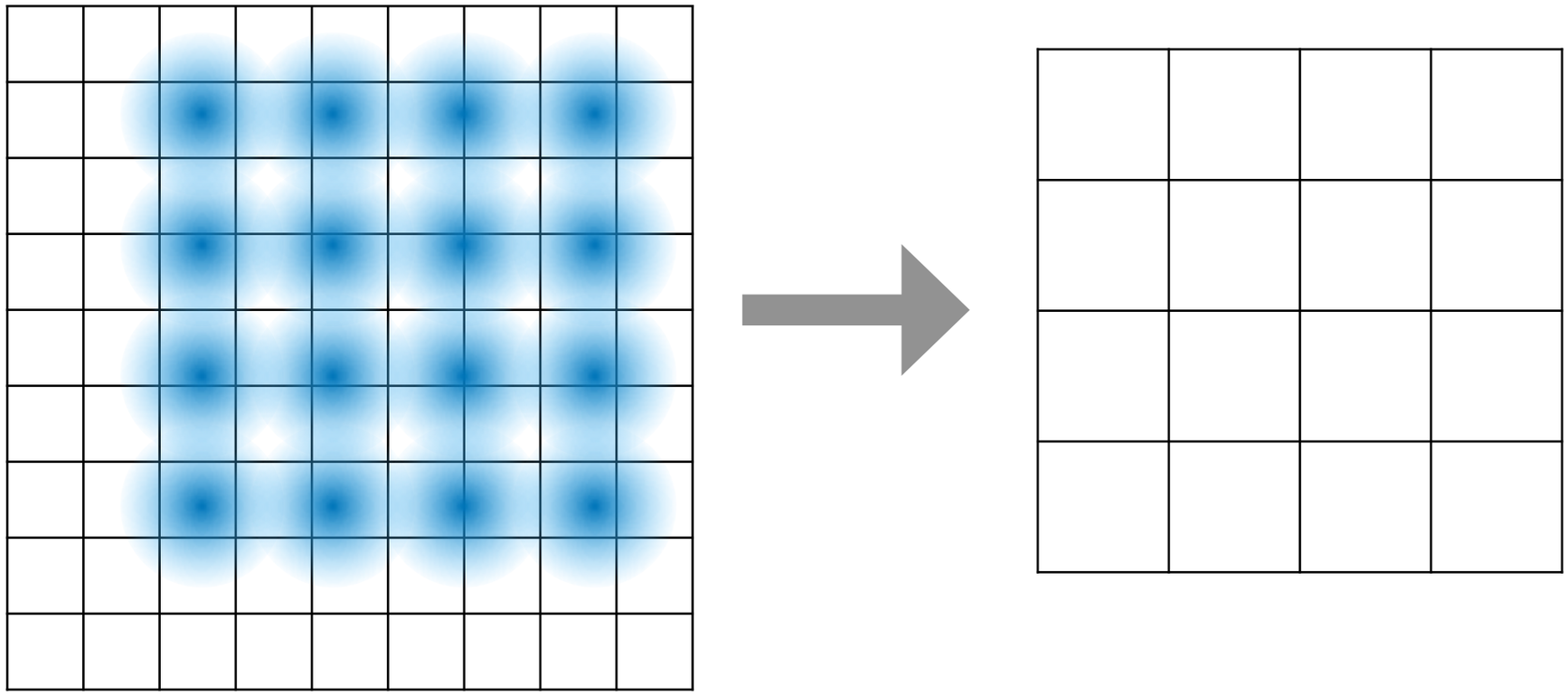
Seq-to-Seq vs Attention

attended sequence length: 60, target sequence length: 20



Spatial Attention: 2D

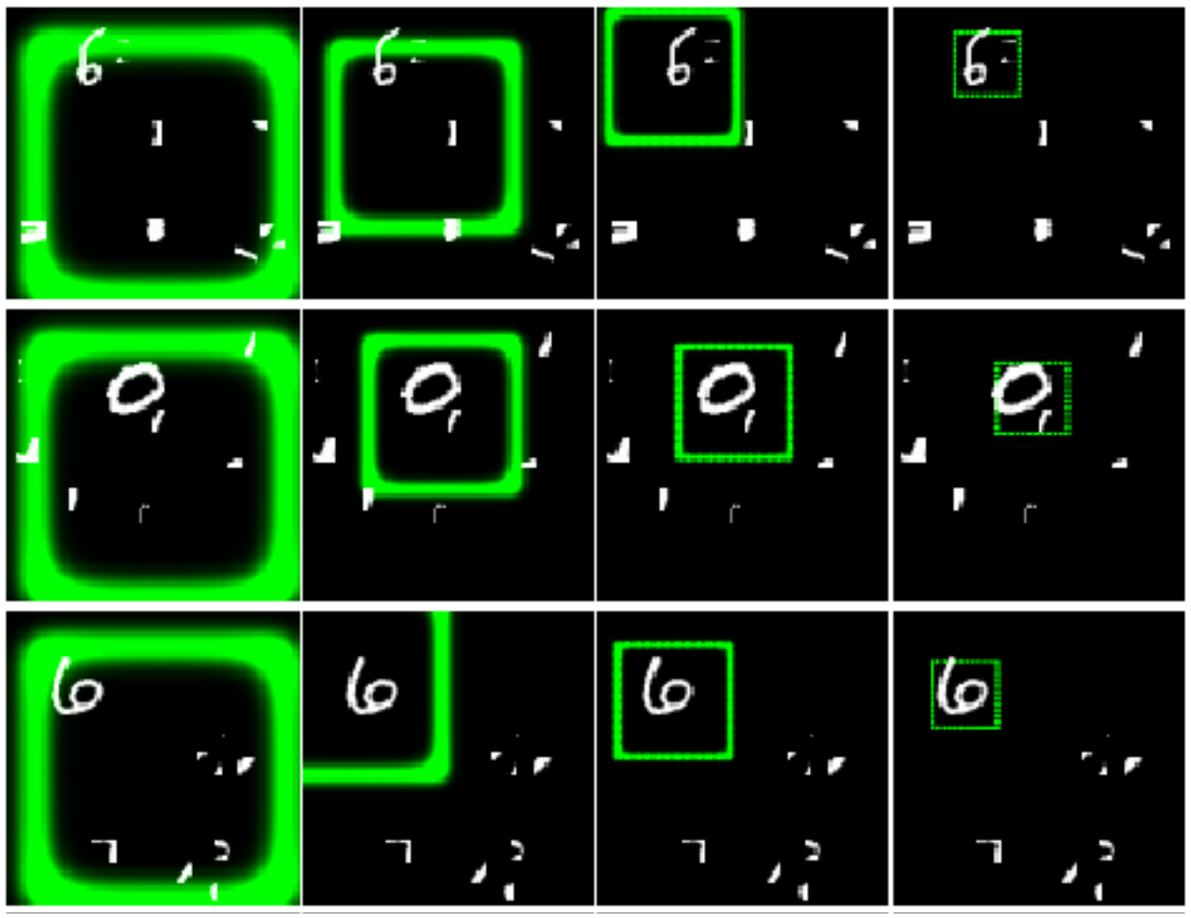
“zoom and shift” focus



anders@watty.io

Spatial Attention: 2D

“zoom and shift” focus



anders@watty.io

Spatial Attention: 2D

"zoom and shift" focus

A woman is throwing a **frisbee** in a park.

