ELEC 677:
Gated Recurrent Neural Network Architectures & Recurrent Neural Network Language Models
Lecture 8

Ankit B. Patel, CJ Barberan
Baylor College of Medicine (Neuroscience Dept.)
Rice University (ECE Dept.)
11-1-2016

Latest News

Yoshua Bengio launches Element AI

- Deep learning incubator in Montreal

Comma.ai stops development of self-driving car device

- US federal highway safety officials said the device was not safe to use
  - Connectivity issues
  - Needing a person if there were connectivity issues

Microsoft releases beta of Microsoft Cognitive Toolkit

- Another deep learning framework, like Tensorflow, Theano, Torch
  - Python or C++

RNN Training

Why Training is Unstable

$$x^{(l)} = W^{(l-1)} y^{(l-1)} + b^{(l-1)}$$

$$y^{(l)} = f(x^{(l)})$$

Let the activation function $f(x) = \alpha x + \beta$,

$$Var\left(y^{(l)}\right) = \alpha^2 n_{l-1} \sigma_{l-1}^2 \left(Var\left(y^{(l-1)}\right) + \beta^2 I_{n_l}\right).$$

$$Var\left(\frac{\partial cost}{\partial y^{(l-1)}}\right) = \alpha^2 n_l \sigma_{l-1}^2 Var\left(\frac{\partial cost}{\partial y^{(l)}}\right).$$

Variance of activations/gradients grows multiplicatively

[Xu, Huang, Li]

Interesting Question

- Are there modifications to an RNN such that it can combat these gradient problems?

RNNs with Longer Term Memory

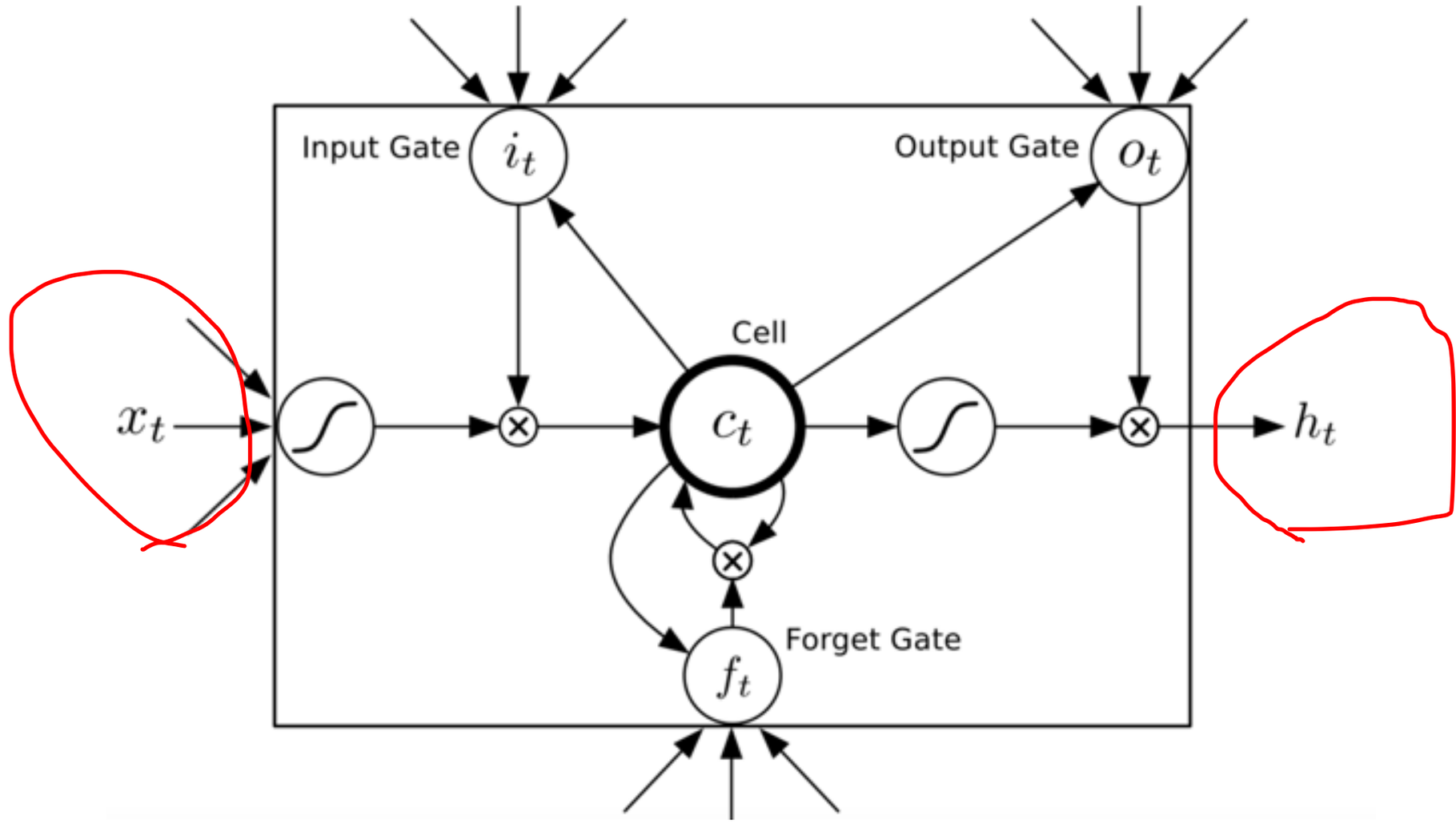
Motivation

- The need to remember certain events for arbitrarily long periods of time (Non-Markovian)

- The need to forget certain events

Long Short Term Memory

- 3 gates
- Input
- Forget
- Output



[Zygmunt Z.]

LSTM Formulation

seen as logistic regression

除非你确定左右对称, 50-50, 不然用bias

$$i_t = \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i\right)$$

$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f\right)$$

maybe input is noise

$$c_t = f_t c_{t-1} + i_t \tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right)$$

-1,1, center = 0, otherwise, ct -> inf

$$o_t = \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o\right)$$

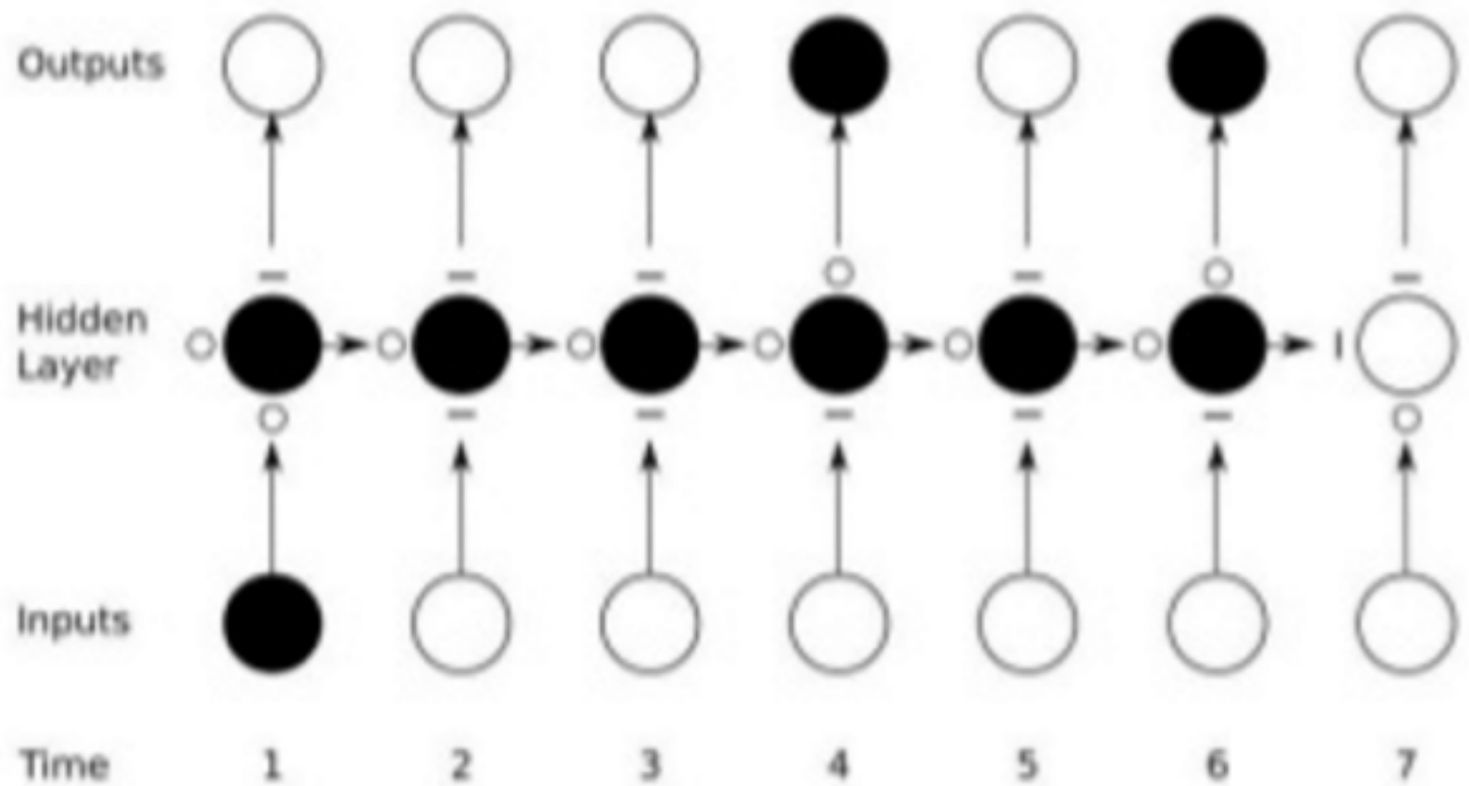
like close your mouth

$$h_t = o_t \tanh(c_t)$$

$$y_t = W_{ho}h_t + b_o$$

[Alex Graves, Navdeep Jaitly]

