Given a prompt, write an essay

Question: What fundamental economic and political change, if any, is needed for an effective response to climate change?

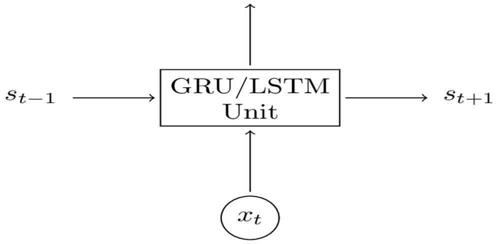
Do we want to go through the same process we have been through for decades with no changes?

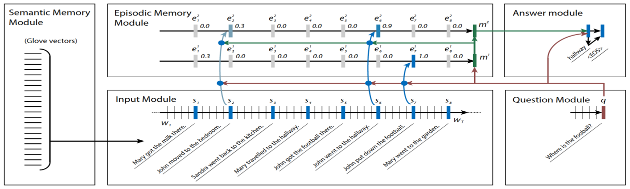
Is there a way to build a sustainable energy system that is both affordable and environmentally responsible? The basic premise behind this article is that we need to transform the economics of our energy system by investing in the necessary infrastructure so that it will be affordable for everyone.

…

https://www.economist.com/open-future/2019/10/01/how-to-respond-to-climate-change-if-you-are-an-algorithm

Recurrent Neural Networks (RNNs)





Lyle Ungar

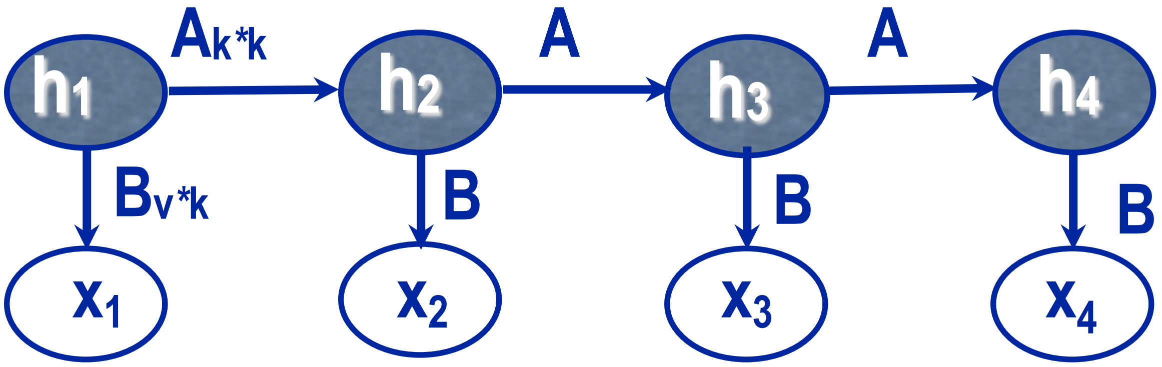
RNN model
Seq2seq
encoder/decoder
attention

Recurrent Neural Nets

- **Needed if you have inputs of varying length**
  - E.g. sequence of observations
    - speech
    - text
    - robots
    - power plants, chemical plants, data centers
- **Generalize HMMs or Linear Dynamical Systems**
  - Hidden state dynamical models, but *nonlinear*

Standard HMM

- **HMM learning problem: Estimate A and B**

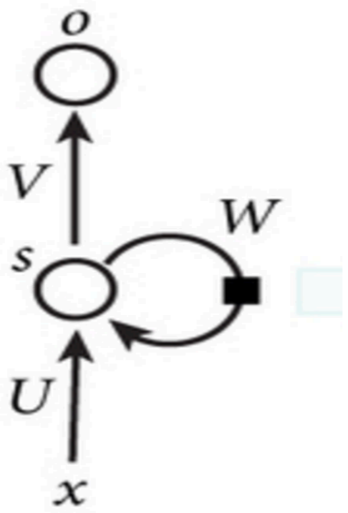


A = Markov transition matrix
B = emission matrix

- **Estimation done via EM**
  - Or spectral methods
- **History is forgotten with an exponential decay**

Simple Recurrent Neural Net

$$s_t = \tanh(Ux_t + Ws_{t-1})$$
$$o_t = \text{softmax}(Vs_t)$$

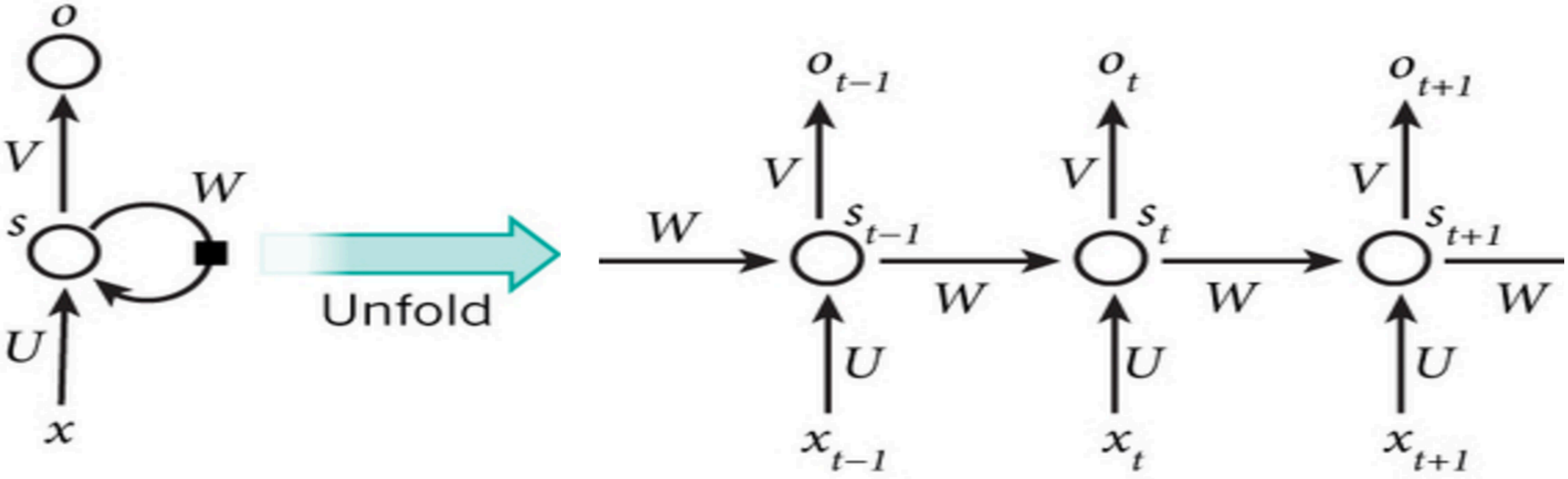


x_t = input (e.g. a word)
s_t = hidden state
o_t = output (e.g. probability of the next word)
y_t = true value (e.g. x_{t+1})

Softmax $\sigma(\mathbf{z})$ transforms the K-dimensional real valued output $\mathbf{z}$ to a distribution – *like logistic regression*

$$\sigma(\mathbf{z})_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \quad \text{for } j = 1, \ldots, K.$$

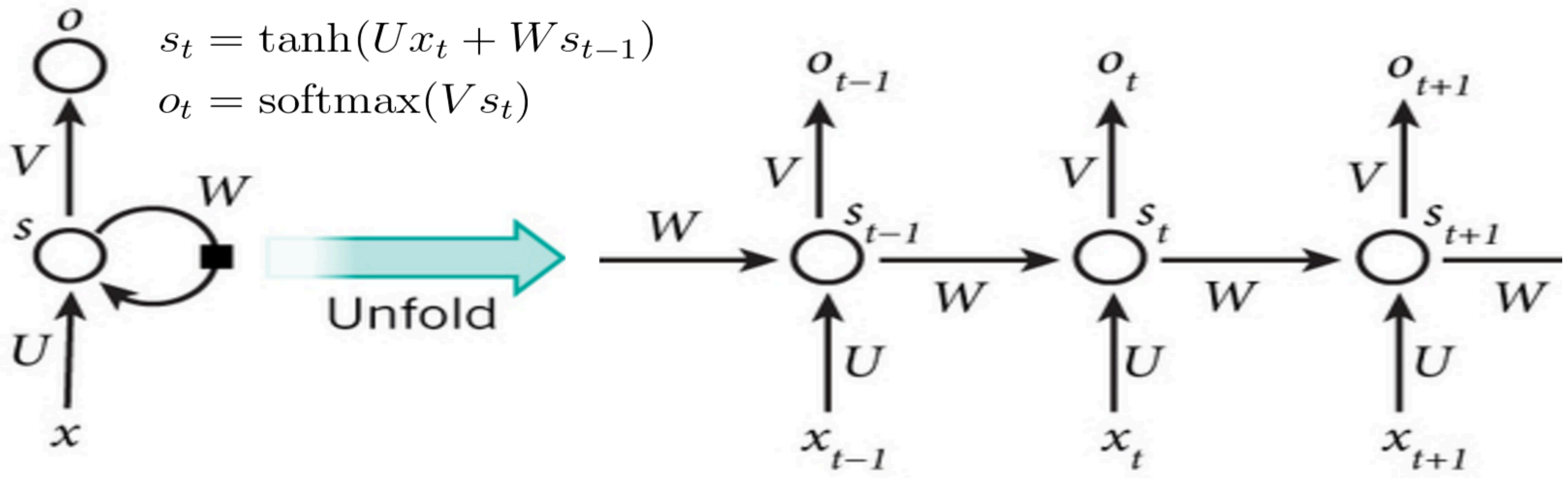
Like HMMs, unroll RNNs in time



x_t = input (e.g. a word)
s_t = hidden state
o_t = output (e.g. probability of the next word)