anders@watty.io

Differentiability

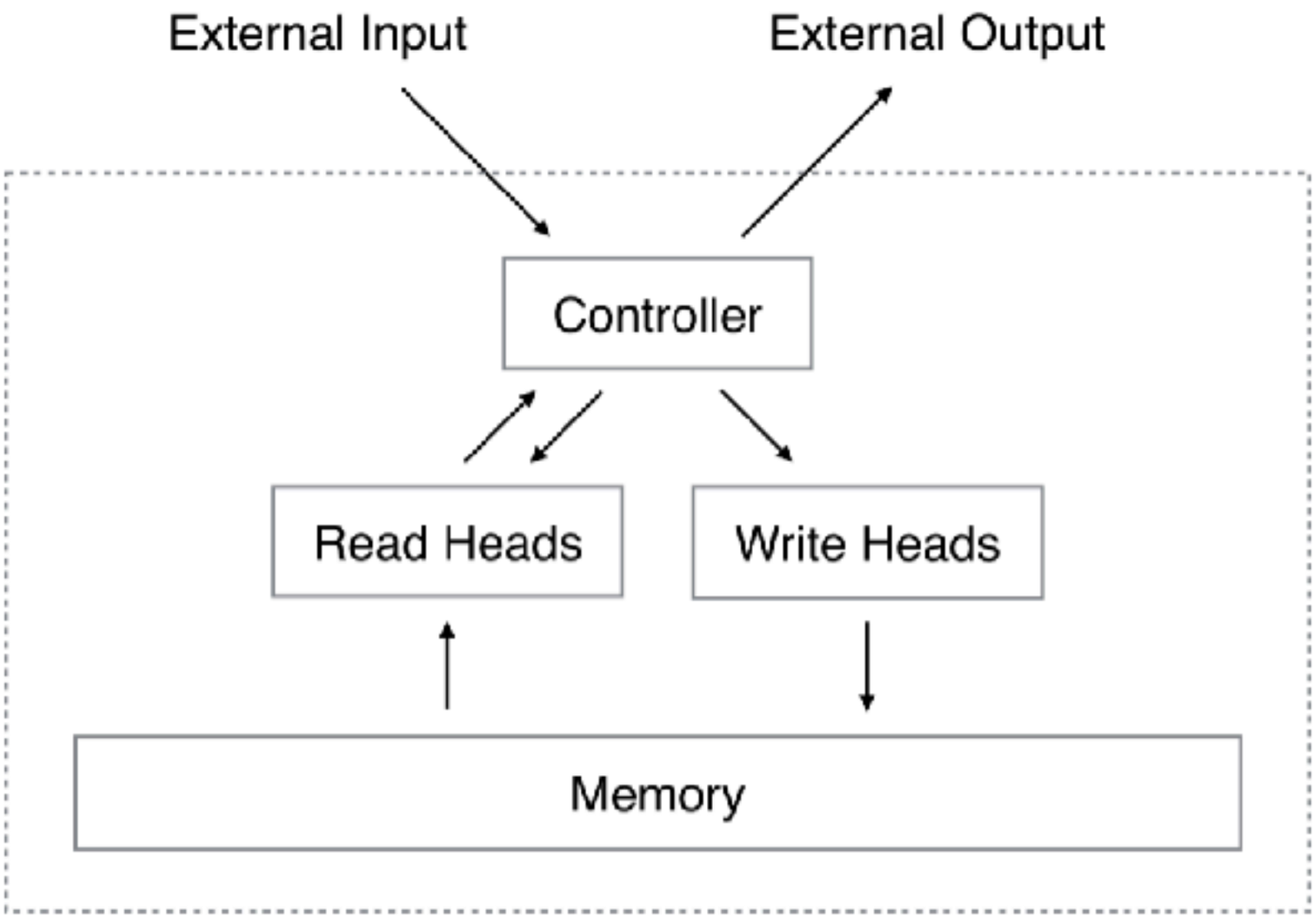
- Very convenient - we can train with SGD and Backprop as usual
- Obvious limitations regarding efficiency and size of the attended

Beyond Spatial Attention

- With attention, we can selectively make use of specific pieces of information when we need it.
- “Information retrieval problems”
- Add support for *storing* information at certain locations -> “Neural Touring Machine”

