Given a prompt, write an essay

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

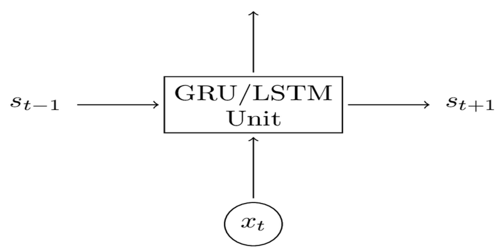
Response: Do we want to go through the same process we have been through for decades with no changes?

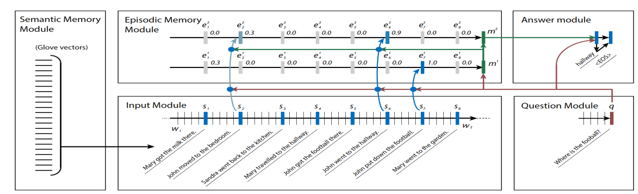
Is there a way to build a sustainable energy system that is both affordable and environmentally responsible? The basic premise behind this article is that we need to transform the economics of our energy system by investing in the necessary infrastructure so that it will be affordable for everyone.

…

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

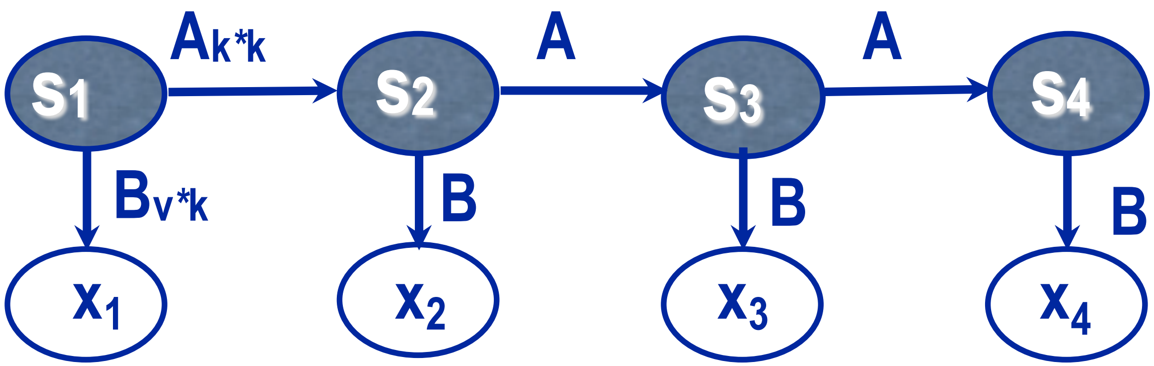
RNN model
Seq2seq
Encoder/decoder
Attention
Masking

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
- ***RNNs are Nonlinear* Hidden state dynamical models**
- **The generalize HMMs or Linear Dynamical Systems**

Standard HMM

- HMM learning problem: Estimate A and B



A = Markov transition matrix
B = emission matrix

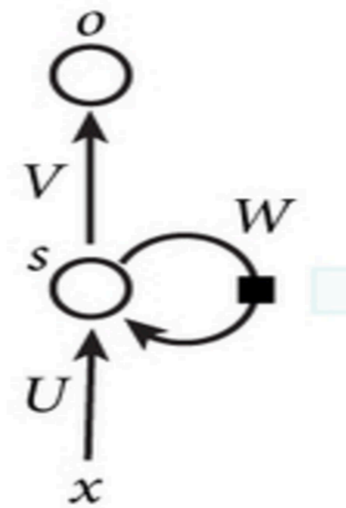
- Estimation done via EM
  - Or spectral methods
- History is forgotten with an exponential decay

Why use hidden states?

- **Markov model on emissions**
  - $p(w_j|w_i)$ is *million x million*
- **Hidden Markov Model**
  - $p(s_j|s_i)$ is *300 x 300* transition matrix
  - $p(w_j|s_i)$ is million * 300 matrix of emeddings

Simple Recurrent Neural Net

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

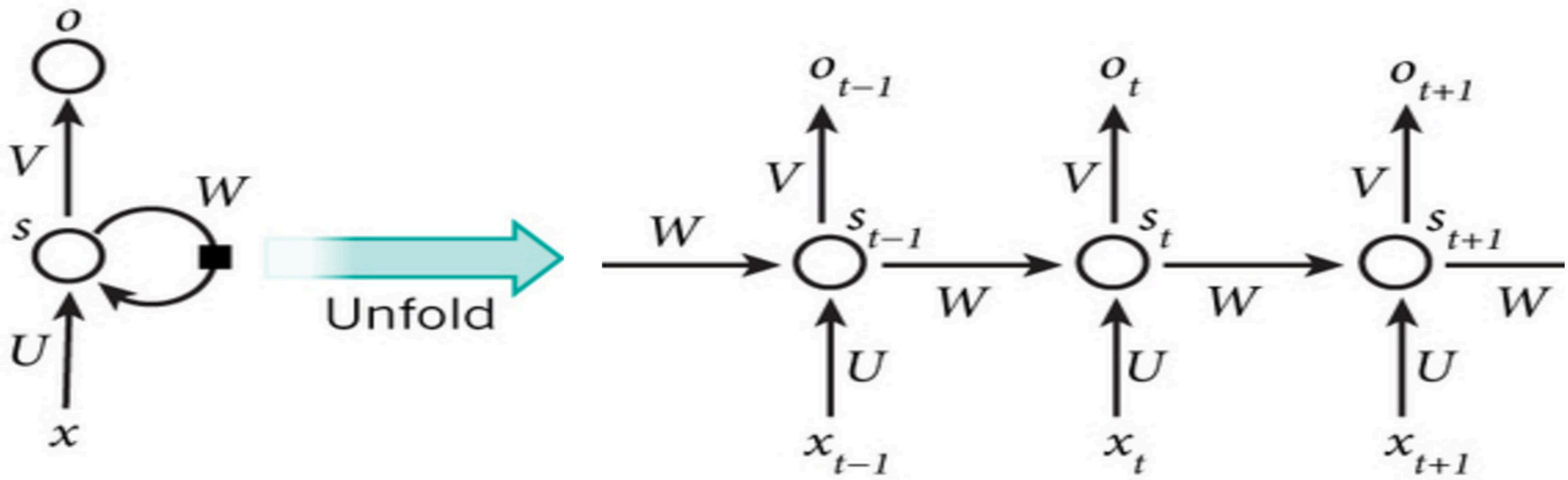


***x_t* = input** (e.g. a word)
***s_t* = hidden state**
***o_t* = output** (e.g. probability of the next word)
y_t = true value (e.g. ***x_{t+1}***)