http://www.nature.com/nature/journal/v521/n7553/full/nature14539.html

Like HMMs, unroll RNNs in time



$$s_t = \tanh(Ux_t + Ws_{t-1})$$

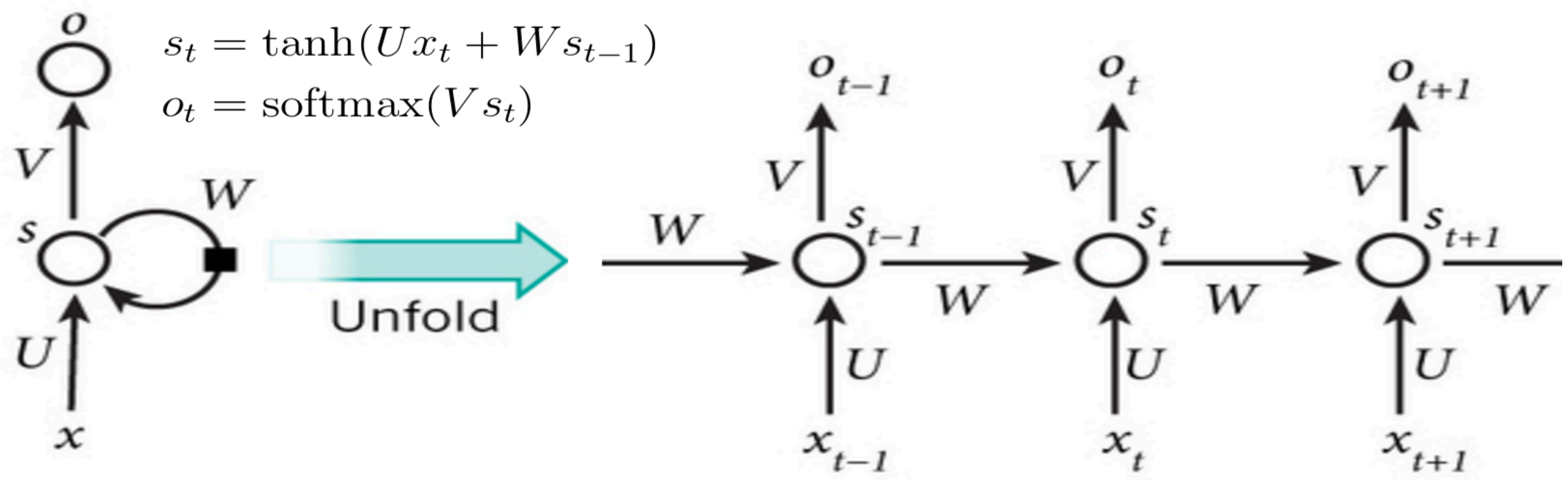
$$o_t = \mathrm{softmax}(Vs_t)$$

x_t = input (e.g. a word) - v
s_t = hidden state - k
o_t = output - v

What are the dimensions of *U, W, V*?

U: *k*v* ***W:*** *k*k* ***V:*** *v*k*

Like HMMs, unroll RNNs in time



$$s_t = \tanh(Ux_t + Ws_{t-1})$$
$$o_t = \mathrm{softmax}(Vs_t)$$

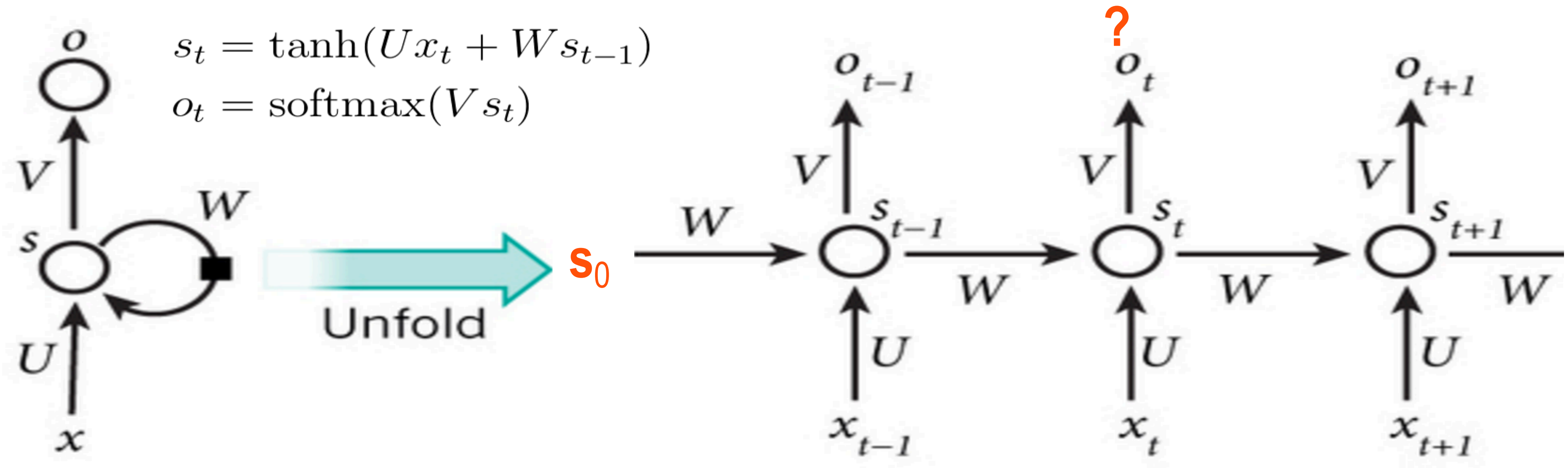
x_t = input (e.g. a word) - v
s_t = hidden state - k
o_t = output - v

What is the usual loss function?

$-\Sigma_t \log(\mathbf{o}_t[y_t])$ *- est. prob.of truth*

where $y_t=i$ gives the true label

Like HMMs, unroll RNNs in time



$$s_t = \tanh(Ux_t + Ws_{t-1})$$

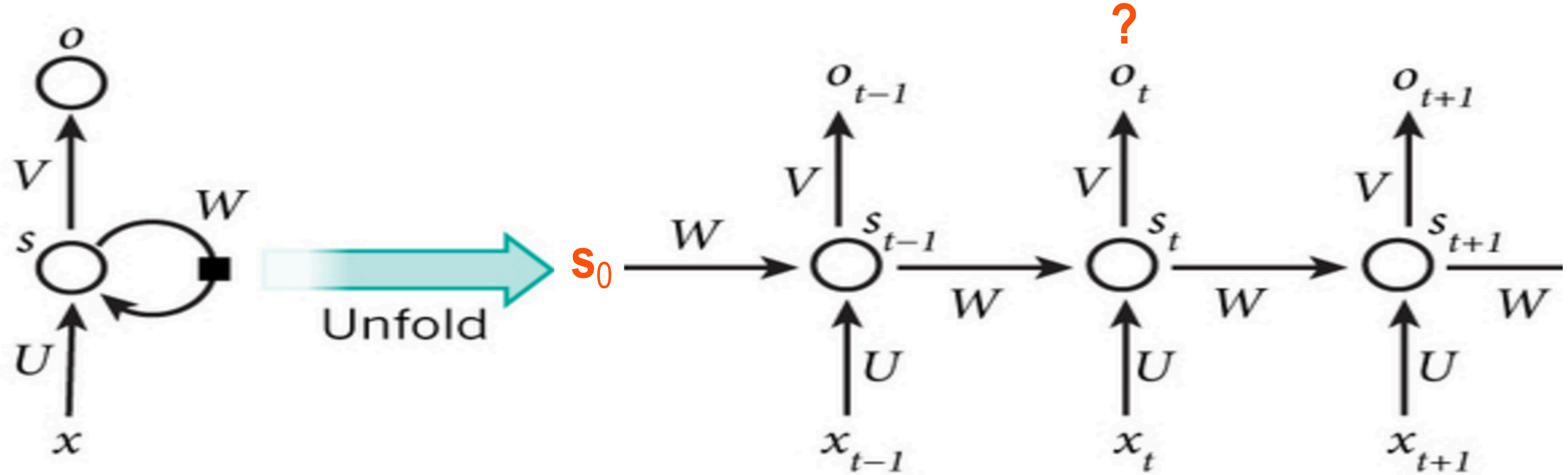
$$o_t = \mathrm{softmax}(Vs_t)$$

x_t = input - v
s_t = hidden state - k
o_t = output - v

If $s_{t-2} = s_0$, what is o_t in terms of s_0 and x?

$$o_t = \mathrm{softmax}(Vs_t) = \mathrm{softmax}(V \tanh(Ux_t + Ws_{t-1}))$$
$$= \mathrm{softmax}(V \tanh(Ux_t + W \tanh(Ux_{t-1} + Ws_{t-2})))$$

RNN gradients



$\mathbf{o}_t = softmax(\mathbf{V}\ tanh(\mathbf{U}\mathbf{x}_t + \mathbf{W}\ tanh(\mathbf{U}\mathbf{x}_{t-1} + \mathbf{W}\mathbf{s}_{t-2})$

Observe $y_t = i$ What is the stochastic gradient step?