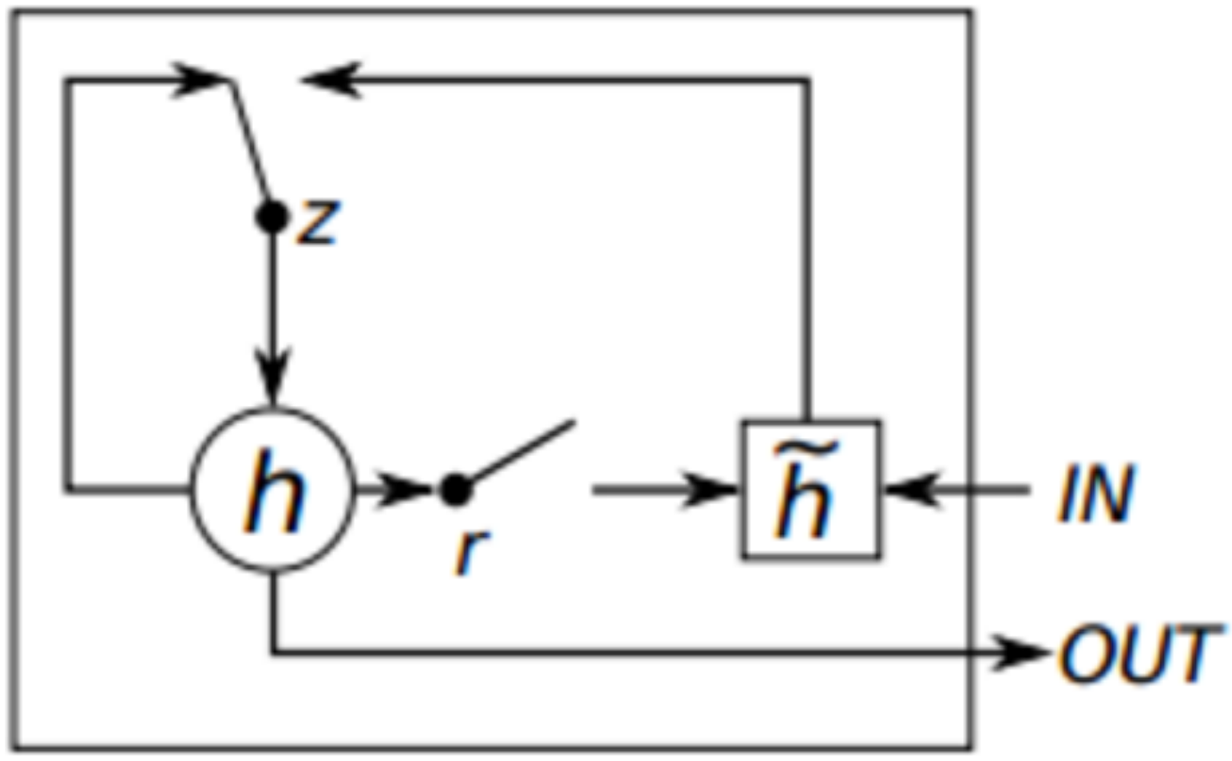
Preserving Gradients



[Hochreiter, Schmidhuber]

Gated Recurrent Unit

- 2 gates
  - Reset
    - Combine new input with previous memory
  - Update
    - How long the previous memory should stay



[Zygmunt Z.]

GRU Formulation

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

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

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

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

exponential moving average

pt^ = (1-z)pt-1^ + z pt

1-z + 1 guarentee sum to 1

[Danny Britz]

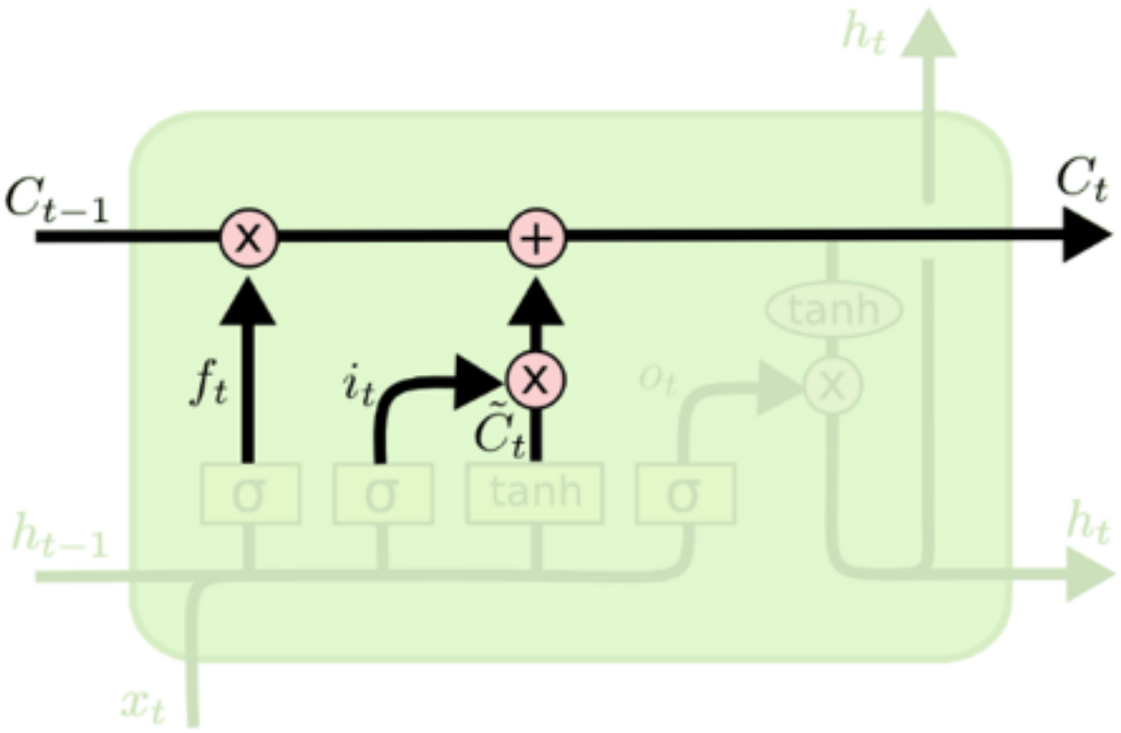
LSTM & GRU Benefits

- Remember for longer temporal durations
  - RNN has issues for remembering longer durations
- Able to have feedback flow at different strengths depending on inputs

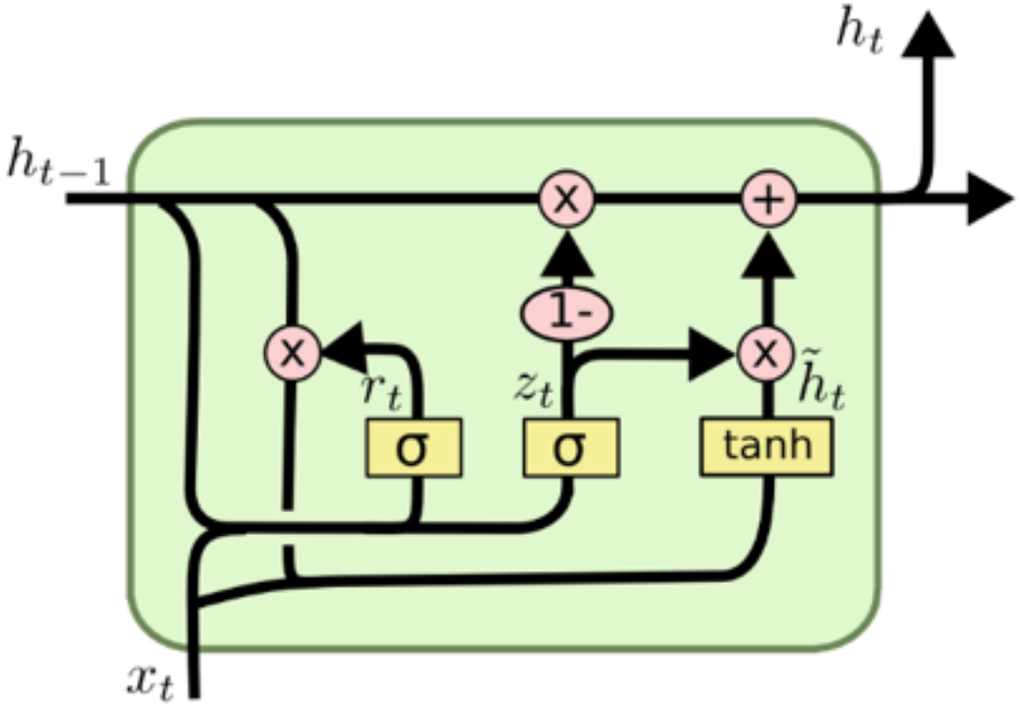
Differences between LSTM & GRU

- GRU has two gates, while LSTM has three gates
- GRU does not have internal memory
- GRU does not use a second nonlinearity for computing the output

Visual Difference of LSTM & GRU



LSTM



GRU

[Chris Olah]

LSTM vs GRU Results

			tanh	GRU	LSTM
Music Datasets	Nottingham	train	3.22	2.79	3.08
		test	**3.13**	3.23	3.20
	JSB Chorales	train	8.82	6.94	8.15
		test	9.10	**8.54**	8.67
	MuseData	train	5.64	5.06	5.18
		test	6.23	**5.99**	6.23
	Piano-midi	train	5.64	4.93	6.49
		test	9.03	**8.82**	9.03
Ubisoft Datasets	Ubisoft dataset A	train	6.29	2.31	1.44
		test	6.44	3.59	**2.70**
	Ubisoft dataset B	train	7.61	0.38	0.80
		test	7.62	**0.88**	1.26

[Chung, Gulcehre, Cho, Bengio]

Other Methods for Stabilizing RNN Training

Why Training is Unstable

$$x^{(l)} = W^{(l-1)} y^{(l-1)} + b^{(l-1)}$$

$$y^{(l)} = f(x^{(l)})$$

Let the activation function $f(x) = \alpha x + \beta$,

$$Var\left(y^{(l)}\right) = \alpha^2 n_{l-1} \sigma_{l-1}^2 \left(Var\left(y^{(l-1)}\right) + \beta^2 I_{n_l}\right).$$

$$Var\left(\frac{\partial cost}{\partial y^{(l-1)}}\right) = \alpha^2 n_l \sigma_{l-1}^2 Var\left(\frac{\partial cost}{\partial y^{(l)}}\right).$$