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

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

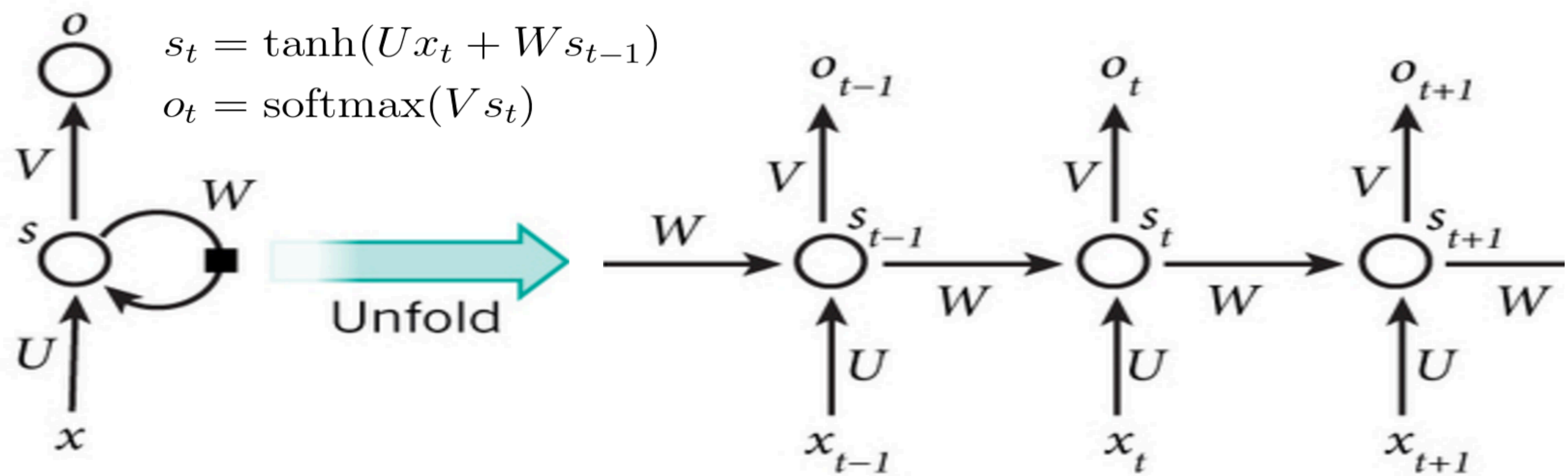
Like HMMs, unroll RNNs in time



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



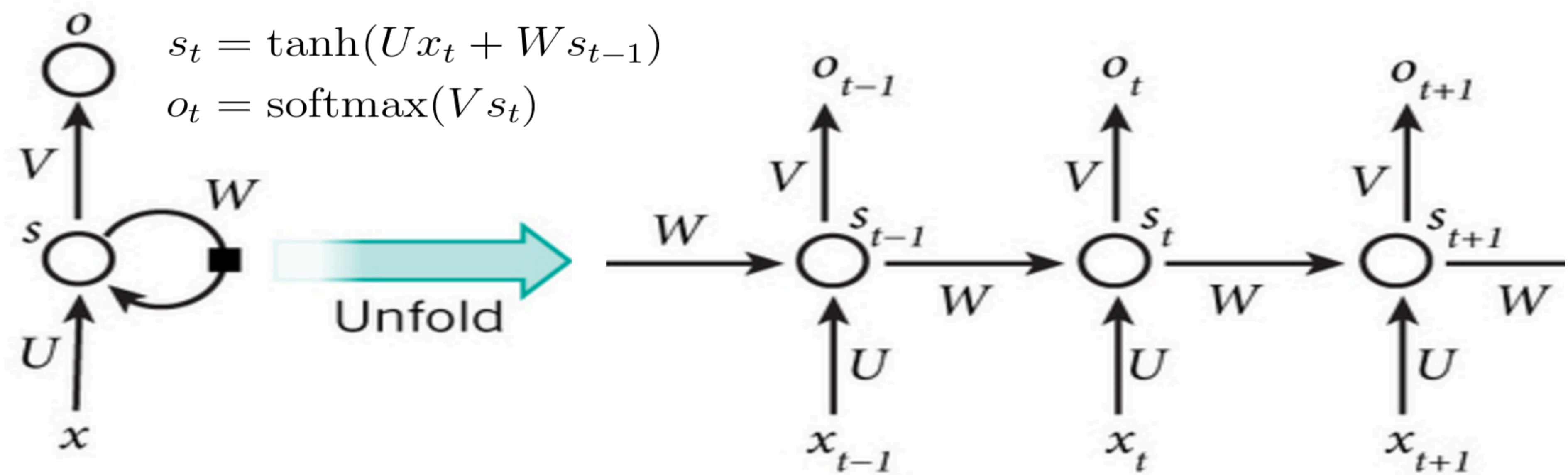
$$s_t = \tanh(Ux_t + Ws_{t-1})$$
$$o_t = \mathrm{softmax}(Vs_t)$$

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



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

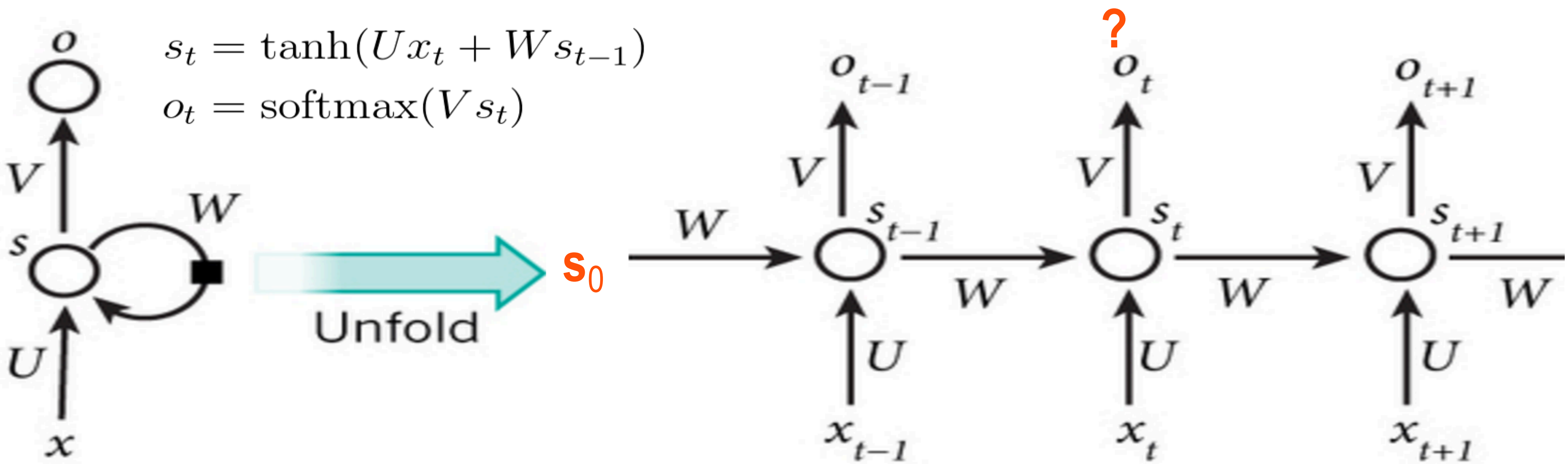
What is the usual loss function?