$Err = -log(\mathbf{o}_t[i])$

Find $d\ Err/d\mathbf{V}$, $d\ Err/d\mathbf{U}$, $d\ Err/d\mathbf{W}$

RNN Gradients

- $\mathbf{o}_t = softmax(\mathbf{V}\, tanh(\mathbf{U}\mathbf{x}_t + \mathbf{W}\, tanh(\mathbf{U}\mathbf{x}_{t-1} + \mathbf{W}\mathbf{s}_{t-2})$
- **Observe** $y_t = i$ What is the stochastic gradient step?
- $Err = -log(\mathbf{o}_t[i])$

$$d\, Err/d\mathbf{V} = -(d\, log(\mathbf{o}_t[i]/d\mathbf{o}_t[i]) \quad d\mathbf{o}_t[i]/d\mathbf{V}$$

$$= -(1/\mathbf{o}_t[i]) \qquad d\, softmax(\mathbf{z})/d\mathbf{z} \quad dz/d\mathbf{V}$$

$$\mathbf{z} = \mathbf{V}\, tanh(\mathbf{U}\mathbf{x}_t + \mathbf{W}\, tanh(\mathbf{U}\mathbf{x}_{t-1} + \mathbf{W}\mathbf{s}_{t-2})$$

$$d\, softmax(\mathbf{z})/dz_j = -1/(\Sigma_k e^{z_k})^2\ e^{z_j}\ e^{z_k} \qquad \text{for k not equal to j}$$

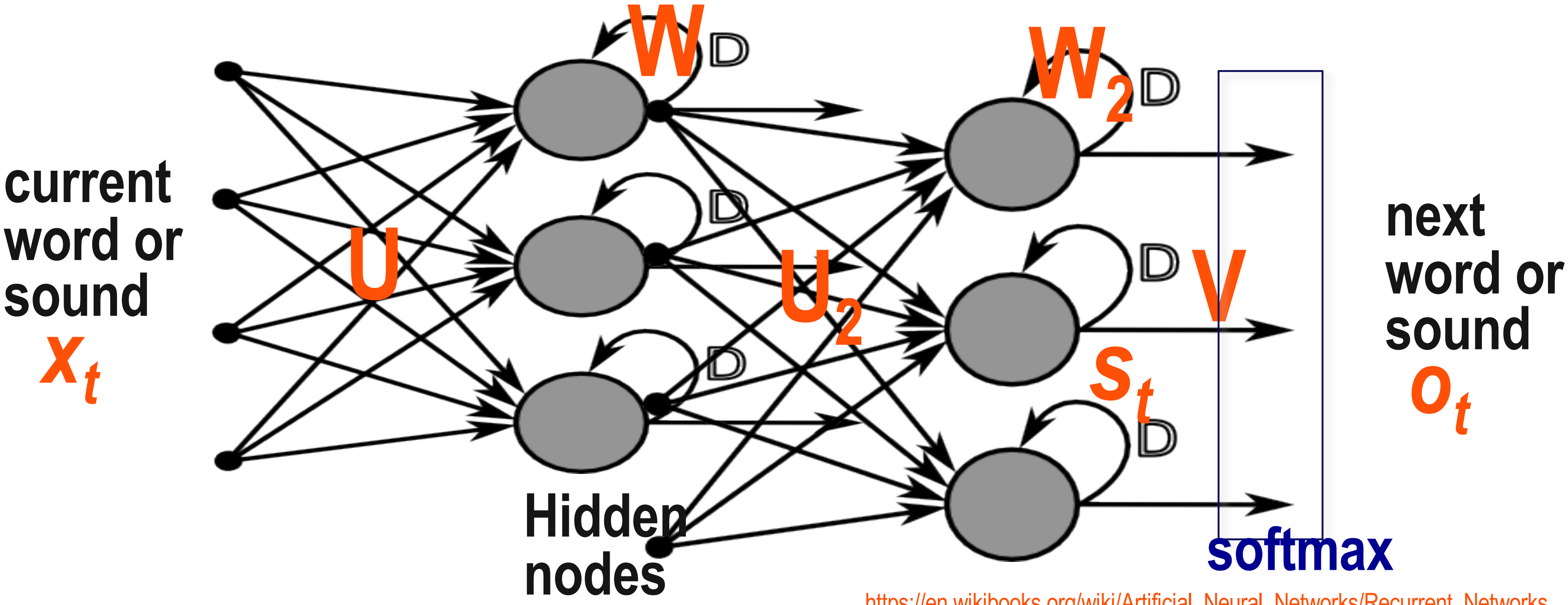
$$= -1/(\Sigma_k e^{z_k})^2\ e^{2z_j} + e^{z_j}/(\Sigma_k e^{z_k}) \qquad \text{for k=j}$$

$$\sigma(\mathbf{z})_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \quad \text{for } j = 1, \ldots, K.$$

Recurrent Neural Nets (RNNs)

$$s_t = \tanh(Ux_t + Ws_{t-1})$$
$$o_t = \text{softmax}(Vs_t)$$

Can use multiple layers



https://en.wikibooks.org/wiki/Artificial_Neural_Networks/Recurrent_Networks

Gated RNNs

- Standard RNNs, like HMMs, tend to forget things exponentially fast
- Solution: Gated RNN
  - Stores hidden state

You don’t need to know this; it’s just a bunch of weights and transformations.

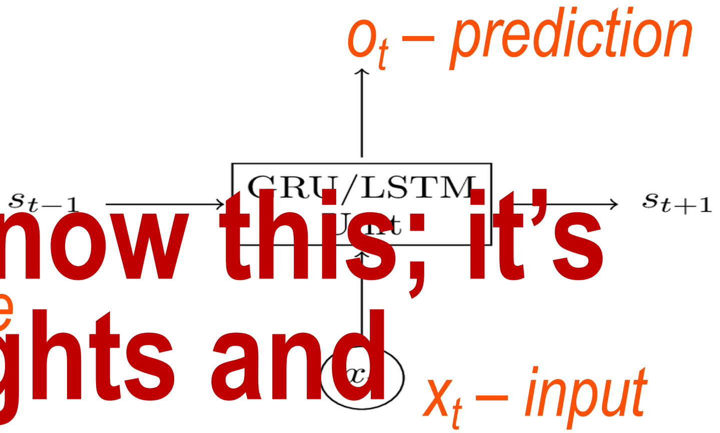
$z = \sigma(U^z x_t + W^z s_{t-1})$ z: update gate

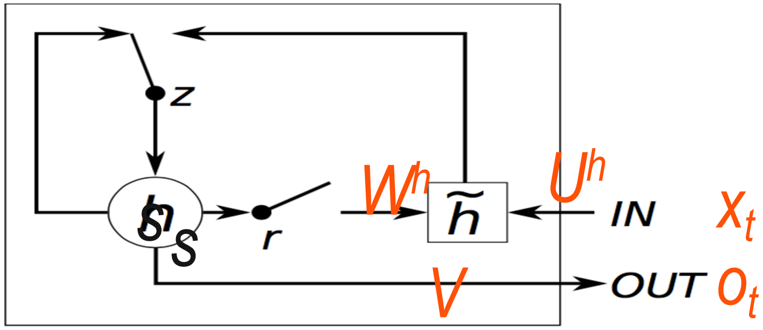
$r = \sigma(U^r x_t + W^r s_{t-1})$ r: reset gate