Variance of activations/gradients grows multiplicatively

[Xu, Huang, Li]

Stabilizing Activations & Gradients

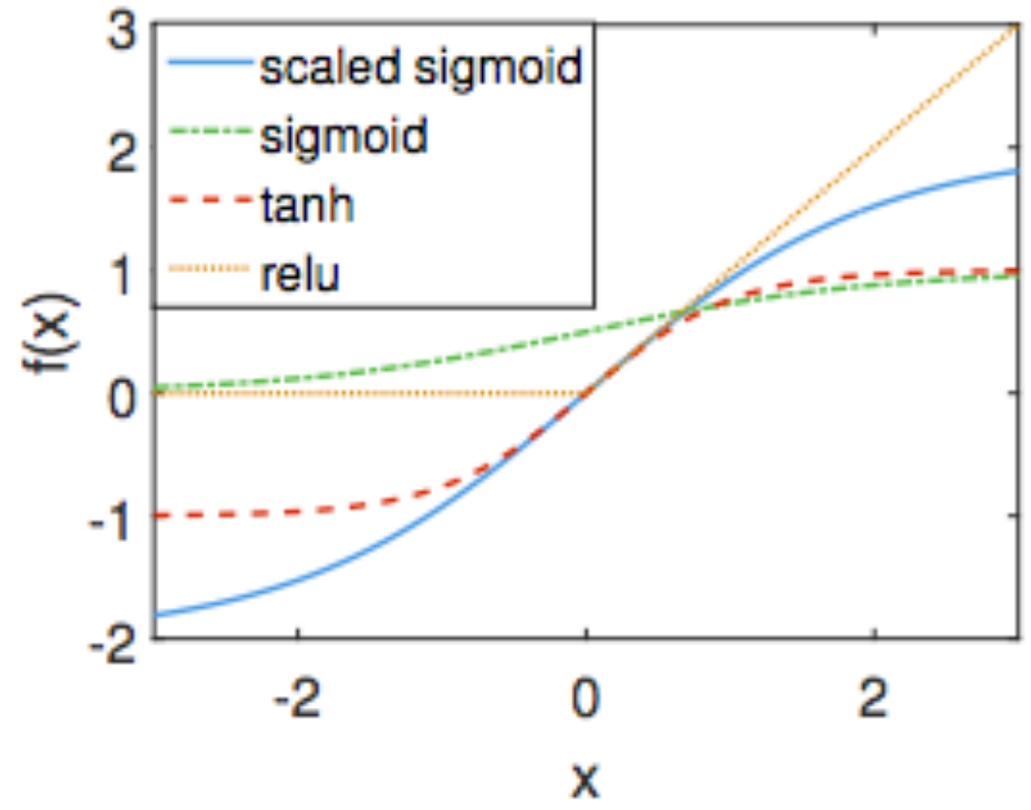
$$\mathbf{Var}\left(y^{(l)}\right) = \mathbf{Var}\left(y^{(l-1)}\right) \quad \textbf{and} \quad \mathbf{Var}\left(\frac{\partial \text{cost}}{\partial y^{(l)}}\right) = \mathbf{Var}\left(\frac{\partial \text{cost}}{\partial y^{(l-1)}}\right);$$

$$n_l \sigma_{l-1}^2 \approxeq 1 \quad \textbf{and} \quad n_{l-1} \sigma_{l-1}^2 \approxeq 1;$$

We want $\alpha = 1$ **and** $\beta = 0.$

[Xu, Huang, Li]

Taylor Expansions of Different Activation Functions



$$\text{sigmoid}(x) = \frac{1}{2} + \frac{x}{4} - \frac{x^3}{48} + O(x^5)$$

$$\tanh(x) = 0 + x - \frac{x^3}{3} + O(x^5)$$

$$\text{relu}(x) = 0 + x \quad \text{for } x \geq 0.$$

why sigmoid is difficult to be stable?
- not 0 center.

[Xu, Huang, Li]

Layer Normalization

- Similar to batch normalization

why not batch normalization in RNN?
- inputs may be sentences with different length, thus can't be scaled.
https://arxiv.org/pdf/1607.06450.pdf

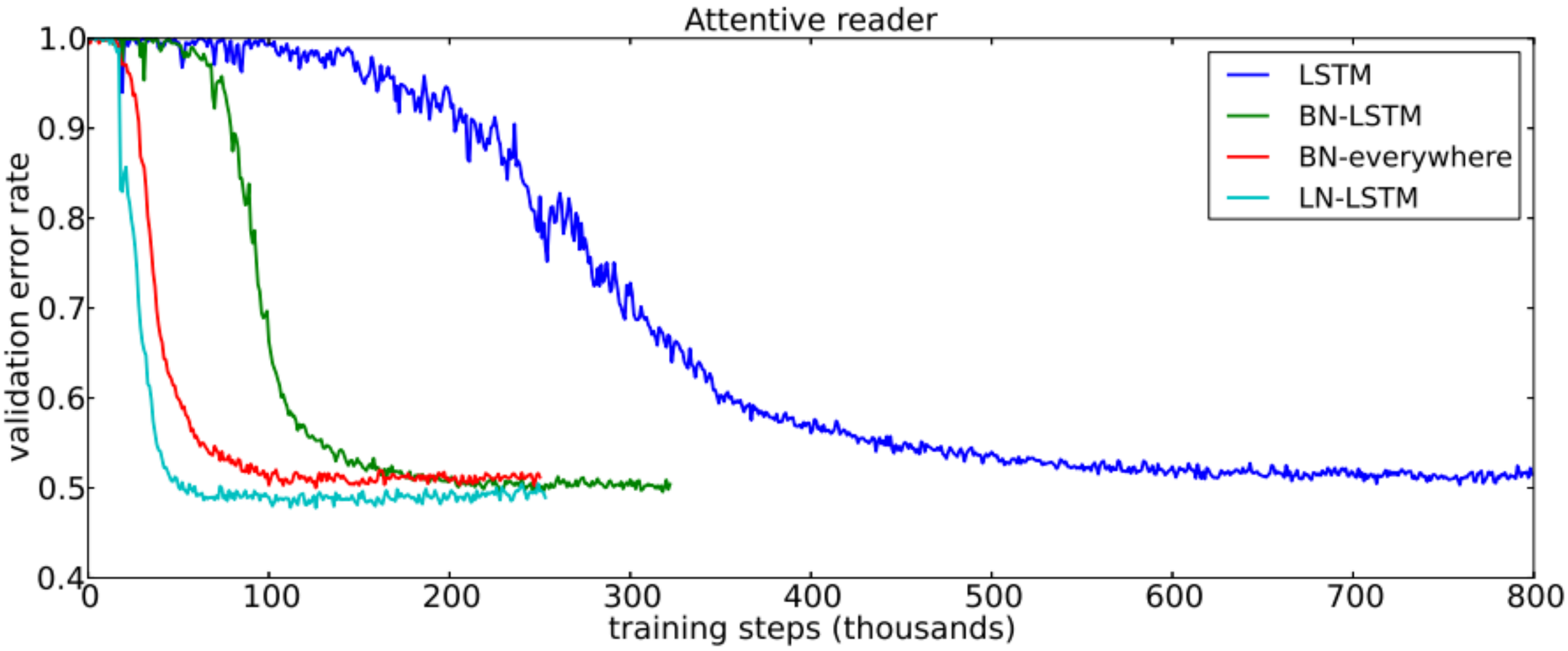
  - Apply it to RNNs to stabilize the hidden state dynamics

$$\mathbf{h}^t = f\left[\frac{\mathbf{g}}{\sigma^t} \odot (\mathbf{a}^t - \mu^t) + \mathbf{b}\right] \qquad \mu^t = \frac{1}{H}\sum_{i=1}^{H} a_i^t \qquad \sigma^t = \sqrt{\frac{1}{H}\sum_{i=1}^{H}(a_i^t - \mu^t)^2}$$

H-1 to make unbiased, but does not mater here for H super large.

[Ba, Kiros, Hinton]

Layer Normalization Results



[Ba, Kiros, Hinton]

Streamlining Normalization

Algorithm 5 Streaming Normalization Layer: Forward

Require: layer input $\mathbf{x}$ (a mini-batch), NormOP N(.,.), function S(.) to compute NormStats for every element of $\mathbf{x}$, running estimates of NormStats and/or related information packed in a structure/table H_1, function F to update H_1 and generate current estimate of NormStats $\hat{\mathrm{s}}$.

Ensure: layer output $\mathbf{y}$ (a mini-batch) and H_1 (it is stored in this layer, instead of feeding to other layers), always maintain the latest $\hat{\mathrm{s}}$ in case of testing

if *training* **then**

$\mathrm{s} = \mathrm{S}(\mathbf{x})$

$\{H_1, \hat{\mathrm{s}}\} = \mathrm{F}(H_1, \mathrm{s})$

$\mathrm{y} = \mathrm{N}(\mathbf{x}, \hat{\mathrm{s}})$

else {*testing*}

$\mathrm{y} = \mathrm{N}(\mathbf{x}, \hat{\mathrm{s}})$

end if

[Liao, Kawaguchi, Poggio]

Streamlining Normalization

Algorithm 6 Streaming Normalization Layer: Backpropagation

Require: $\frac{\partial E}{\partial y}$ (a mini-batch) where E is objective, layer input $\mathbf{x}$ (a mini-batch), NormOP N(.,.), function S(.) to compute NormStats for every element of $\mathbf{x}$, running estimates of NormStats $\hat{\mathrm{s}}$, running estimates of gradients and/or related information packed in a structure/table H_2, function G to update H_2 and generate the current estimates of gradients of NormStats $\frac{\partial E}{\partial \hat{s}}$.

Ensure: $\frac{\partial E}{\partial x}$ (a mini-batch) and H_2 (it is stored in this layer, instead of feeding to other layers)

$\frac{\partial E}{\partial \hat{s}}$ is calculated using chain rule.