anders@watty.io

Neural Touring Machine

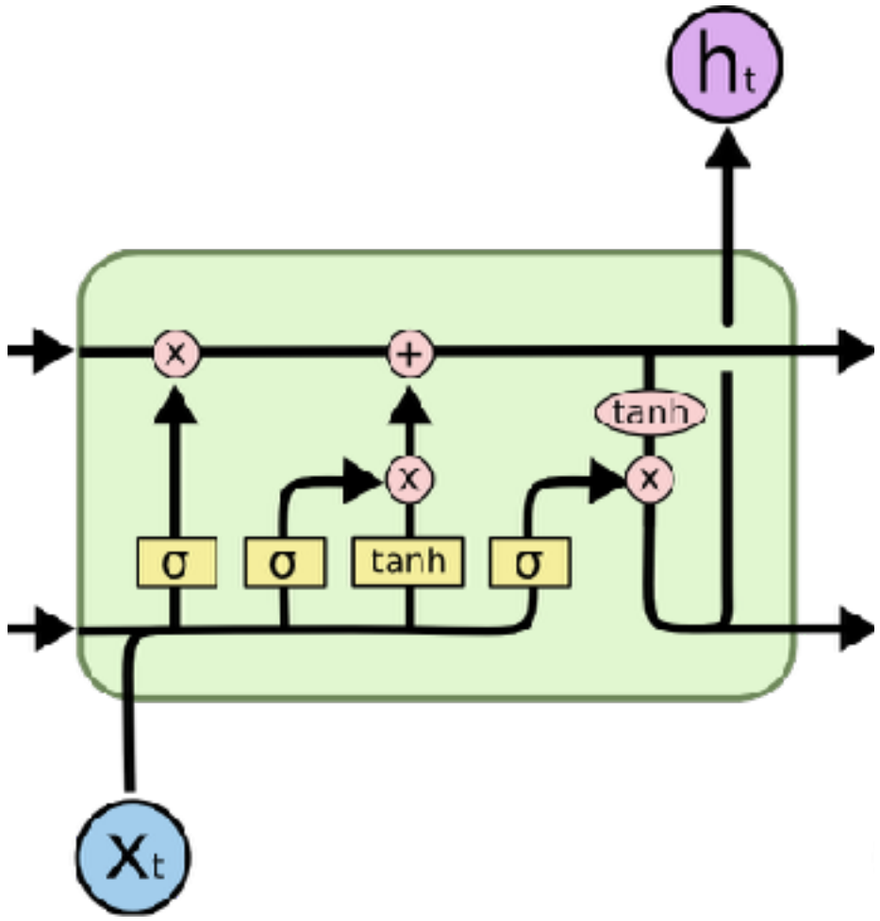


anders@watty.io

Neural Touring Machine

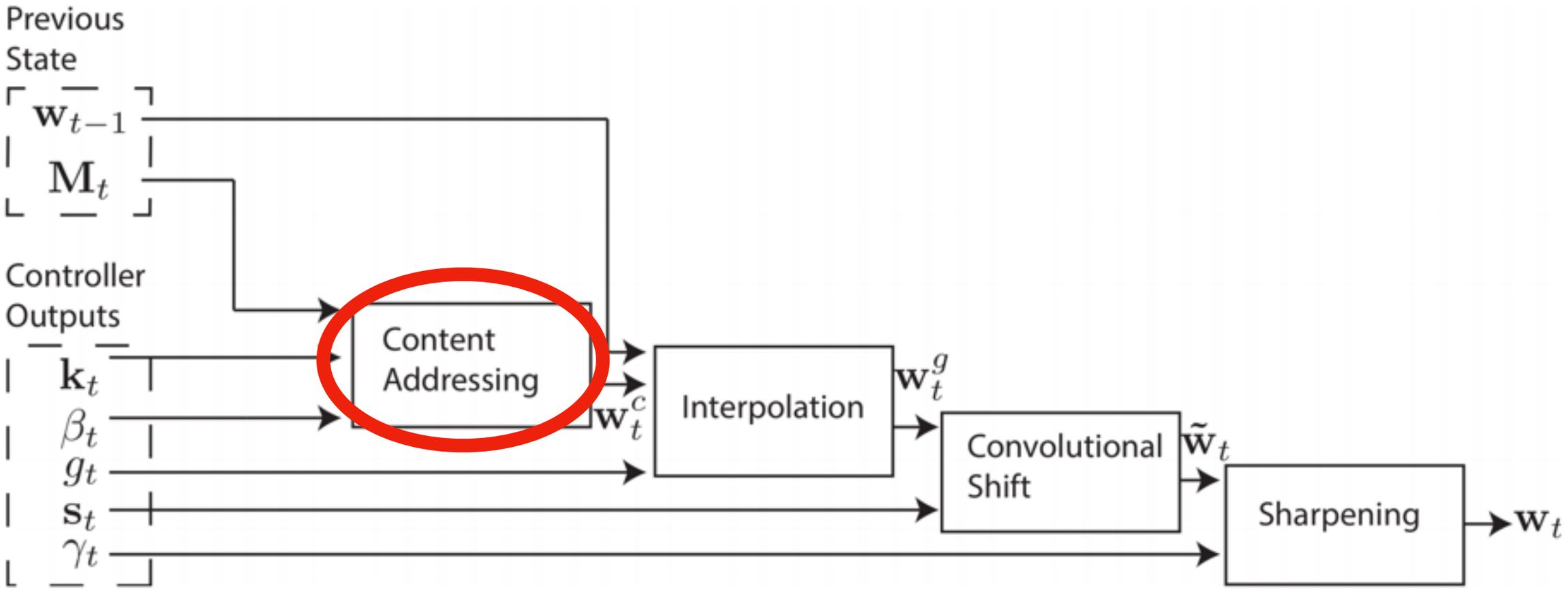
Comparison LSTM

- Similar concept
- NTM considerable more sophisticated (tailored) read/ write mechanisms
- size of NTM memory can be increase without increasing number of parameters to learn (grows quadratically in LSTM)



anders@watty.io

Neural Touring Machine



anders@watty.io

Questions/Discussion

anders@watty.io