$h = \tanh(U^h x_t + W^h(s_{t-1} \circ r))$

$s_t = (1-z) \circ h + z \circ s_{t-1}$ s_t : hidden state

r=0 resets h
z=1 keeps state
z=0 updates it to h
r=1’s, z=0’s gives simple RNN
$\circ$ *is pointwise multiplication*



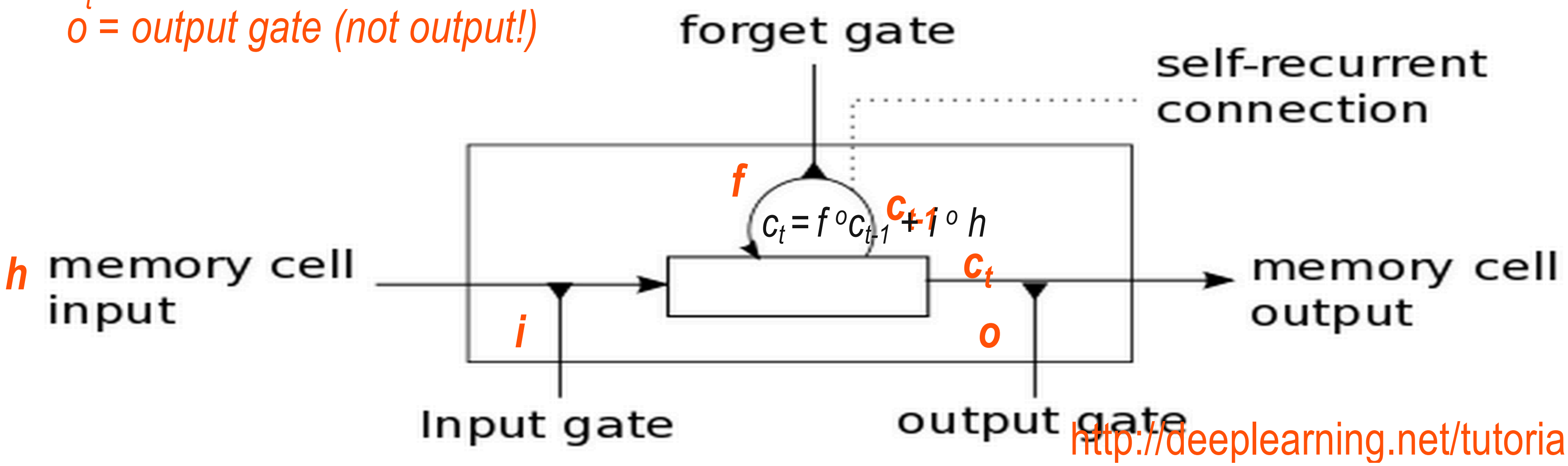


http://deeplearning.net/tutorial/lstm.html

Long Short Term Memory (LSTM)

- **LSTM is a kind of gated RNN**
  - Just with more, different gates
  - **Don’t worry about what they are!!!**

x_t – observation
s_t – hidden state
o = output gate (not output!)

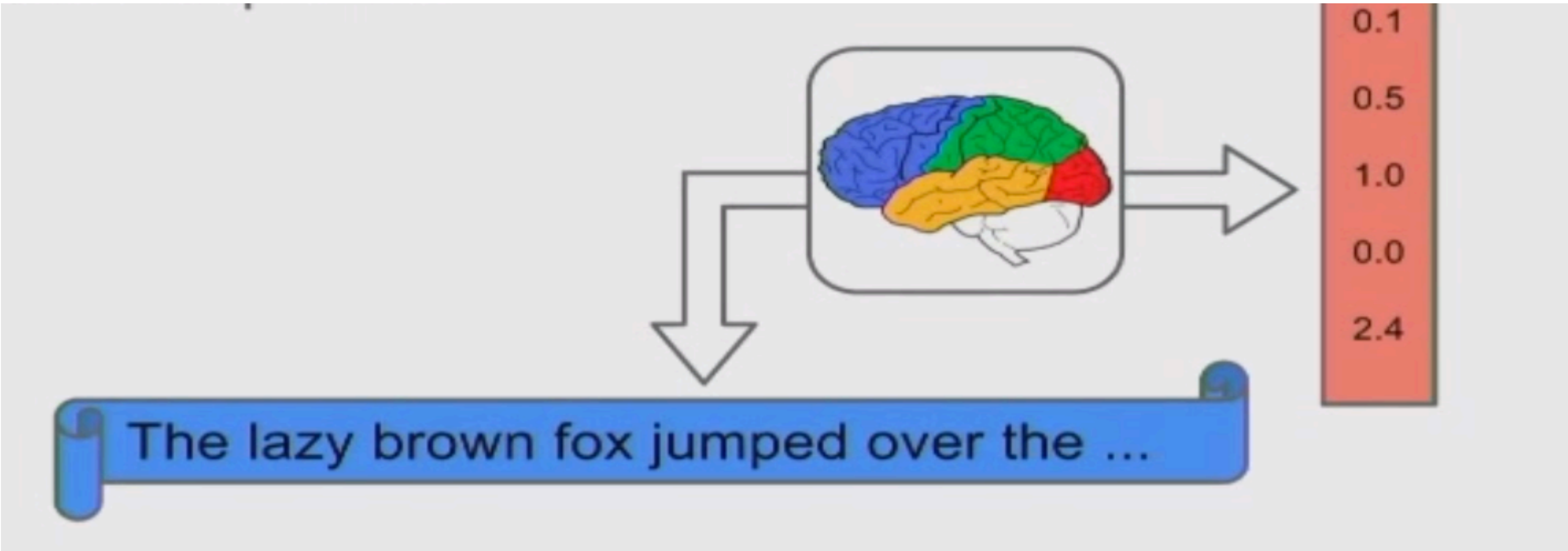


http://deeplearning.net/tutorial/lstm.html

Recurrent Neural Nets

- **Predict a label for each observation**
  - $\mathbf{y}_t = f(\mathbf{x}_t, \mathbf{s}_t)$
- **Predict the next observation given past observations**
  - $\mathbf{y}_t = \mathbf{x}_{t+1} = f(\mathbf{x}_t, \mathbf{s}_t)$
- **Or map one sequence to another sequence**
  - **An encoder**
    - sentence (sequence of words) to vector
  - **A decoder**
    - vector to sentence (sequence of words)

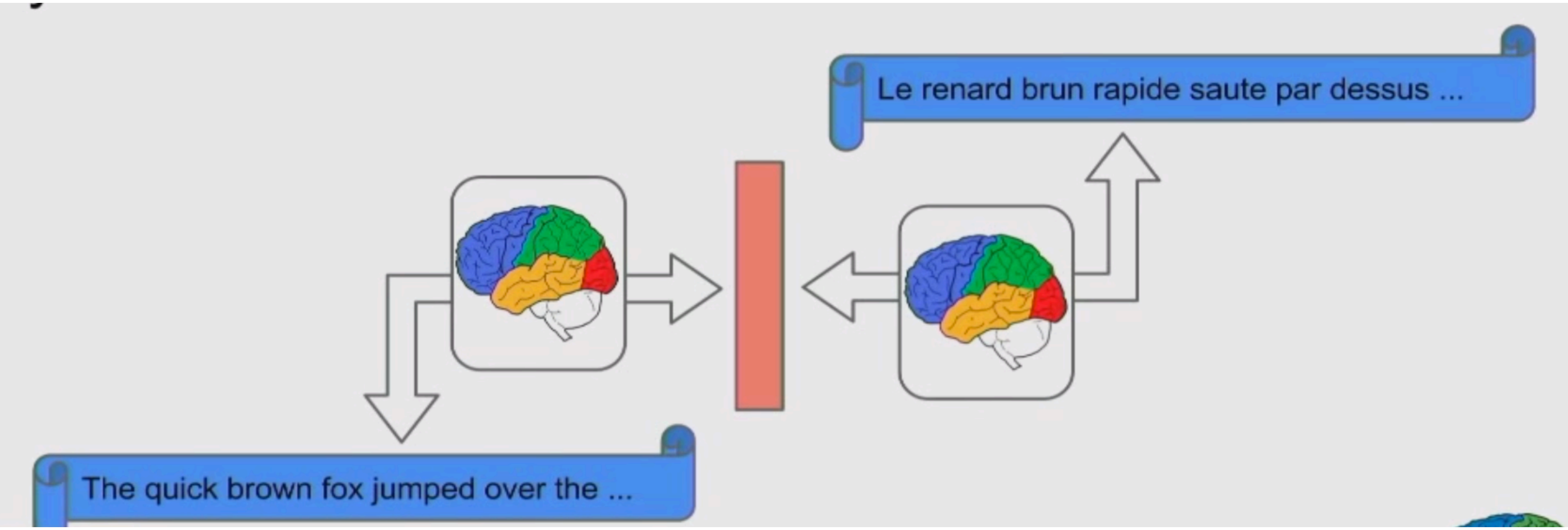
LSTM encodes a sentence



Jeff Dean, google

https://www.youtube.com/watch?v=90-S1M7Ny_o&spfreload=1

Sequence to sequence (Seq2seq)