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

where y_t=i gives the true label

Like HMMs, unroll RNNs in time

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



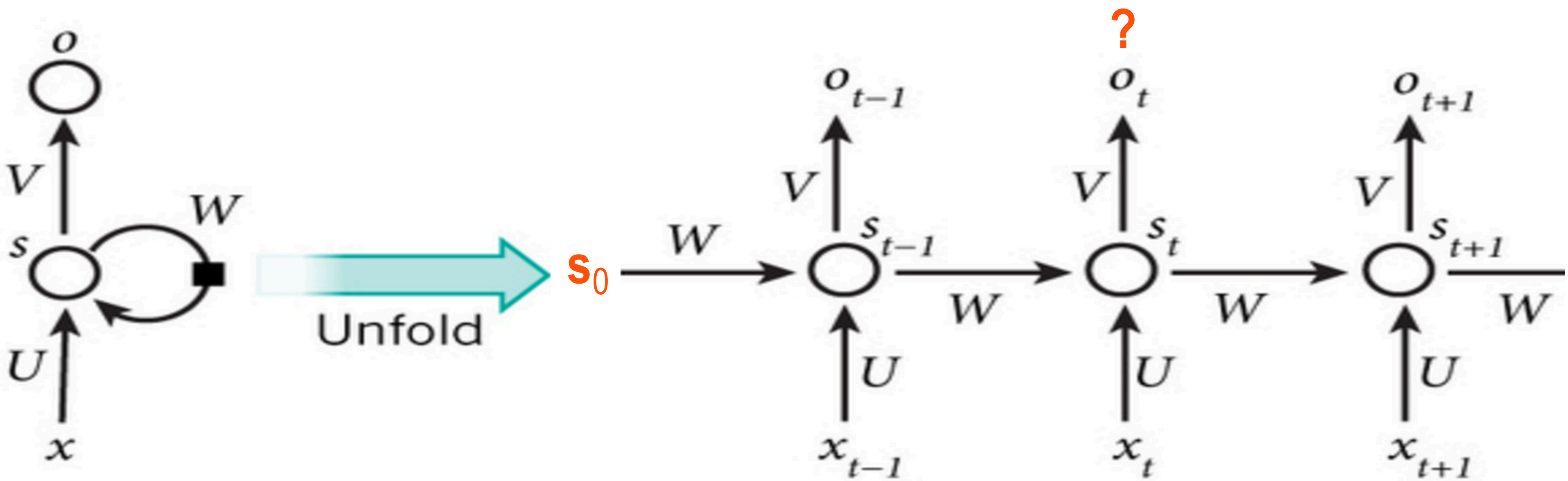
x_t = input - v
s_t = hidden state - k
o_t = output - v

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

$o_t = \mathrm{softmax}(\mathbf{V}s_t) = \mathrm{softmax}(\mathbf{V}\tanh(\mathbf{U}x_t + \mathbf{W}s_{t-1}))$

$= \mathrm{softmax}(\mathbf{V}\tanh(\mathbf{U}x_t + \mathbf{W}\tanh(\mathbf{U}x_{t-1} + \mathbf{W}s_{t-2})))$

RNN gradients



$\mathbf{o}_t$ = *softmax*($\mathbf{V}$ *tanh*($\mathbf{Ux}_t + \mathbf{W}$ *tanh*($\mathbf{Ux}_{t-1} + \mathbf{Ws}_{t-2}$)

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

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

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

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

$$\mathbf{z} = \mathbf{V}\ tanh(\mathbf{Ux}_t + \mathbf{W}\ tanh(\mathbf{Ux}_{t-1} + \mathbf{Ws}_{t-2})$$

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

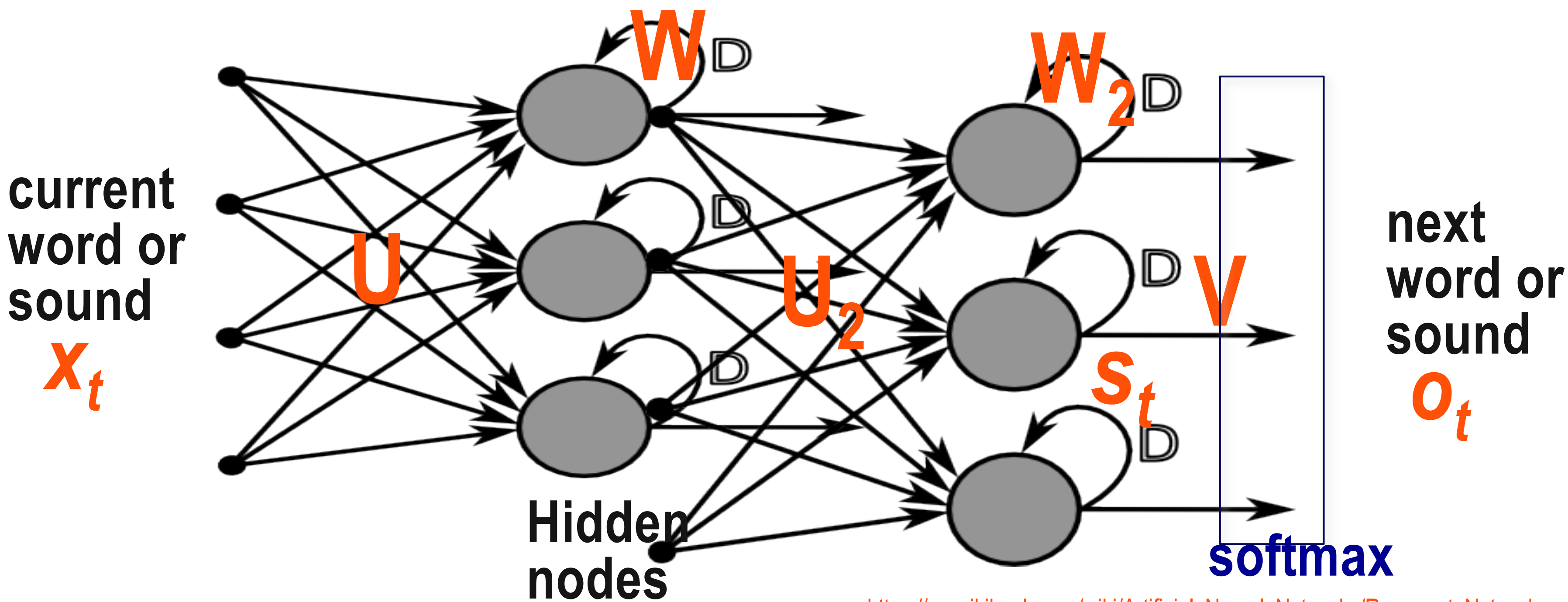
$$= -1/(\Sigma_k e^{z_k})^2\ e^{2z_j} + e^{z_j}/(\Sigma_k e^{z_k}) \quad \text{for k=j}$$

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

Recurrent Neural Nets (RNNs)

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

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 quickly
- Solution: Gated RNN
  - Stores hidden state

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

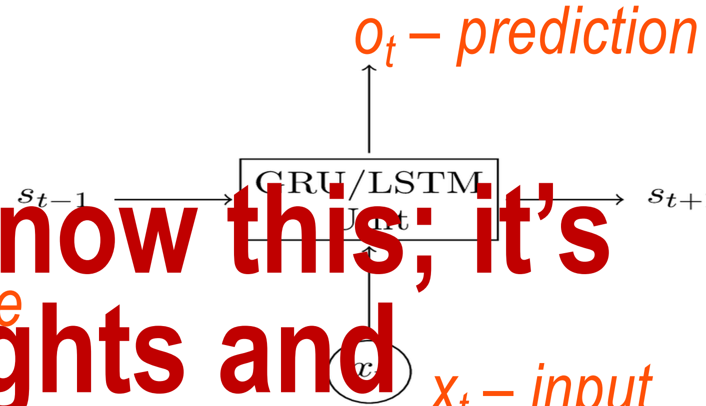
$z = \sigma(U^z x_t + W^z s_{t-1})$ z: update gate

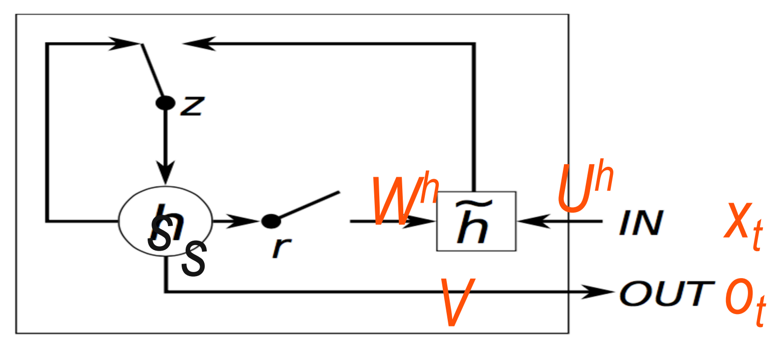
$r = \sigma(U^r x_t + W^r s_{t-1})$ r: reset gate