$\{H_2, \widehat{\frac{\partial E}{\partial \hat{s}}}\} = \mathrm{G}(H_2, \frac{\partial E}{\partial \hat{s}})$

Use $\widehat{\frac{\partial E}{\partial \hat{s}}}$ for further backpropagation, instead of $\frac{\partial E}{\partial \hat{s}}$

$\frac{\partial E}{\partial x}$ is calculated using chain rule.

[Liao, Kawaguchi, Poggio]

Streamlining Normalization for RNNs

Normalized RNN

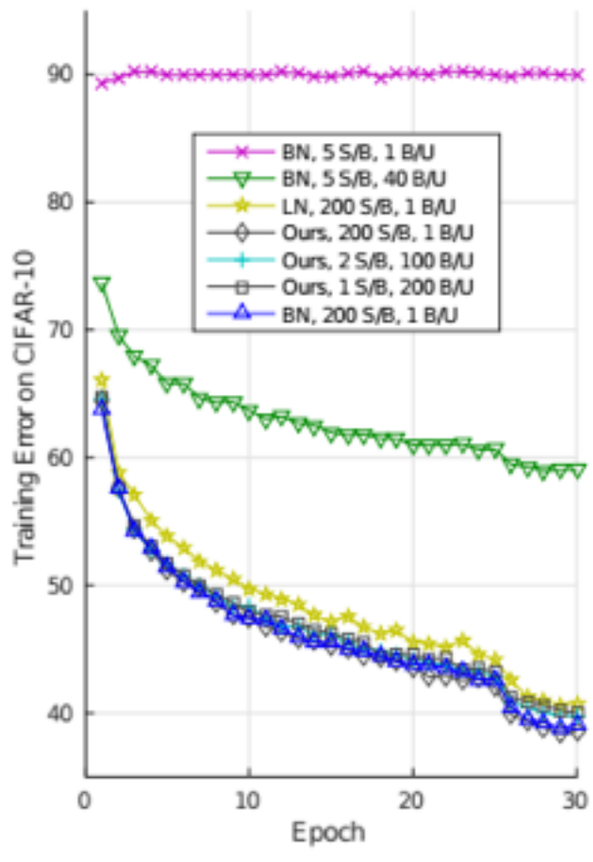
$$h_t = NonLinear(Norm(W_x * x_t) + Norm(W_h * h_{t-1}))$$

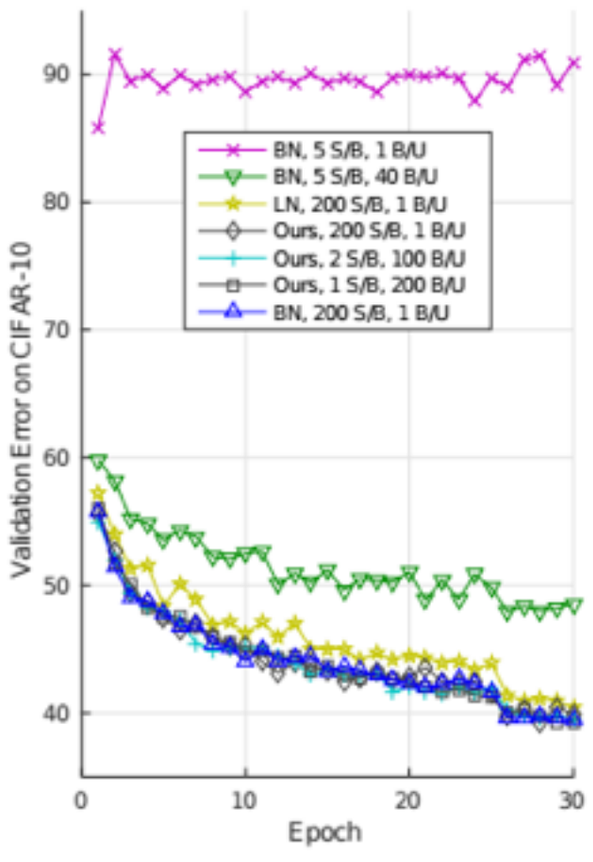
Normalized GRU

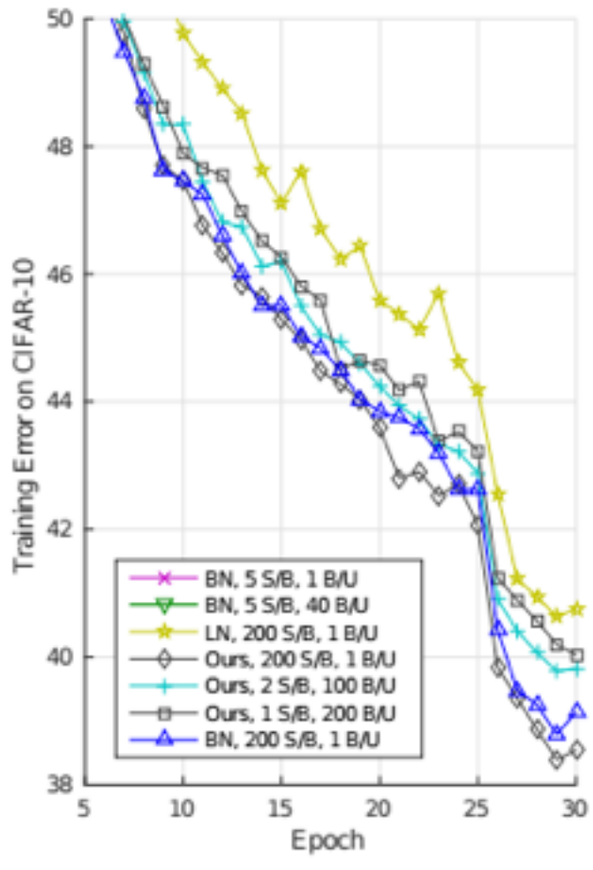
$$g_r = Sigmoid(Norm(W_{xr} * x_t) + Norm(W_{hr} * h_{t-1}))$$
$$g_z = Sigmoid(Norm(W_{xz} * x_t) + Norm(W_{hz} * h_{t-1}))$$
$$h_{new} = NonLinear(Norm(W_{xh} * x_t) + Norm(W_{hh} * (h_{t-1} \odot g_r)))$$
$$h_t = g_z \odot h_{new} + (1 - g_z) \odot h_{t-1}$$

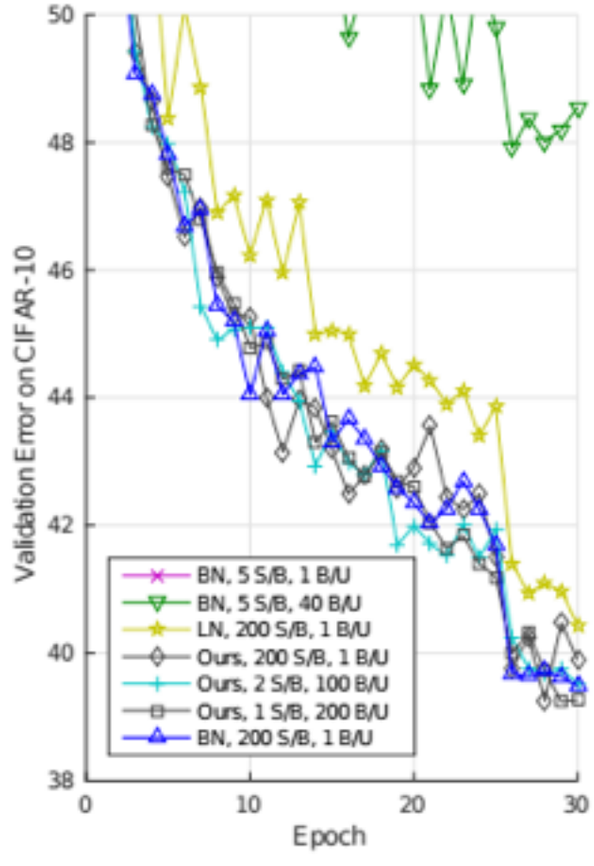
[Liao, Kawaguchi, Poggio]