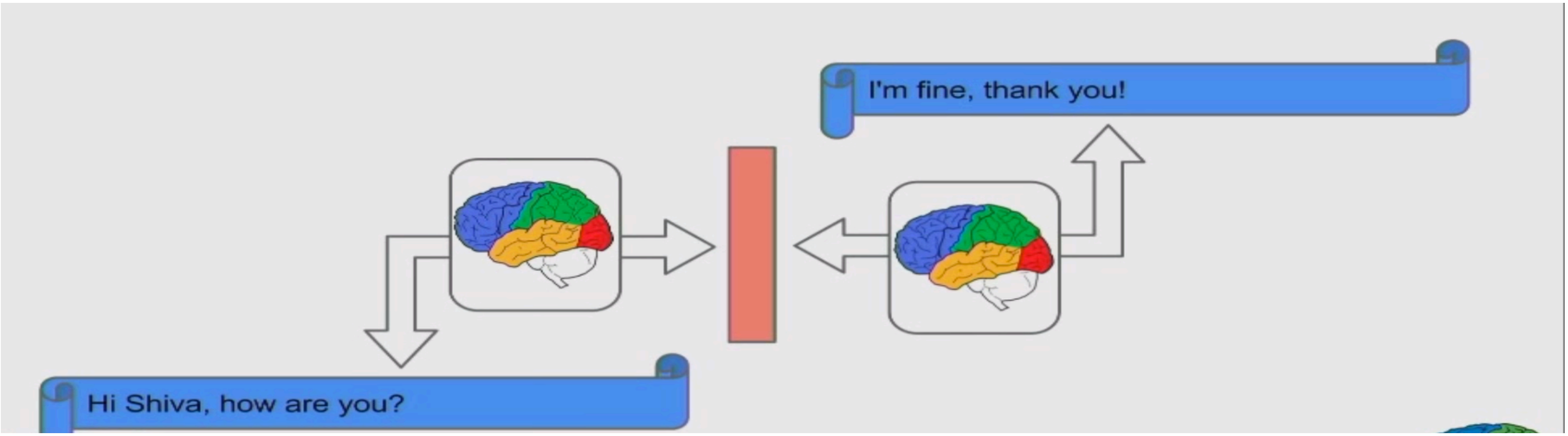
Le renard brun rapide saute par dessus ...

The quick brown fox jumped over the ...

Encode and Decode = translate

Jeff Dean, google

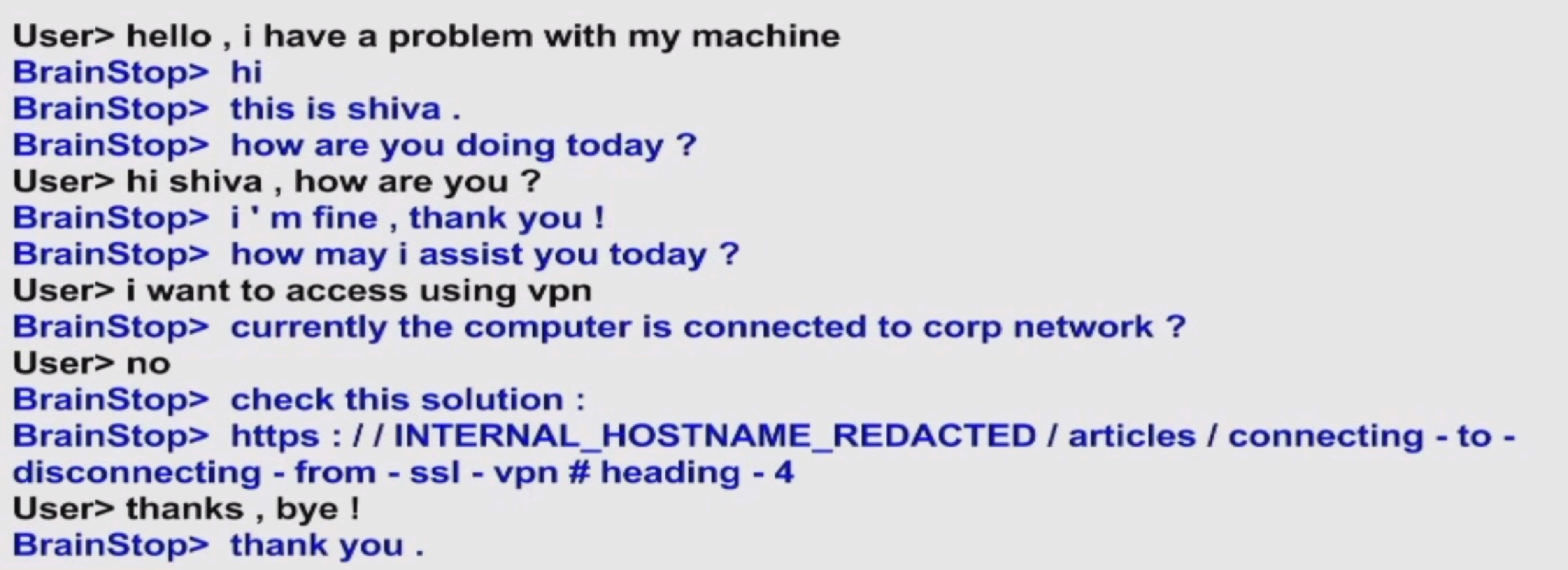
Seq2seq chatbot



Encode and Decode = chatbot

Jeff Dean, google

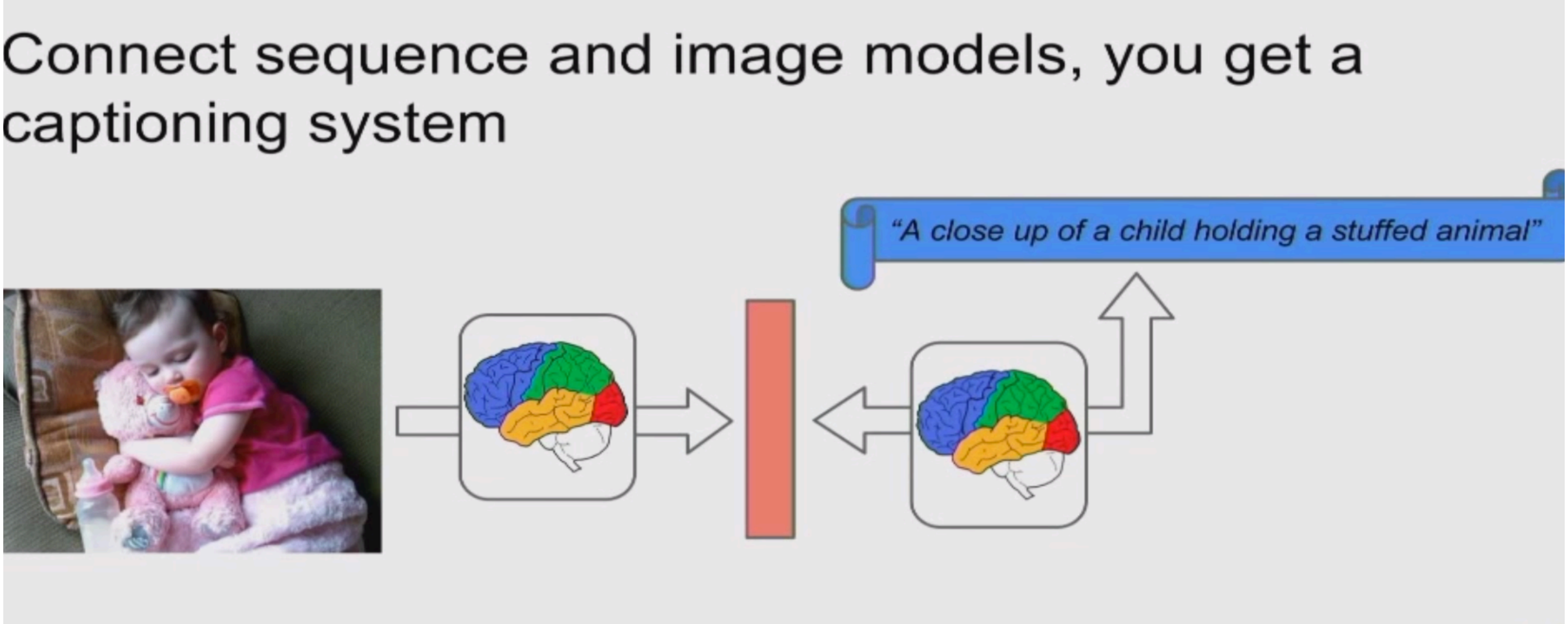
It works (sort of)



```
User> hello , i have a problem with my machine
BrainStop>  hi
BrainStop>  this is shiva .
BrainStop>  how are you doing today ?
User> hi shiva , how are you ?
BrainStop>  i ' m fine , thank you !
BrainStop>  how may i assist you today ?
User> i want to access using vpn
BrainStop>  currently the computer is connected to corp network ?
User> no
BrainStop>  check this solution :
BrainStop>  https : / / INTERNAL_HOSTNAME_REDACTED / articles / connecting - to -
disconnecting - from - ssl - vpn # heading - 4
User> thanks , bye !
BrainStop>  thank you .
```