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

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



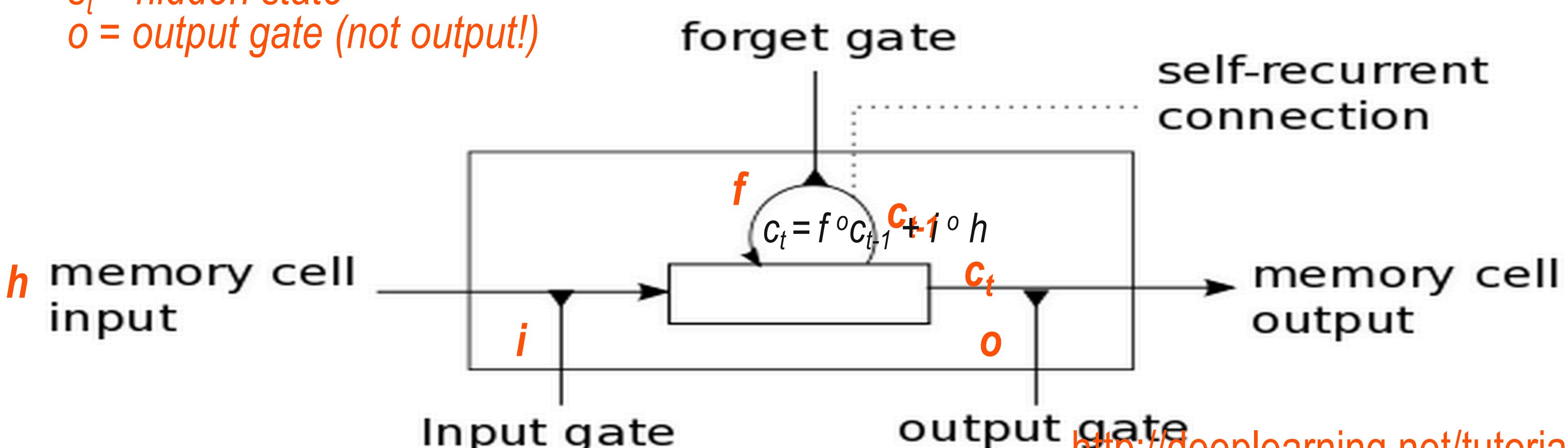


http://deeplearning.net/tutorial/lstm.html

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!)

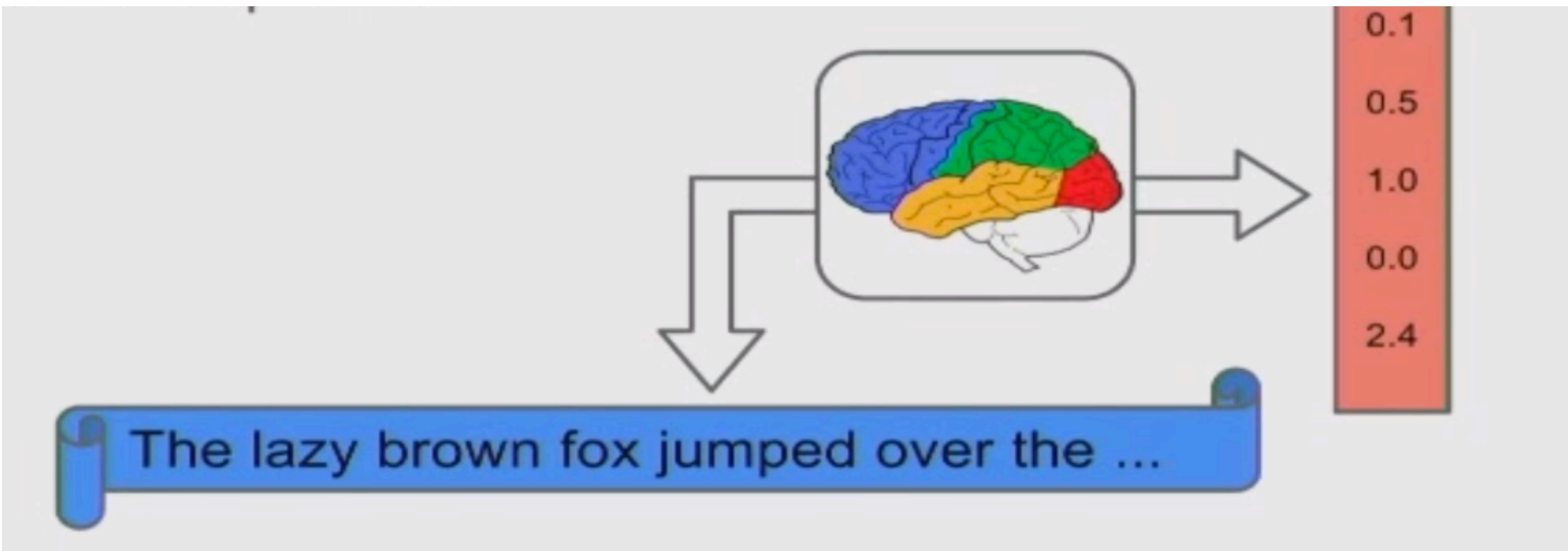


http://deeplearning.net/tutorial/lstm.html

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)