Streamlining Normalization Results



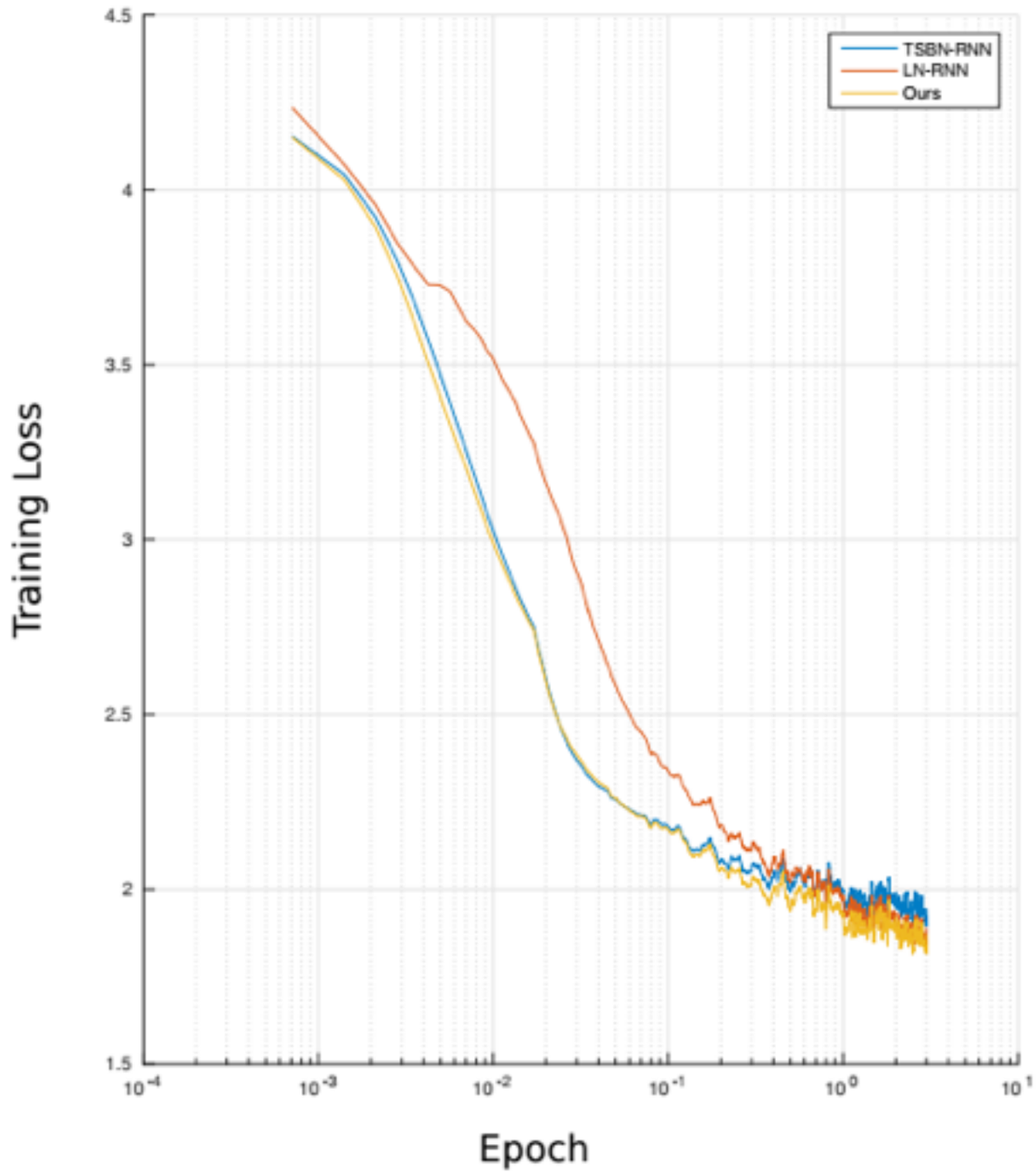


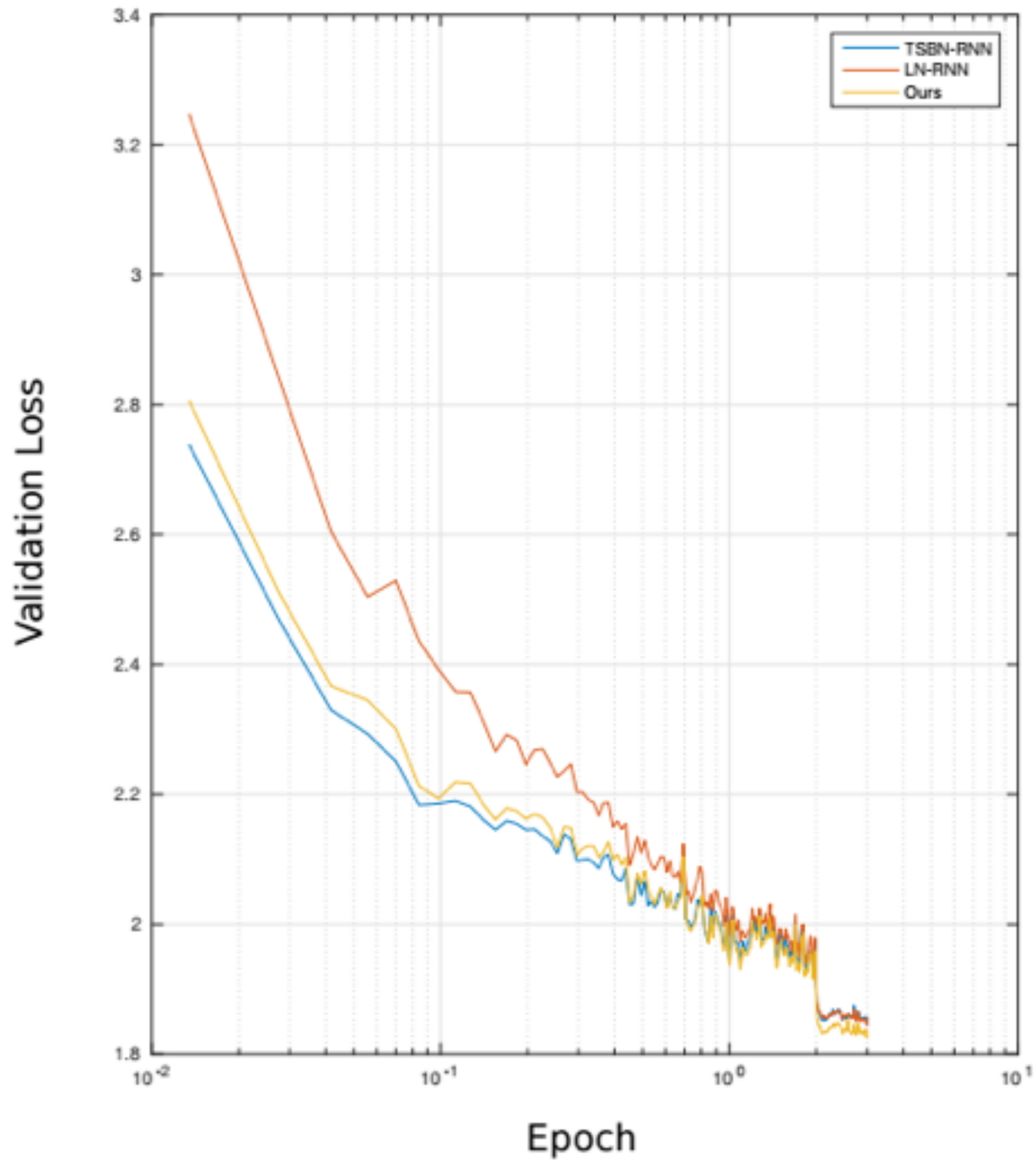




[Liao, Kawaguchi, Poggio]

Streamlining Normalization RNN Results





[Liao, Kawaguchi, Poggio]

Normalization Techniques

Approach	FF & FC	FF & Conv	Rec & FC	Rec & Conv	Online Learning	Small Batch	All Combined
Original Batch Normalization(BN)	✓	✓	✗	✗	✗	Suboptimal	✗
Time-specific BN	✓	✓	Limited	Limited	✗	Suboptimal	✗
Layer Normalization	✓	✗*	✓	✗*	✓	✓	✗*
Streaming Normalization	✓	✓	✓	✓	✓	✓	✓

Table 1: An overview of normalization techiques for different tasks. ✓: works well. ✗: does not work well. FF: Feedforward. Rec: Recurrent. FC: Fully-connected. Conv: convolutional. Limited: time-specific BN requires recording normalization statistics for each timestep and thus may not generalize to novel sequence length. *Layer normalization does not fail on these tasks but perform significantly worse than the best approaches.

[Liao, Kawaguchi, Poggio]

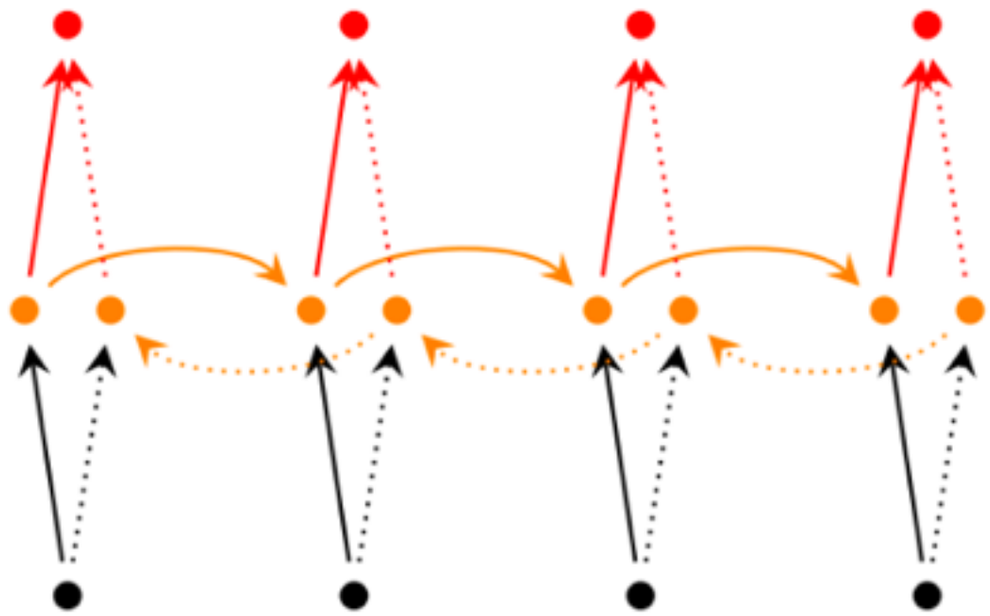
Applications

RNN Applications

- Speech Recognition
- Natural Language Processing
- Action Recognition
- Machine Translation
- Many more to come

Bidirectional RNNs

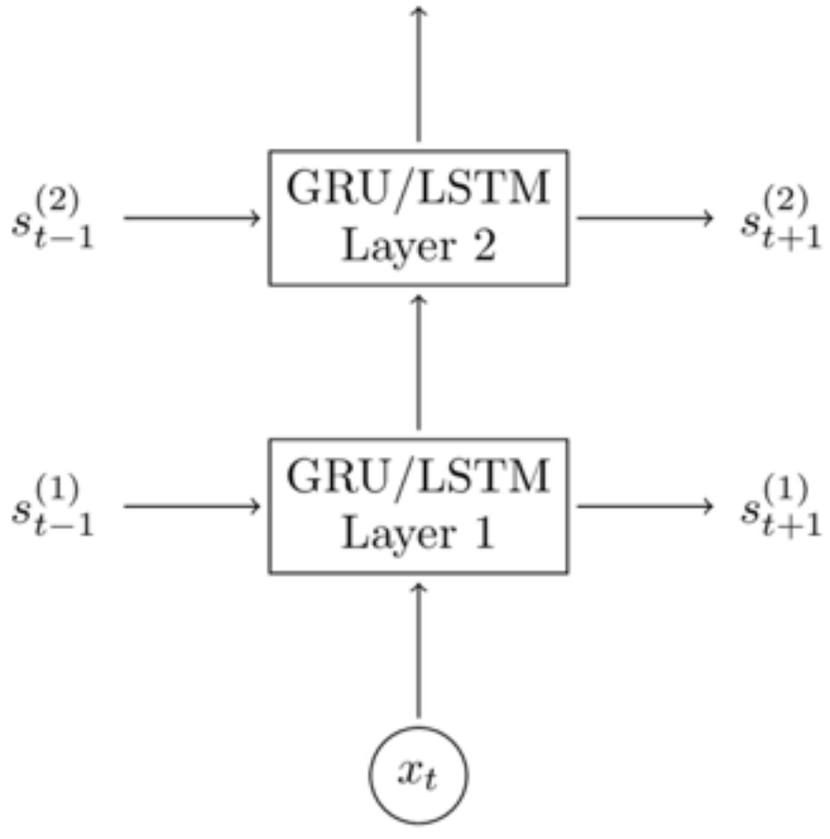
- The output at time *t* does not depend on previous time steps but also the future
  - Two RNNs stacked on top of each other