Jeff Dean, google

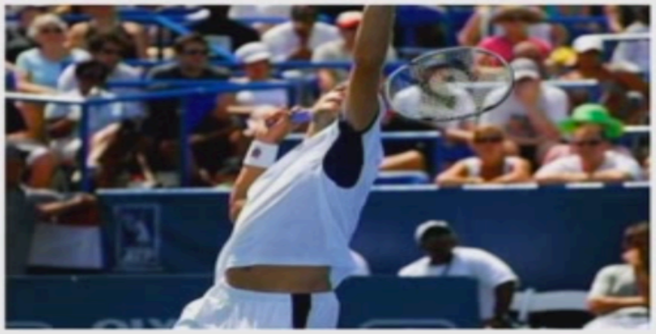
Can map images to words



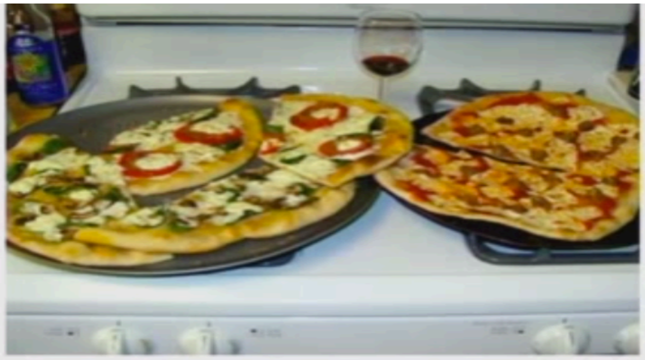
Connect sequence and image models, you get a captioning system

"A close up of a child holding a stuffed animal"

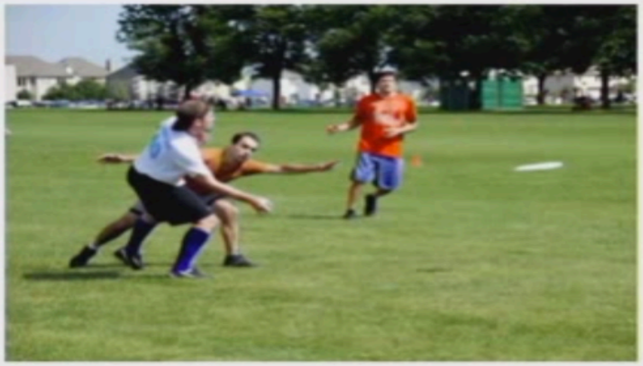
It works (sort of)



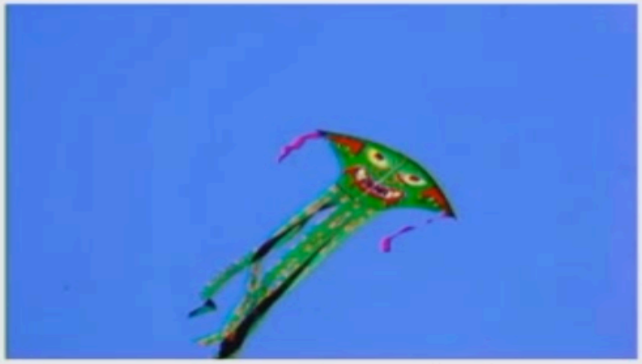
A man holding a tennis racquet on a tennis court.



Two pizzas sitting on top of a stove top oven



A group of young people playing a game of Frisbee



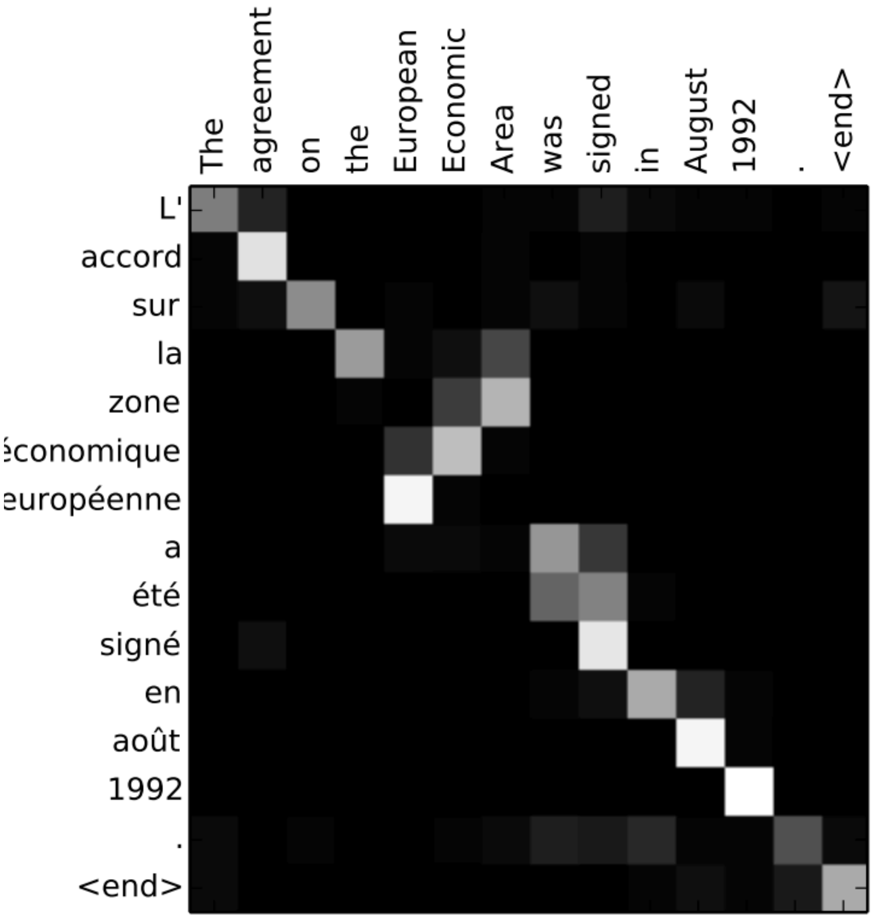
A man flying through the air while riding a snowboard

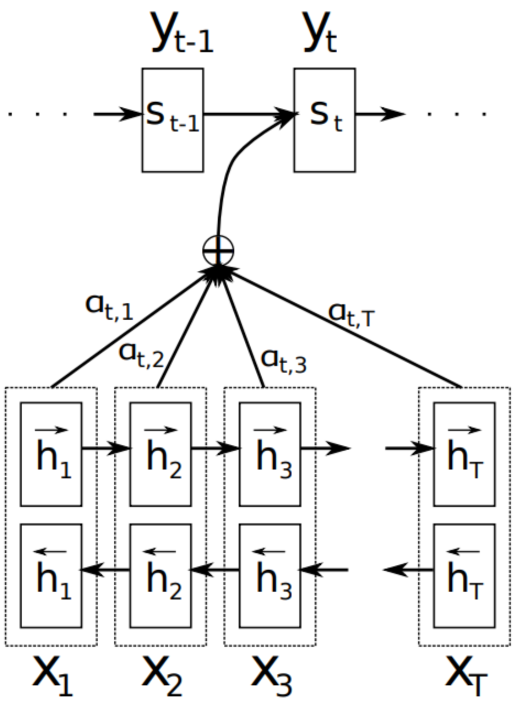
Jeff Dean, google

Language inputs to RNNs

- Words (“one-hot”)
- Characters (“one-hot”)
- Bytecodes (“one-hot”)
- Word embeddings
  - Typically 300 dimensional