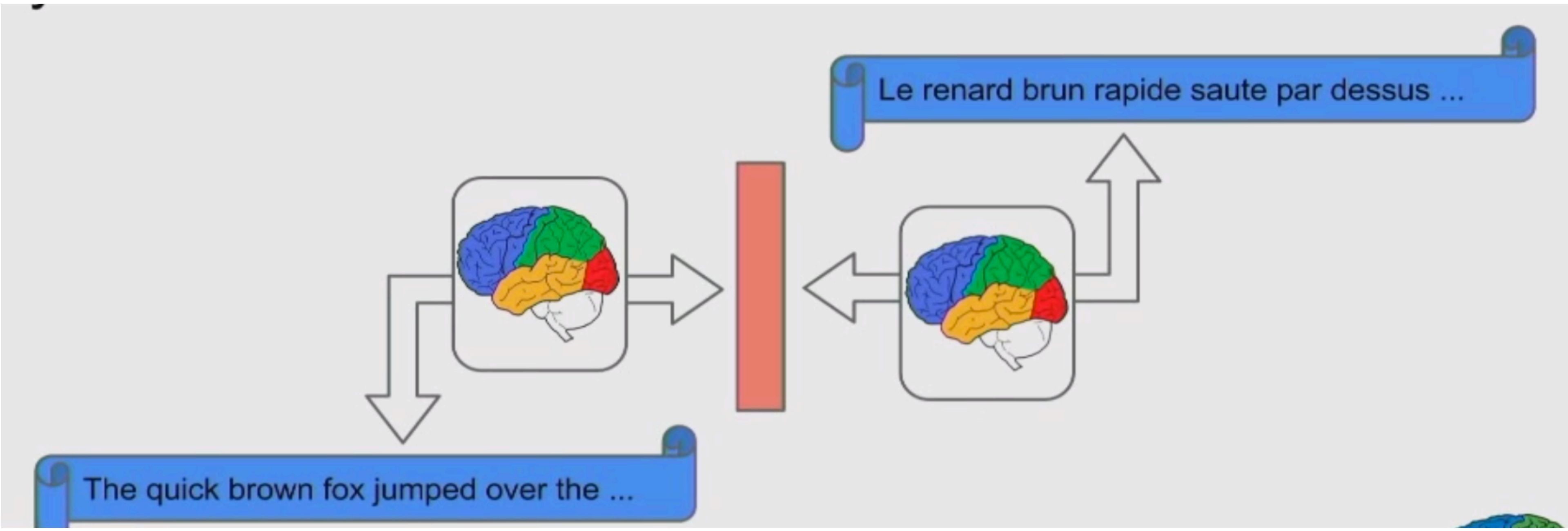
LSTM encodes a sentence



Jeff Dean, google

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

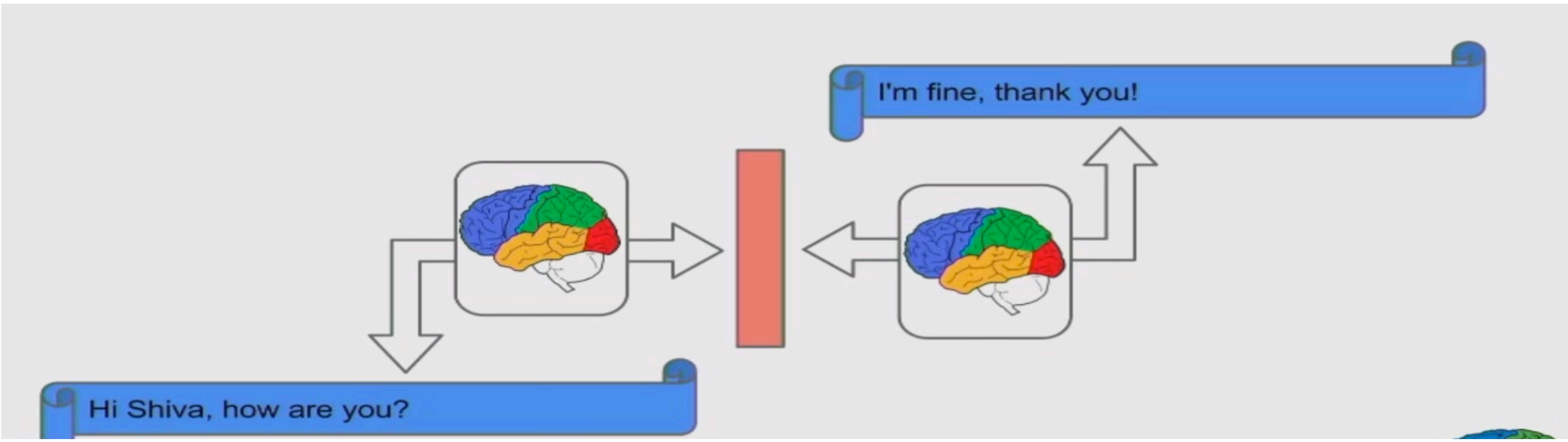
Sequence to sequence (Seq2seq)



Encode and Decode = translate

Jeff Dean, google

Seq2seq chatbot



Encode and Decode = chatbot

Jeff Dean, google

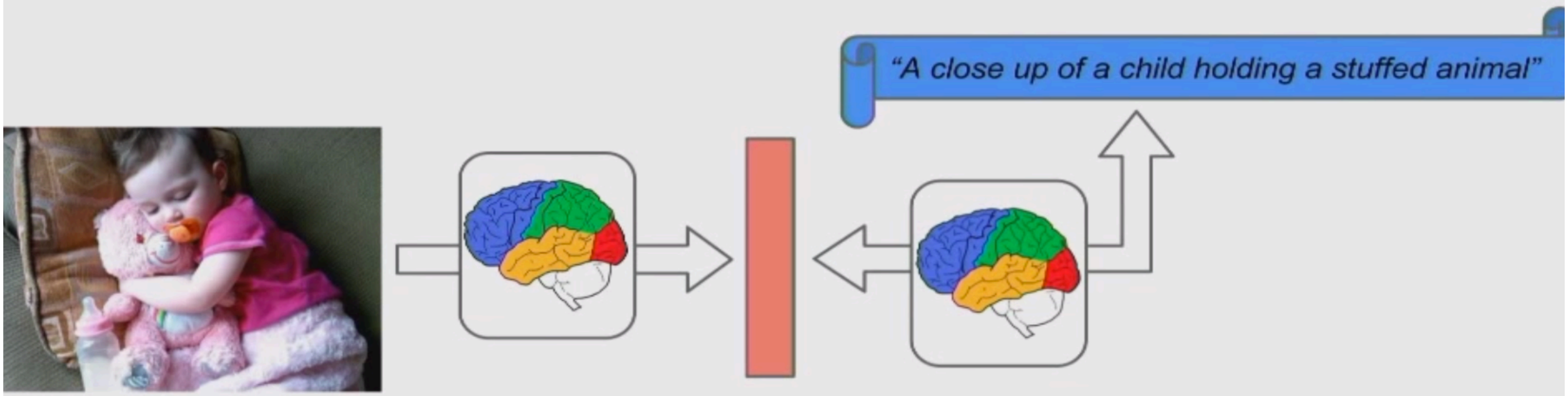
It works (sort of)

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

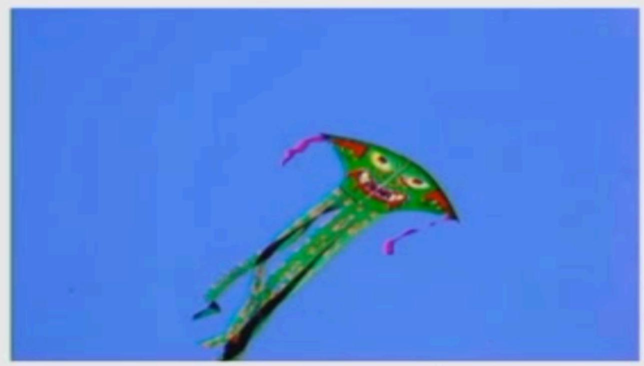


It works (sort of)

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



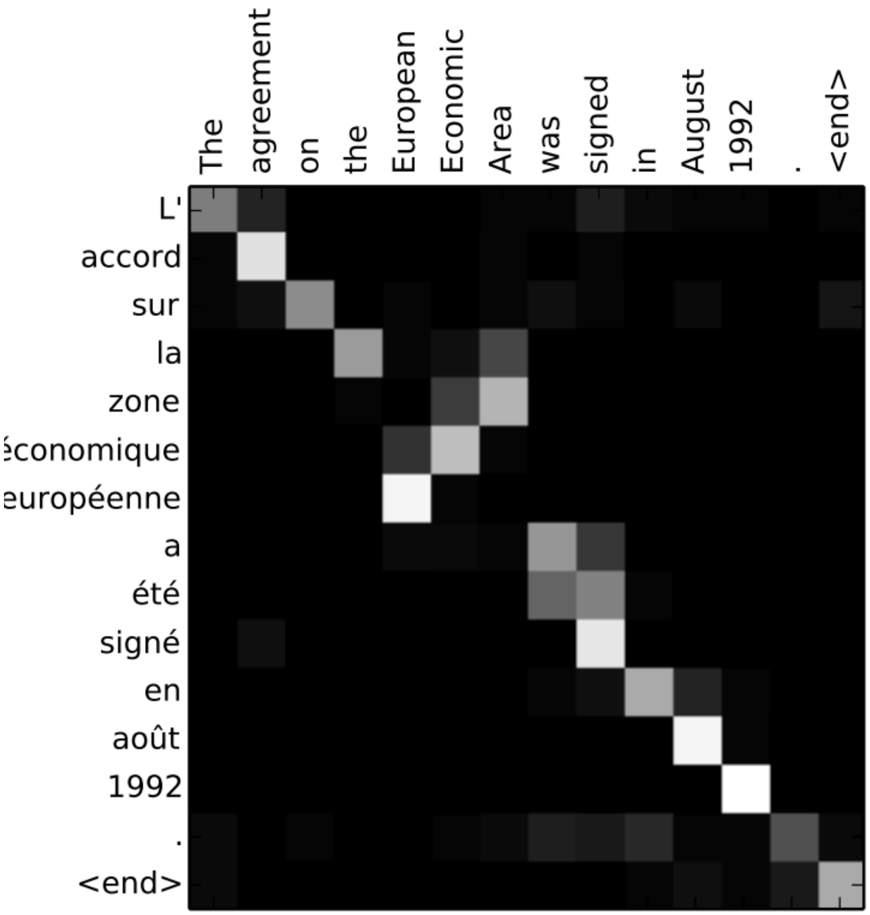
A man flying through the air while riding a snowboard