[Danny Britz]

Deep RNNs

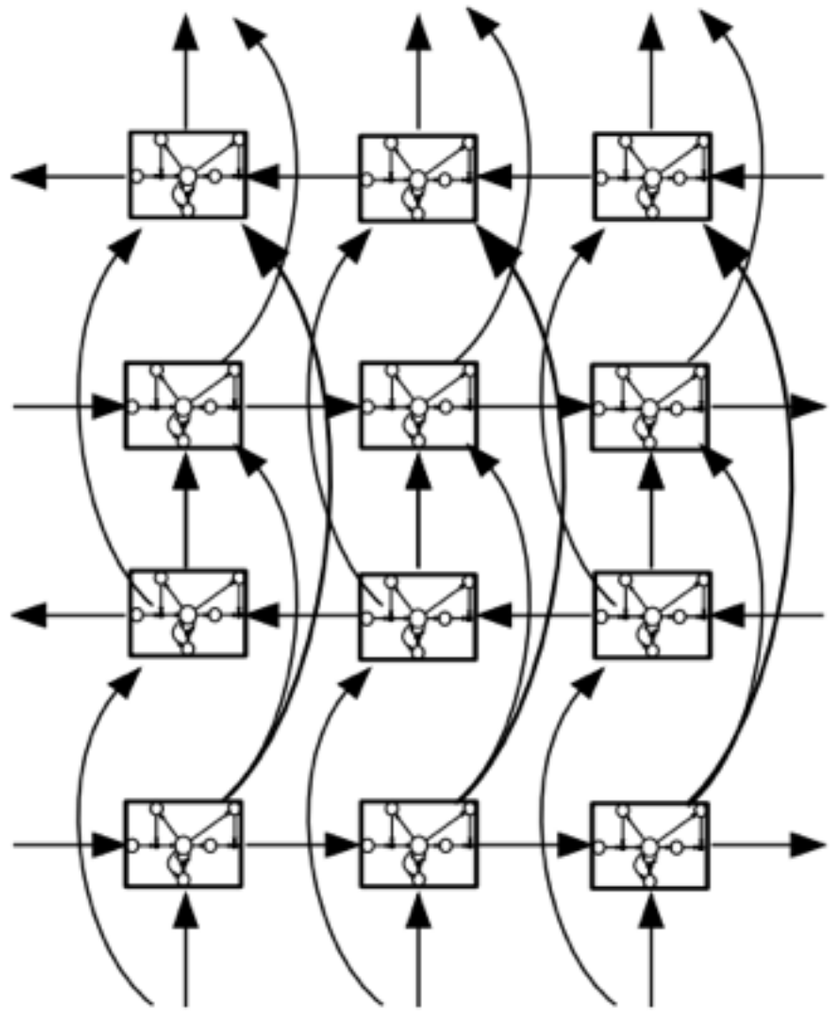
- Stack them on top of each other
  - The output of the previous RNN is the input to the next one



[Danny Britz]

Speech Recognition

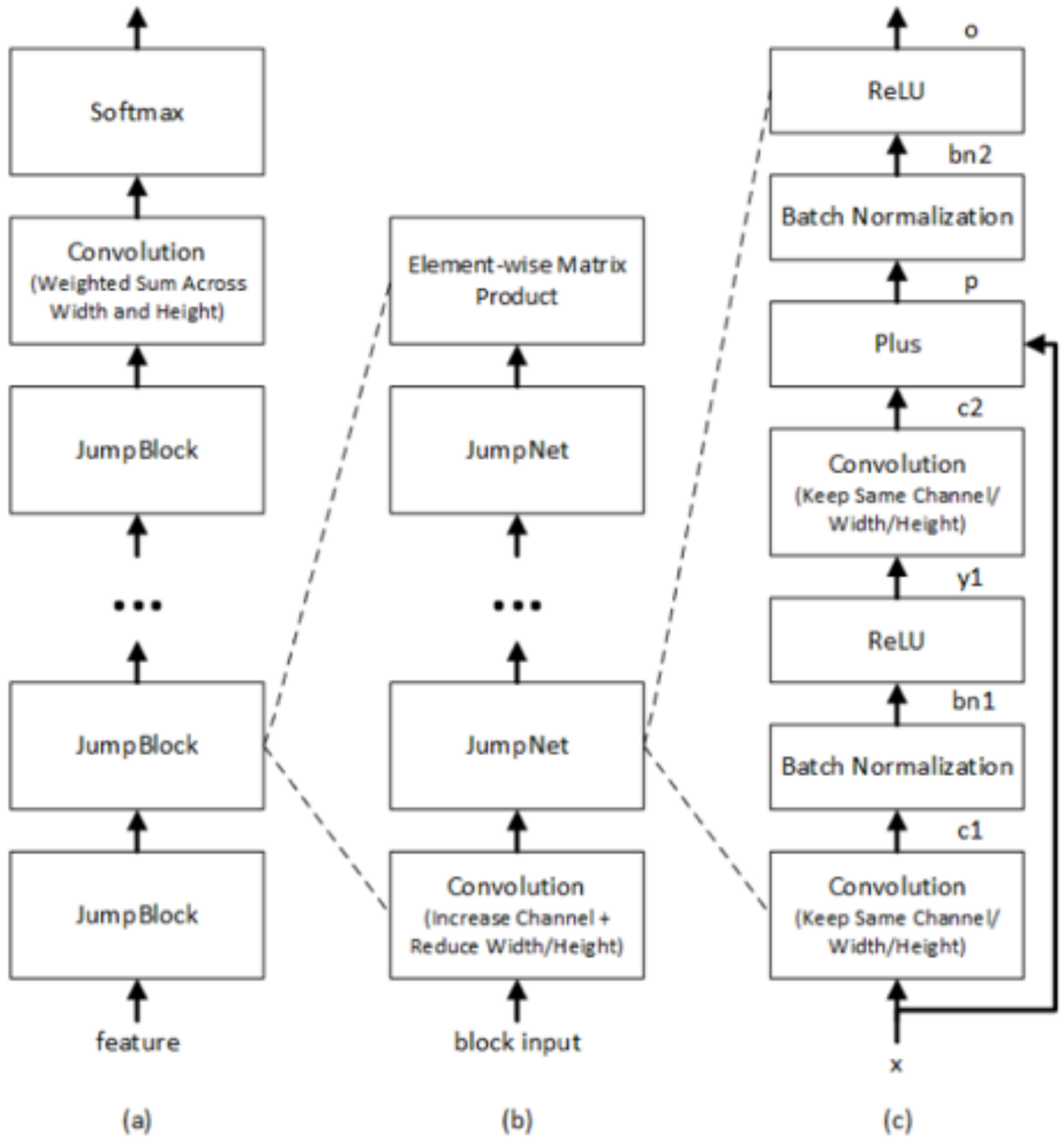
- Infer phonemes from the past and the future
  - Bidirectional Stacked LSTM



[Alex Graves, Navdeep Jaitly, Abdel-rahman Mohamed]

Conversational Speech Recognition

- Achieving human parity
- Combining RNN and CNN

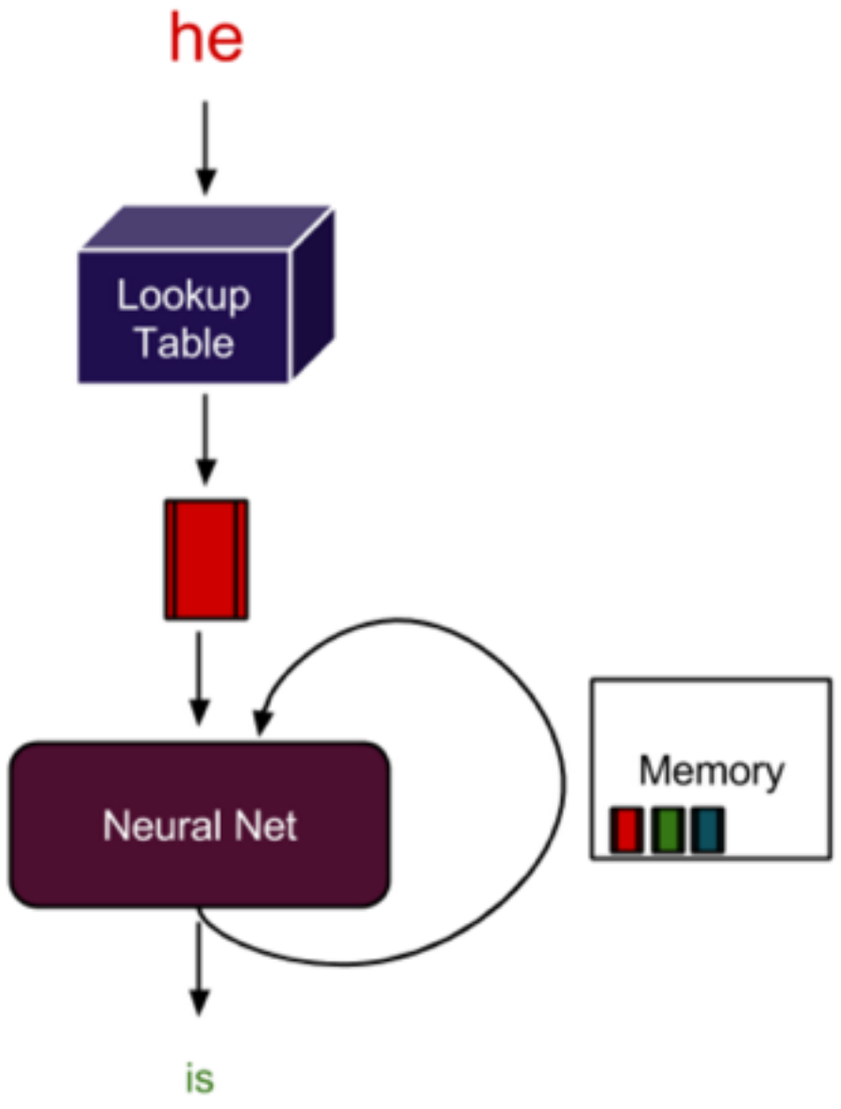


[Xiong et al.]