Attention-based Q&A





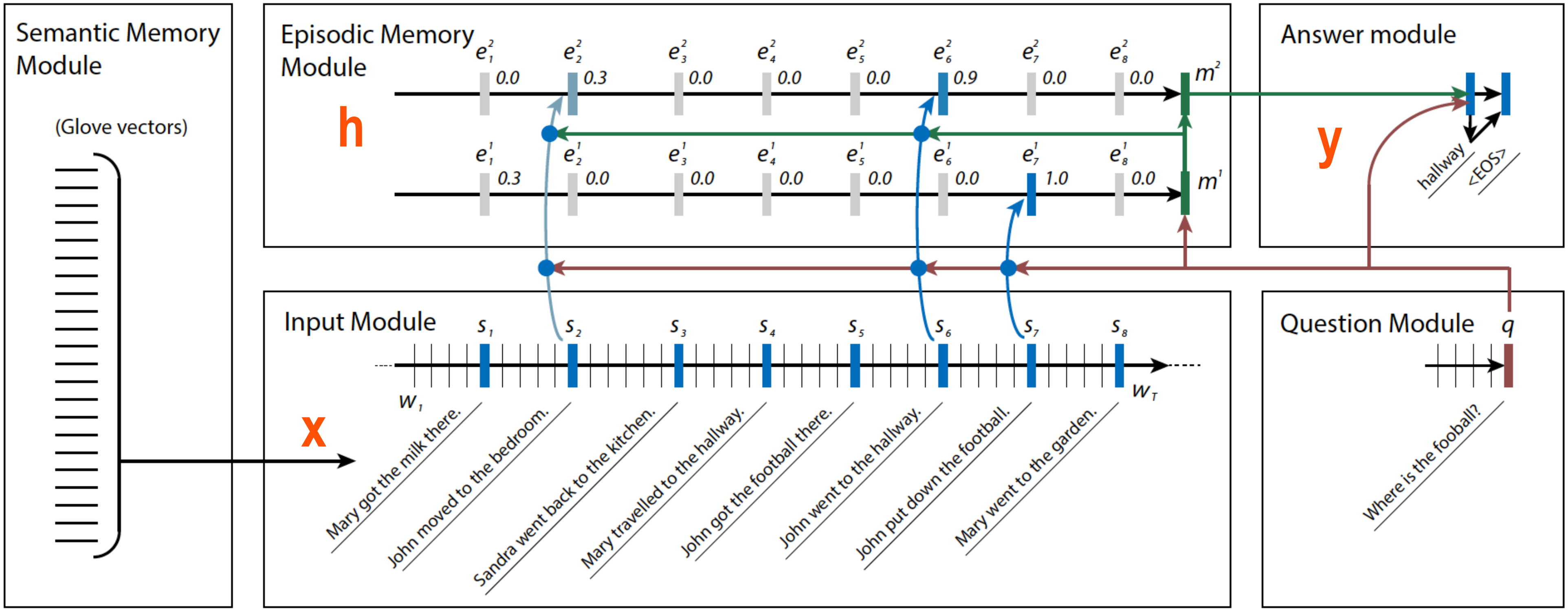
Target output

Forward and backward hidden states

Input

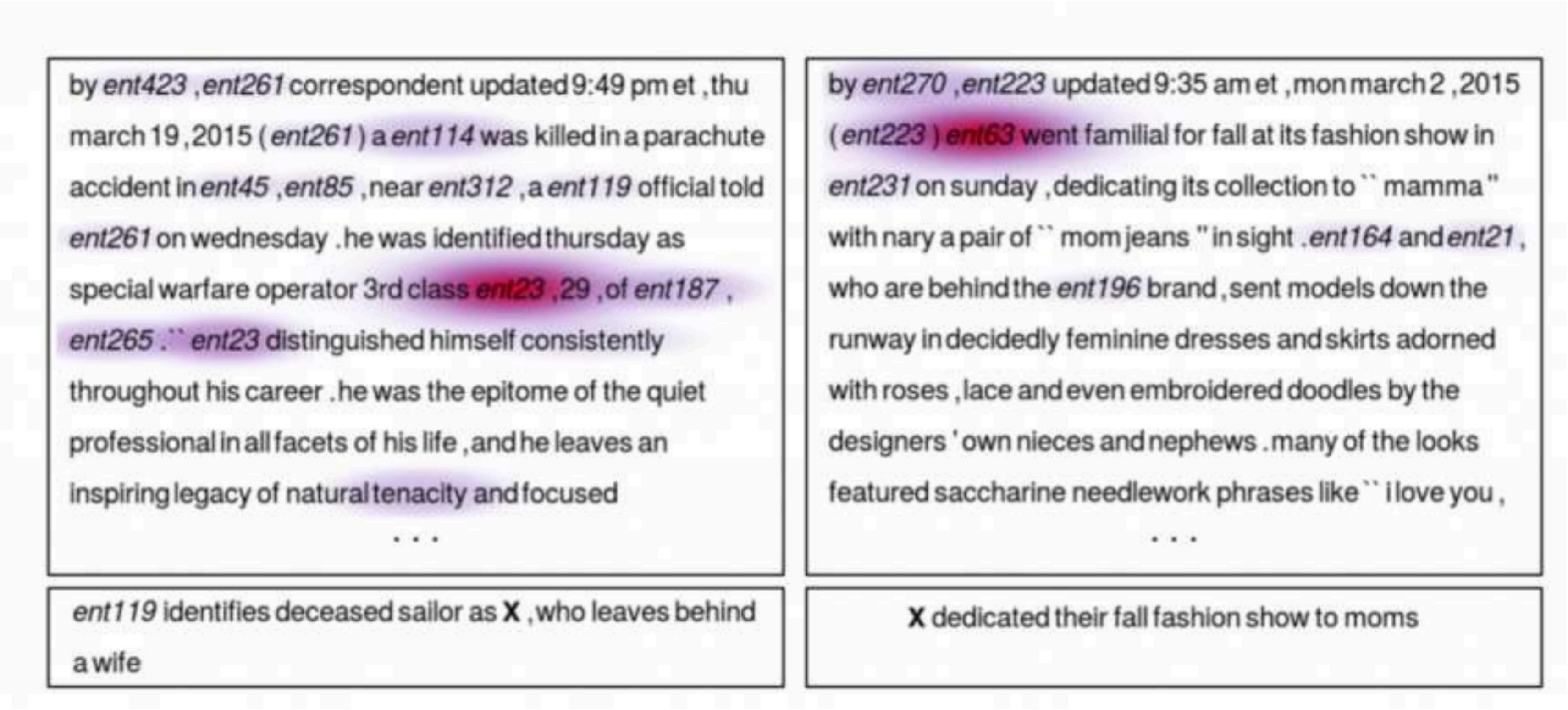
Neural machine translation by jointly learning to align and translate 2015

Attention-based Q&A



Ask Me Anything: Dynamic Memory Networks for Natural Language Processing

Attention-based Q&A



by *ent423* , *ent261* correspondent updated 9:49 pm et , thu march 19 , 2015 (*ent261*) a *ent114* was killed in a parachute accident in *ent45* , *ent85* , near *ent312* , a *ent119* official told *ent261* on wednesday . he was identified thursday as special warfare operator 3rd class *ent23* , 29 , of *ent187* , *ent265* . `` *ent23* distinguished himself consistently throughout his career . he was the epitome of the quiet professional in all facets of his life , and he leaves an inspiring legacy of natural tenacity and focused

. . .

ent119 identifies deceased sailor as **X** , who leaves behind a wife

by *ent270* , *ent223* updated 9:35 am et , mon march 2 , 2015 (*ent223*) *ent63* went familial for fall at its fashion show in *ent231* on sunday , dedicating its collection to `` mamma '' with nary a pair of `` mom jeans '' in sight . *ent164* and *ent21* , who are behind the *ent196* brand , sent models down the runway in decidedly feminine dresses and skirts adorned with roses , lace and even embroidered doodles by the designers ' own nieces and nephews . many of the looks featured saccharine needlework phrases like `` i love you ,

. . .