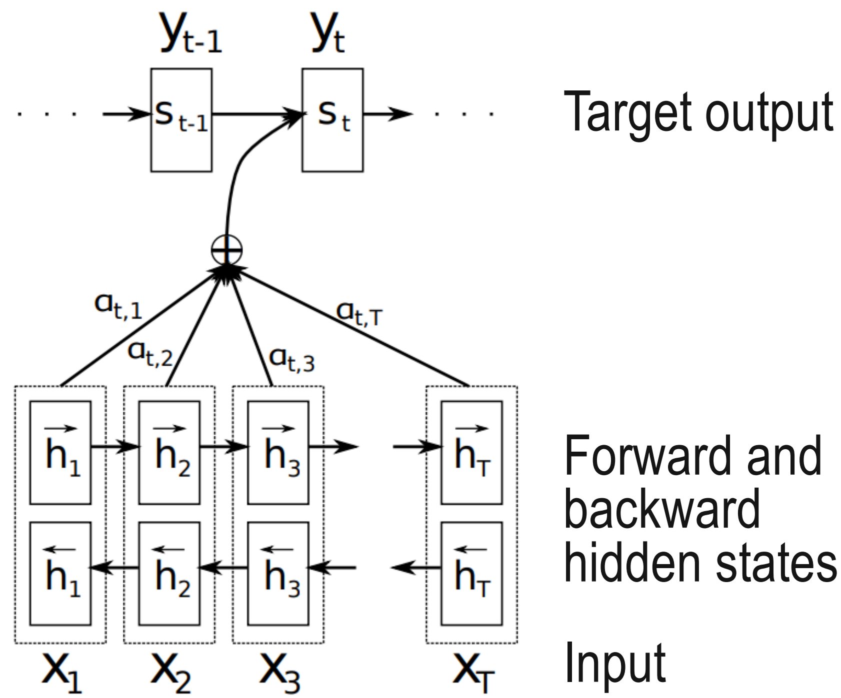
Jeff Dean, google

Language inputs to RNNs

- **Words (“one-hot”)**
- **Characters (“one-hot”)**
- **Bytecodes (“one-hot”)**
- **Word embeddings**
  - Typically 300 dimensional for context-independent (Word2vec)
  - Much bigger for context-sensitive (BERT)