Natural Language Processing

- Infer character distribution from past characters in the sequence

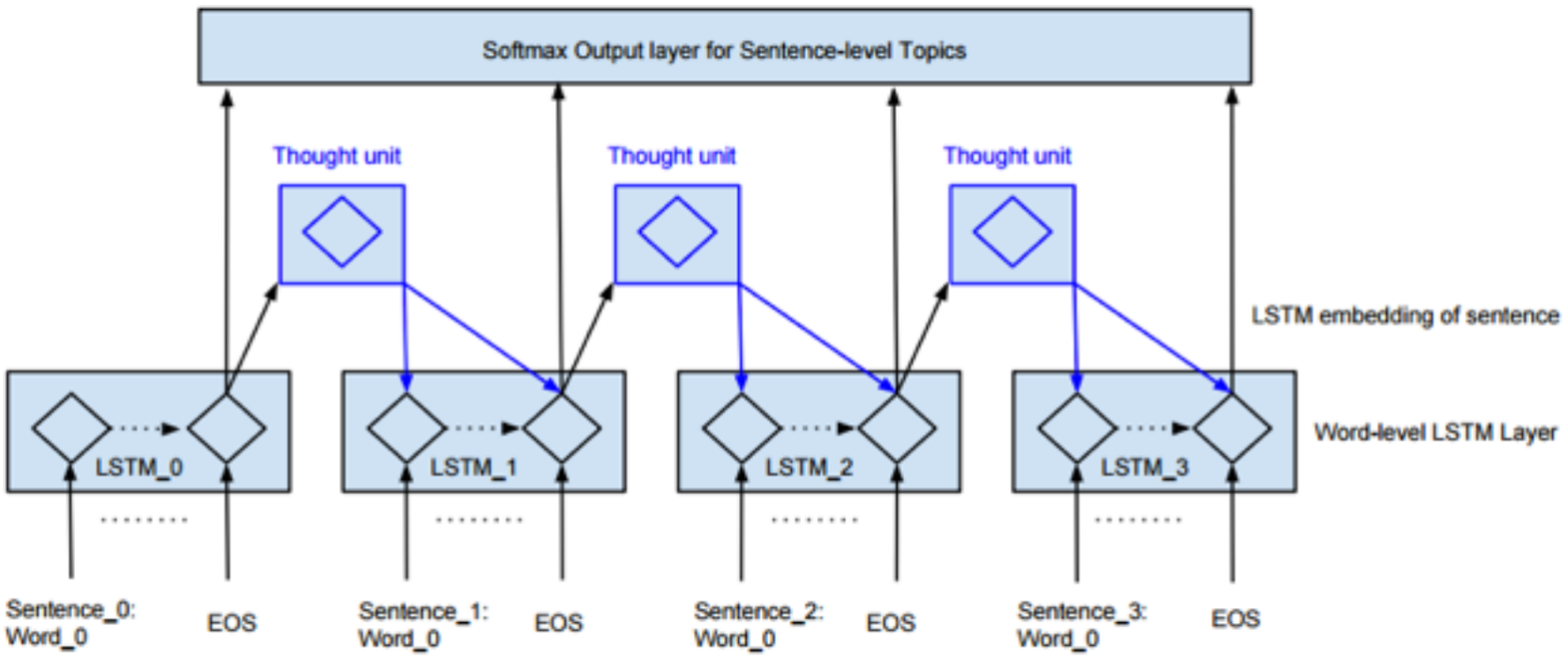
- Remember for longer durations



[Soumith Chantala]

Contextual LSTM for NLP Tasks

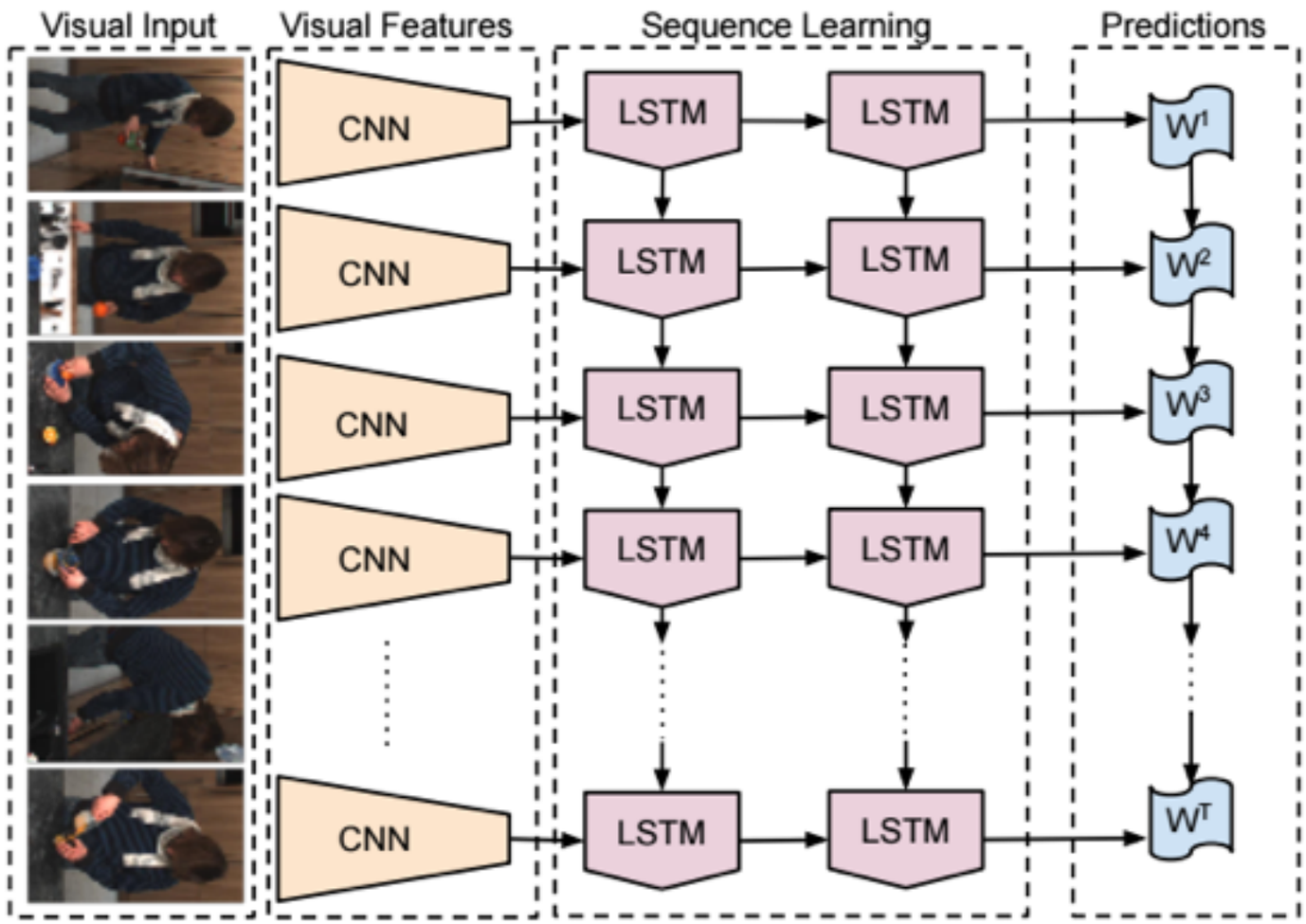
- Using a word embedding with LSTM layers to learn sentiments
- Incorporating a thought unit



[Ghost et al.]