X dedicated their fall fashion show to moms

Teaching Machines to Read and Comprehend 2015

Attention-based Q&A

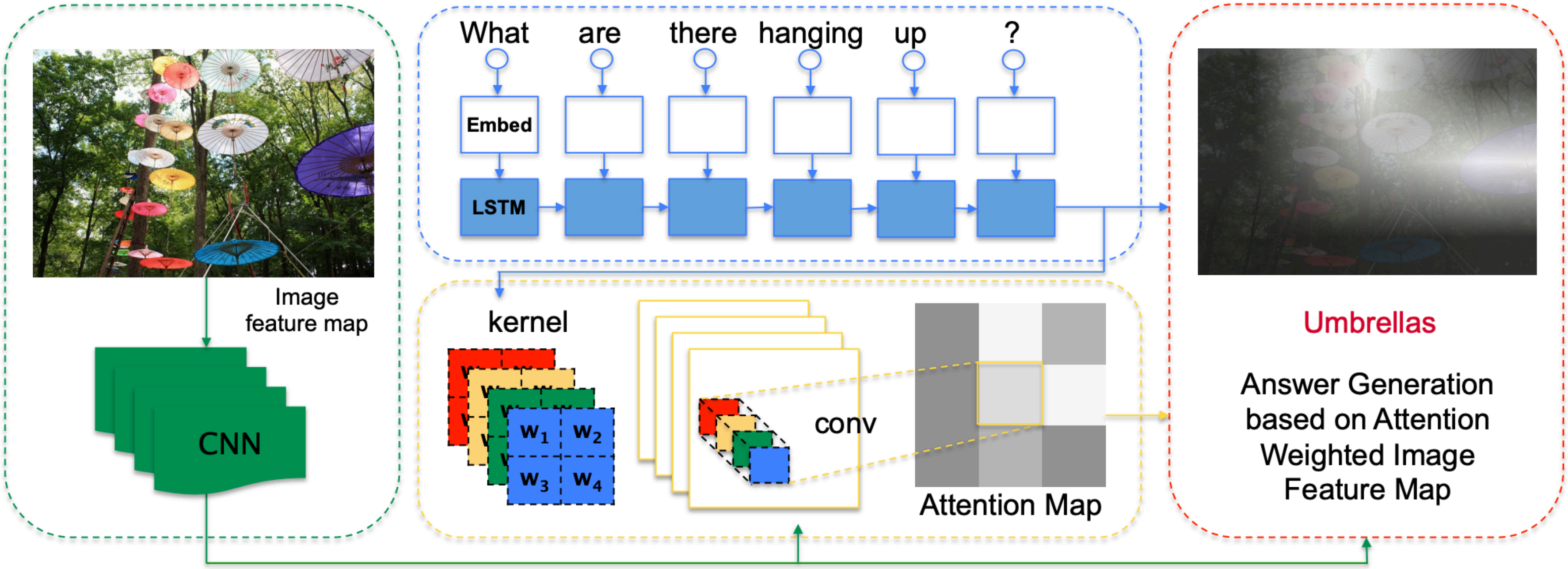
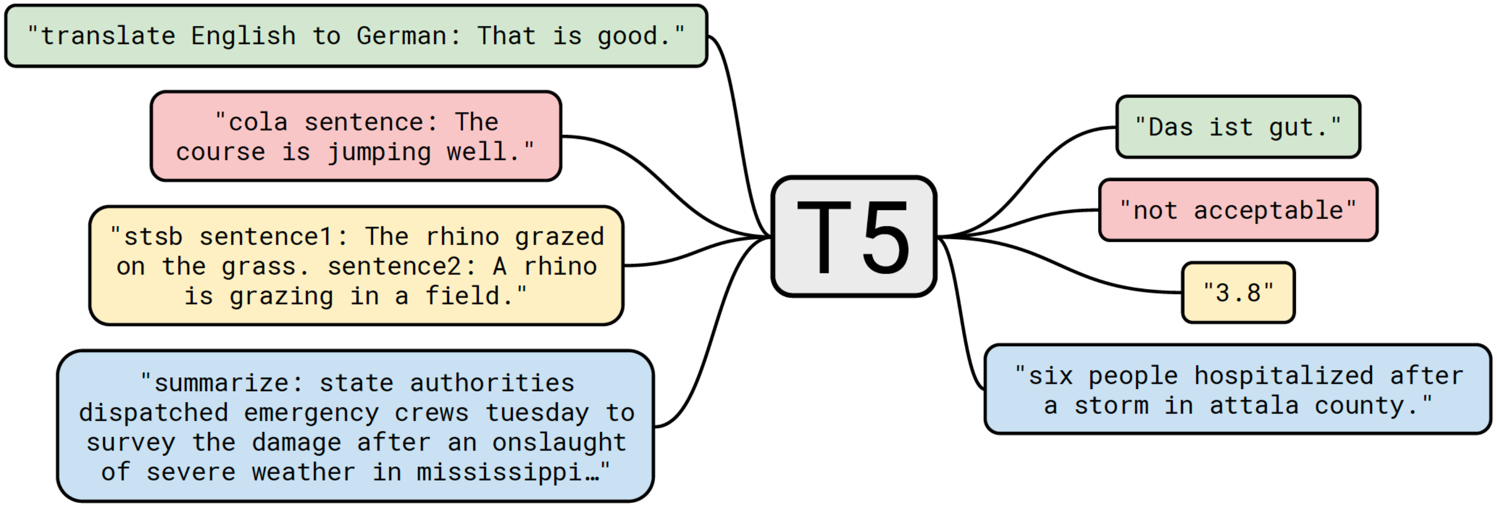


Figure 2. The framework of ABC-CNN. The green box denotes the image feature extraction part using CNN; the blue box is the question understanding part using LSTM; the yellow box illustrates the attention extraction part with configurable convolution; the red box is the answer generation part using multi-class classification based on attention weighted image feature maps. The orange letters are corresponding variables explained in Eq. (1) - (6).

ABC-CNN: An Attention Based Convolutional Neural Network for Visual Question Answering 2016

Transformer – seq2seq extension



"translate English to German: That is good."

"cola sentence: The course is jumping well."

"stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."

"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi…"

T5

"Das ist gut."

"not acceptable"

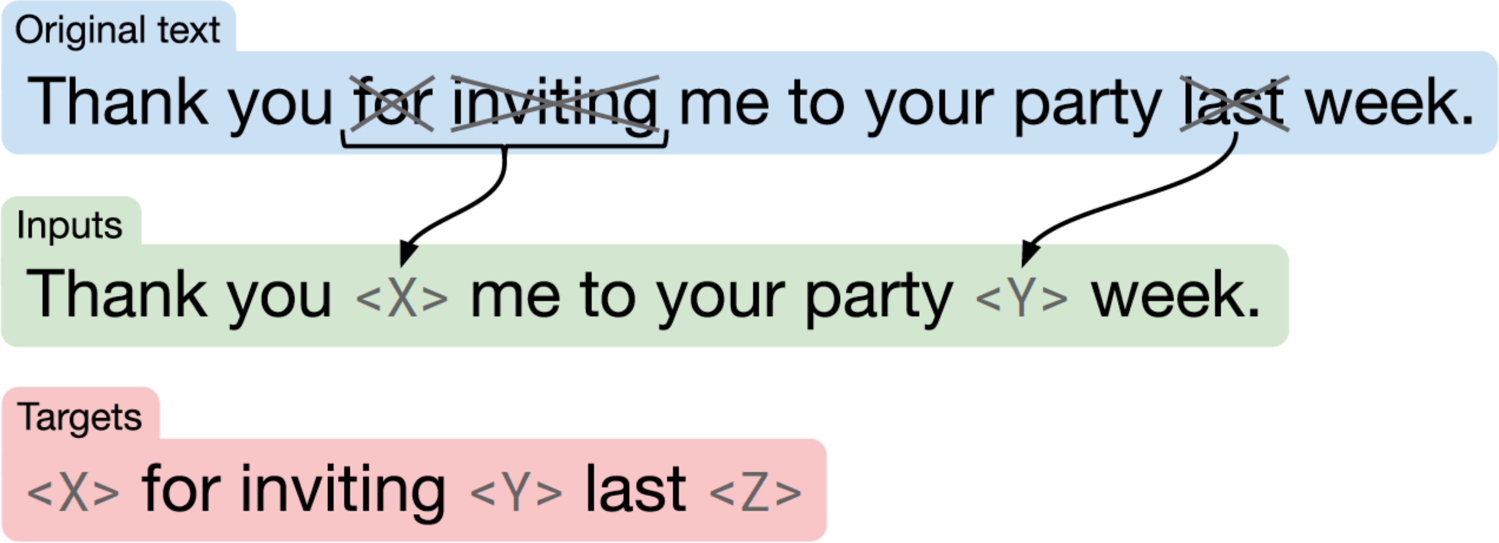
"3.8"

"six people hospitalized after a storm in attala county."

Uses "self attention"

Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer 2019 – building on Attention is All you Need

Train using "denoising"



Original text

Thank you ~~for inviting~~ me to your party ~~last~~ week.

Inputs

Thank you <X> me to your party <Y> week.

Targets

<X> for inviting <Y> last <Z>

Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, 2019

Big data

- **750 GB text**
- **Base model: 220 million parameters**
  - Each in encoder and decoder
- **Big model: 11 billion parameters**

Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, 2019

Generate text from language model

- **Input: prompt**
- **Output: text**
- **https://talktotransformer.com/**

https://transformer.huggingface.co/
https://gpt2.apps.allenai.org/?text=Joel%20is
https://demo.allennlp.org/next-token-lm?text=Lyle%20teaches

Dynamic Network Summary

- **Gated Neural Nets generalize HMMs, Kalman filters**
  - But are far more powerful!
- **They have replaced HMMs for speech to text and machine translation**
- **Lots of black magic “engineering”**
  - Unclear what matters about the network structure
    - Number and size of layers, regularization
    - Forms of gating (LSTM …), attention …
  - Gradient descent is tricky
- **Good software: tensorflow, pytorch …**