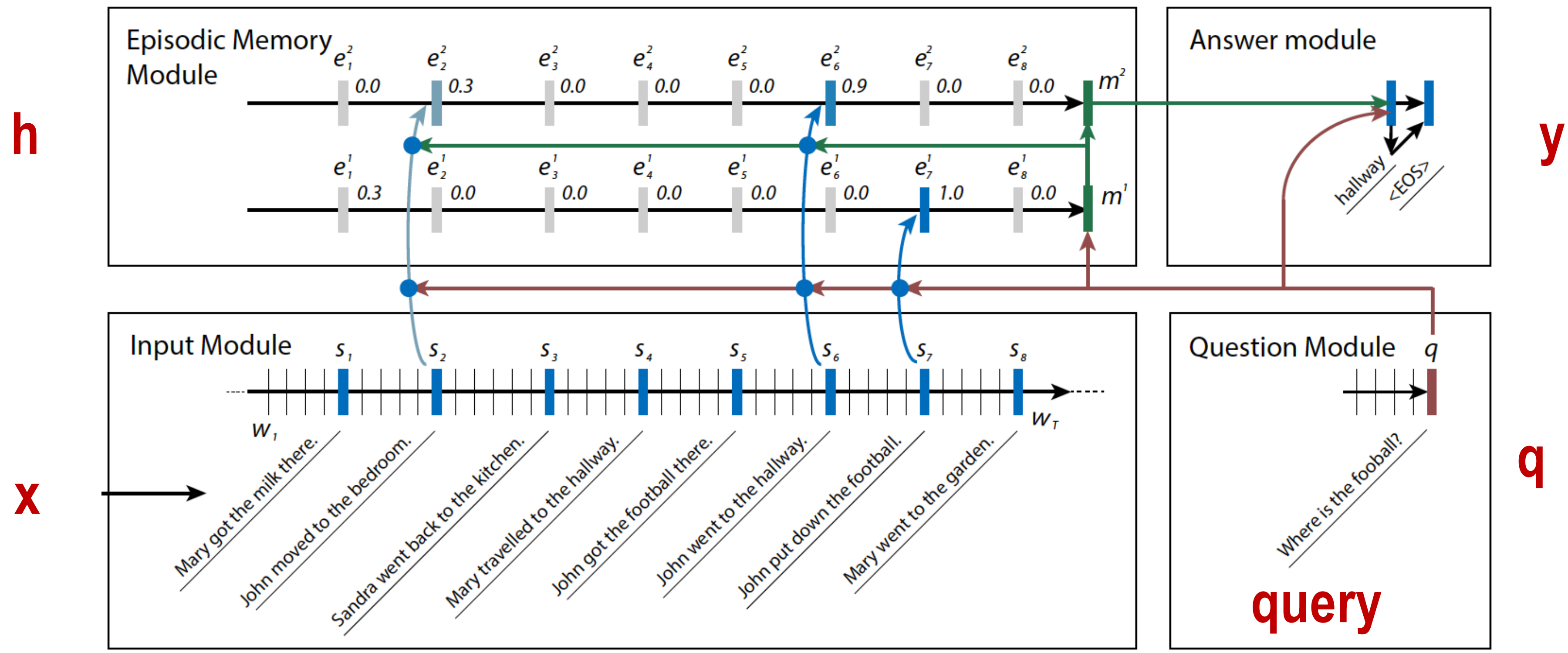
Attention-based Machine Translation





Neural machine translation by jointly learning to align and translate 2015

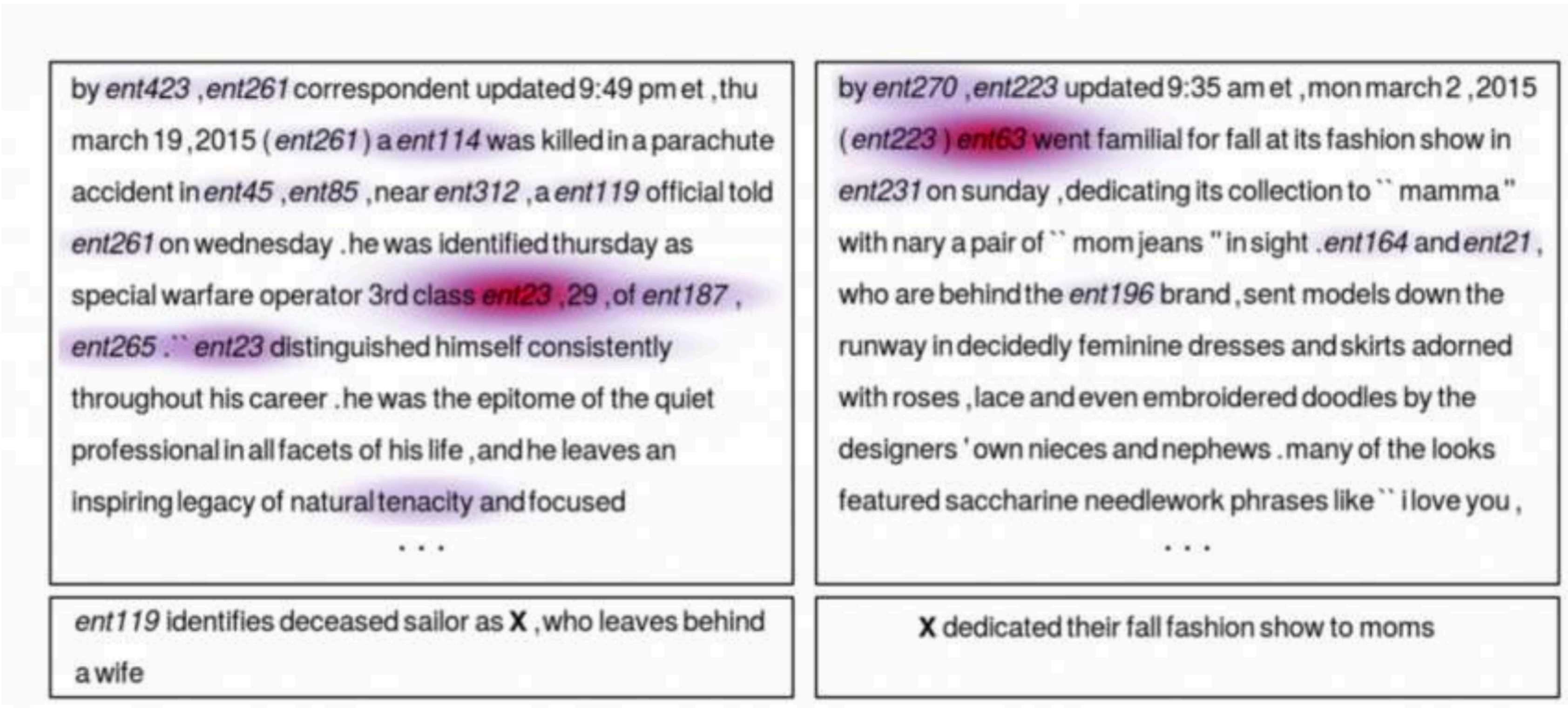
Attention-based Q&A



Find similarity between query and input, **x**

Ask Me Anything: Dynamic Memory Networks for Natural Language Processing

Attention-based Q&A



Teaching Machines to Read and Comprehend 2015

Attention-based Q&A

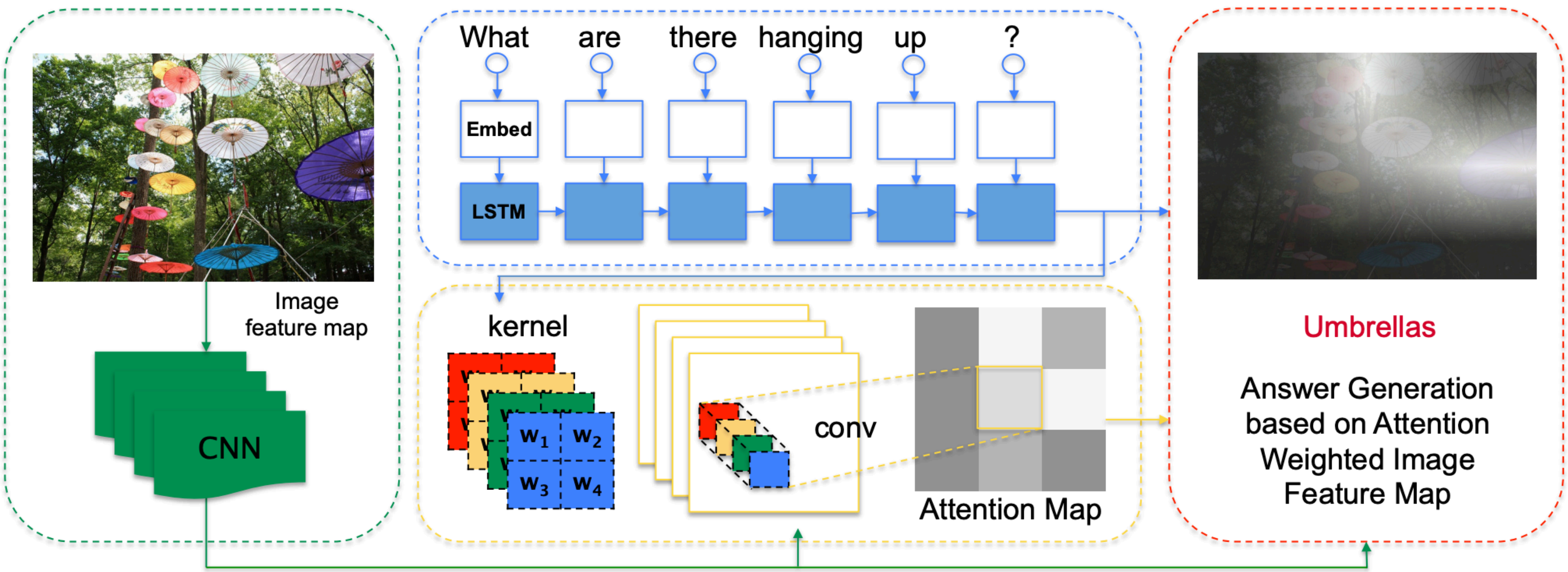
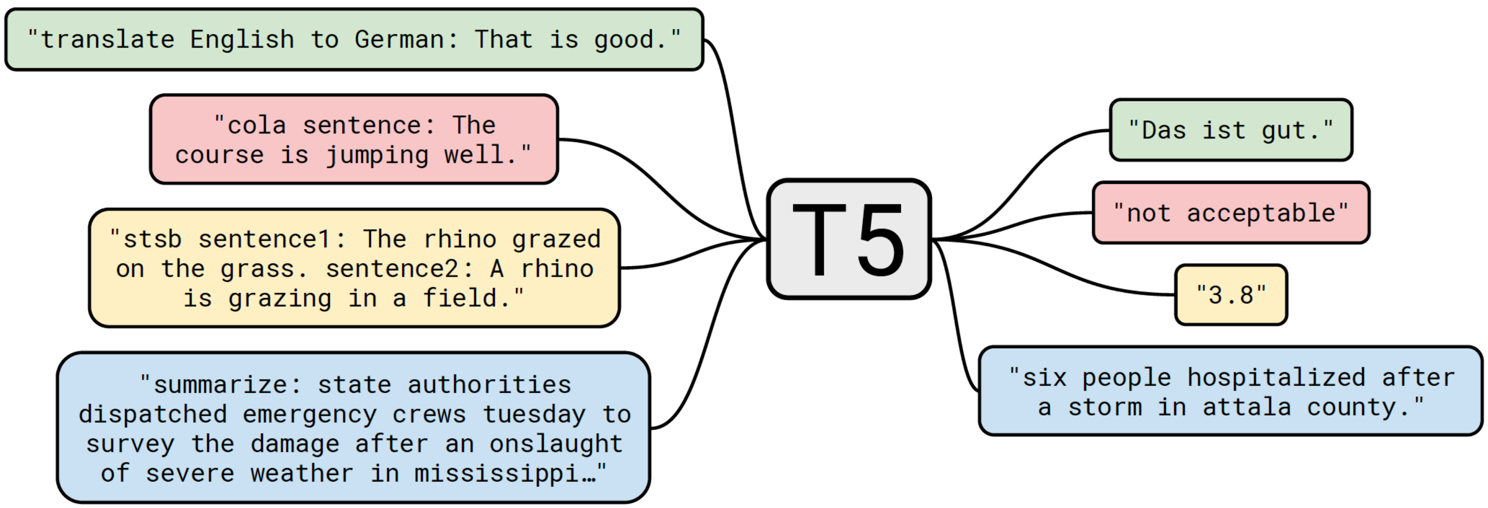


Figure 2. The framework of ABC-CNN. The green box denotes the image feature extraction part using CNN; the blue box is the question understanding part using LSTM; the yellow box illustrates the attention extraction part with configurable convolution; the red box is the answer generation part using multi-class 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

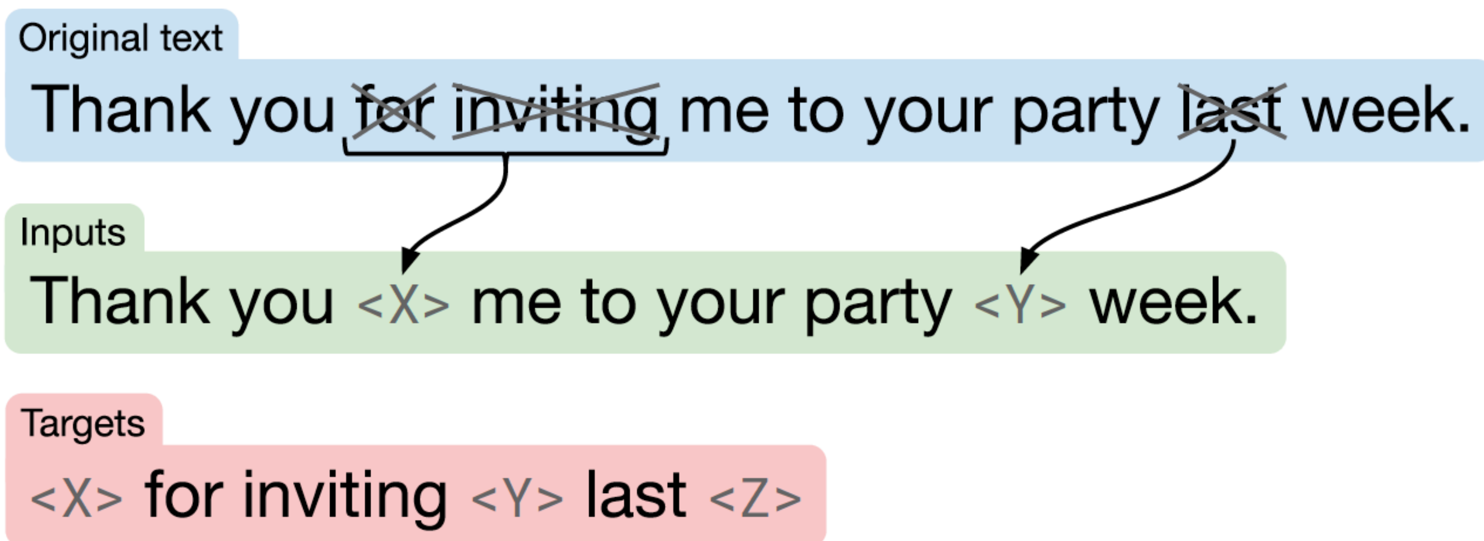
Transformer – seq2seq extension



Uses "self attention"

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

Transformer trained using masking



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

GTP-3

- **Autoregressive Transformer model**
  - **Input:** part of a sentence — **Output:** the next word
  - **Input**: … Time flies like ___ — **Output:** an
  - **Input:** … flies like an ___ — **Output:** arrow

- **Use it**
  - **Input**: part of a story — **Output:** the rest of the story

2020

GPT-3 is big

- **175 billion adjustable parameters**
- **Trained on 500 billion words of text**
  - 410 billion tokens from Common Crawl
  - 67 billion tokens from Books
  - 19 billion tokens from WebText2
  - 3 billion tokens from Wikipedia
- **Training cost $4.6M**

http://jalammar.github.io/how-gpt3-works-visualizations-animations/

GPT-3 can finish a story

***Human**: I am feeling bored today. Grandma, tell me a story about the time the Cat stole the sun.*

***AI**: Okay, Let’s begin. First, you must know that the sun is actually a cat. Also, you must know that the sun is actually not a cat. But it’s also not a piece of cheese, or an orange, or even a gold ring. …*

https://analyticsindiamag.com/openai-gpt-3-model-machine-learning-products/

GPT-3 can finish a poem

Q: Shall I compare thee to a summer’s day?
Thou art more lovely and more temperate:
Rough winds do shake the darling buds of May,
And summer’s lease hath all too short a date;
A: winter’s day, when beams of sun are few,
And skies are grey with clouds intemperate,
Is to thy beauty somehow sordid grown.
Her lovely-lovely eye is quite shut up,

https://www.gwern.net/GPT-3

GPT-3 can translate legalese

Legal Clause: Sale of all or substantially all of the assets of the Company or a sale of shares involving a change in control (each, a "Corporate Transaction") will be treated in the same way as a liquidation and the proceeds of sale will be distributed as set out in paragraph 3. If the holders of Series A Shares have received any Special Dividend it shall be set off against their Liquidation Preference.

Plain English: If the company is sold, or a new owner takes control, the proceeds of the sale will be distributed as in the liquidation clause above. Any special dividend paid will be treated as an initial payment towards the Series A investors.

GPT-3 makes really dumb mistakes

Q: Who was president of the United States in 1600?
A: Queen Elizabeth I was president of the United States in 1600.

It is great at completing patterns based on the 500 billion words it has read, but it isn't able to do reasoning, or to answer questions about things it hasn't seen. It writes great prose, and makes dumb mistakes

https://mindmatters.ai/2020/07/gpt-3-is-mindblowing-if-you-dont-question-it-too-closely/

GPT memorizes a lot

- **Q. What do you call a droid that takes the long way around?**
  - A. R2 detour.
- **Q. Why was 6 afraid of 7?**
  - A. Because 7 8 9!
- **Q. What did the Dalmatian say after lunch?**
  - A. That hit the spot!

https://www.gwern.net/GPT-3#dad-jokes

GPT is less impressive otherwise

- **Q. When does a joke become a "dad" joke?**
  - A. When it's not funny.
- **Q. How do you stop an astronaut's baby from crying?**
  - A. You go into space!
- **Q. How does a vampire start a letter?**
  - A. Dear blood.

https://www.gwern.net/GPT-3#dad-jokes

What you should know

- **Recurrent Neural Nets (LSTMs) generalize HMMs, Kalman filters**
- **They have replaced HMMs for speech-to-text and machine translation**
  - Now used (as “Transformers”) for natural language *generation*
- **Lots of black magic “engineering”**
  - Number and size of layers, regularization
  - Forms of gating (LSTM …), attention, masking
  - Gradient descent can be tricky
  - “Fine tune” a model that was trained on an enormous data set