Action Recognition

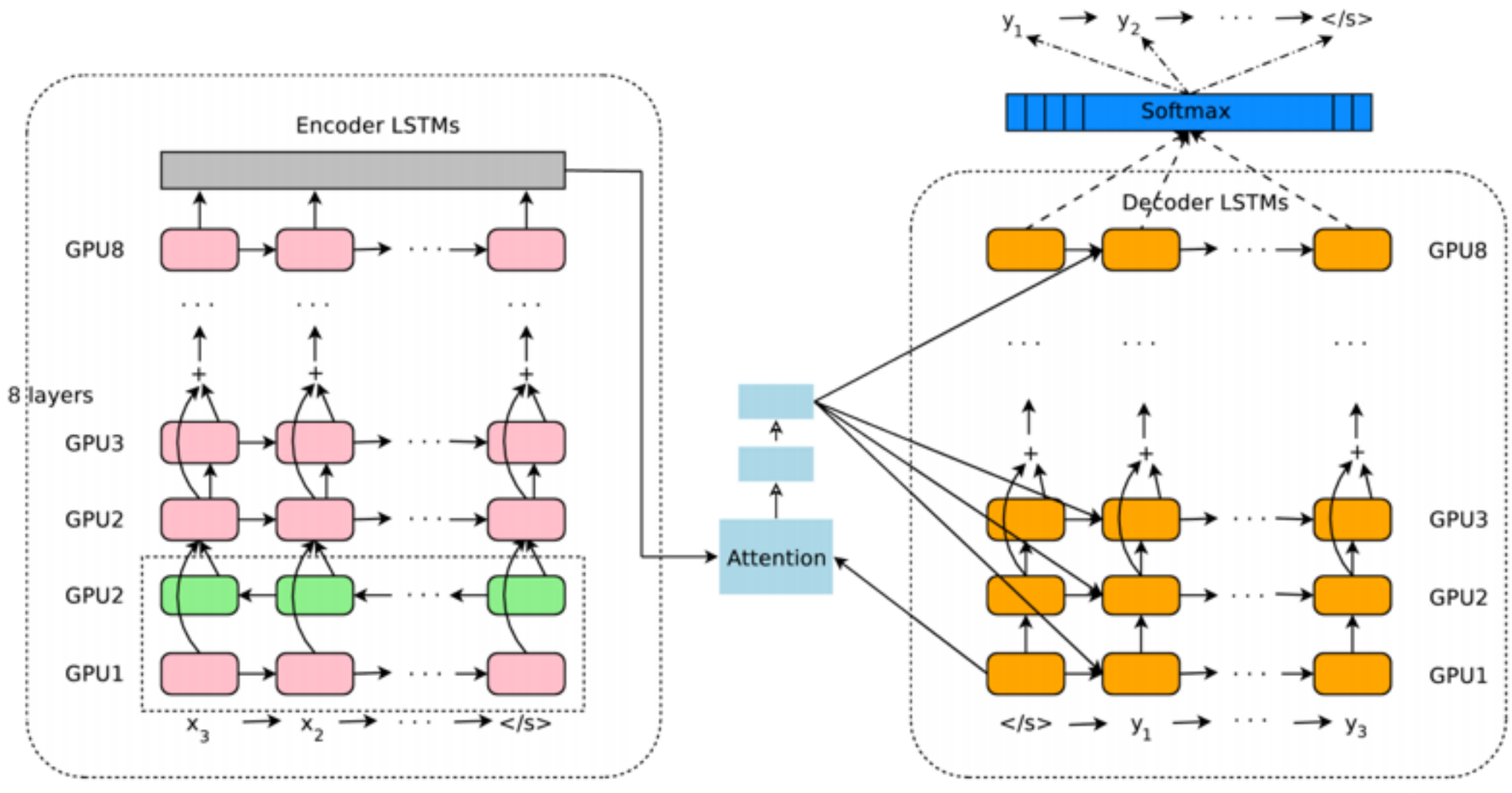
- Using LSTMs and CNN for videos
- CNN creates a feature vector that is feed into LSTM



[Donahue et al.]

Google's Neural Machine Translation System

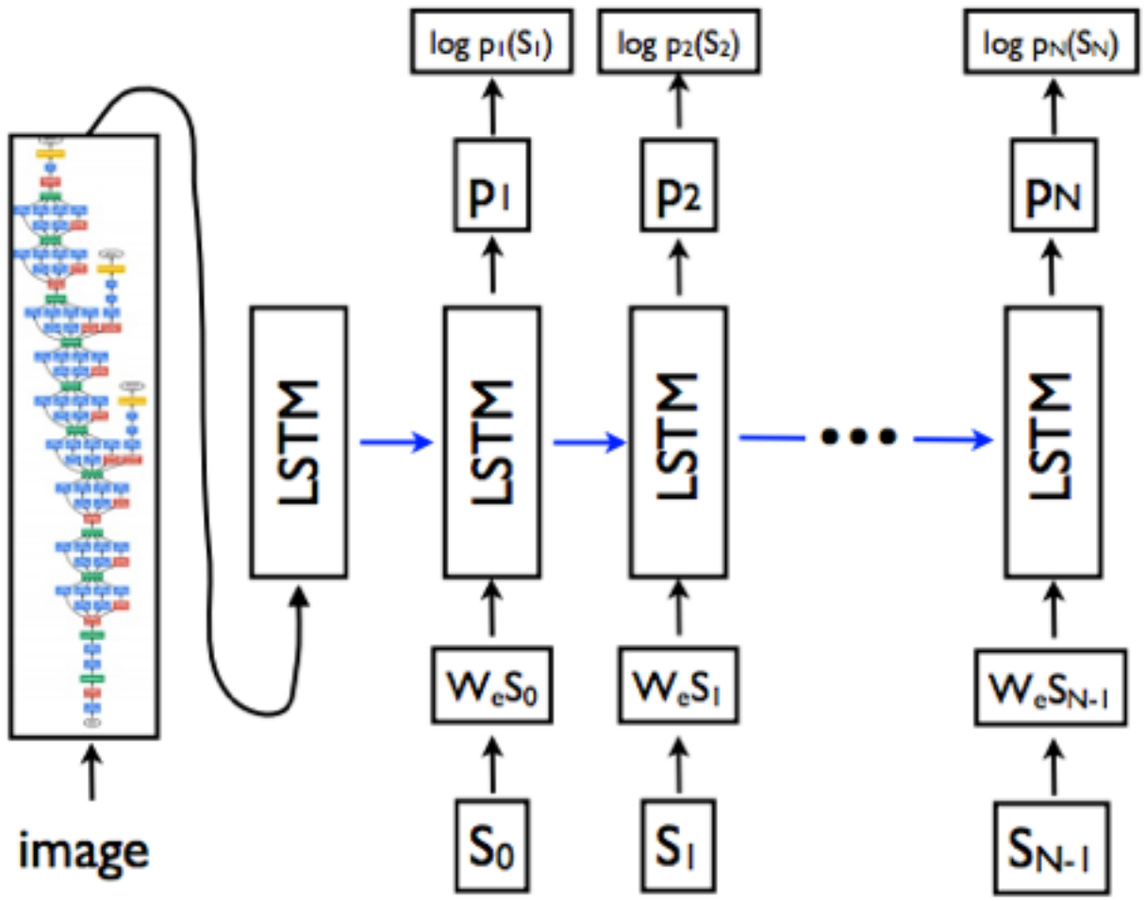
- Encoder and Decoder LSTMs
- Attention model



[Yonghui Wu et al.]

Image Captioning Pt 1

- Combination of CNN and LSTM to caption images
  - Using a pretrained CNN for visual features



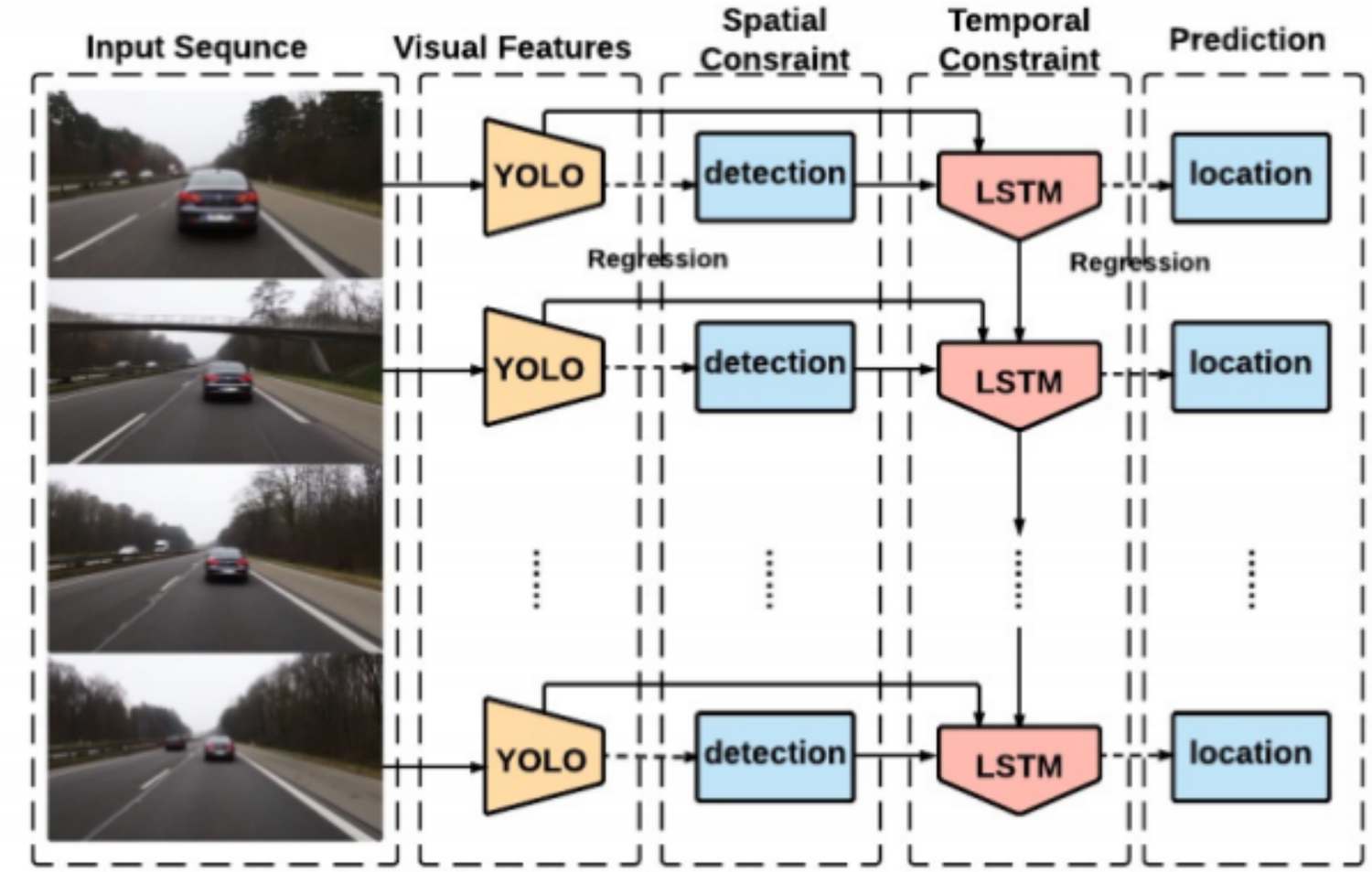
[Vinyals, Toshev, Bengio, Erhan]

Image Captioning Pt 2

[Vinyals, Toshev, Bengio, Erhan]

Object Tracking

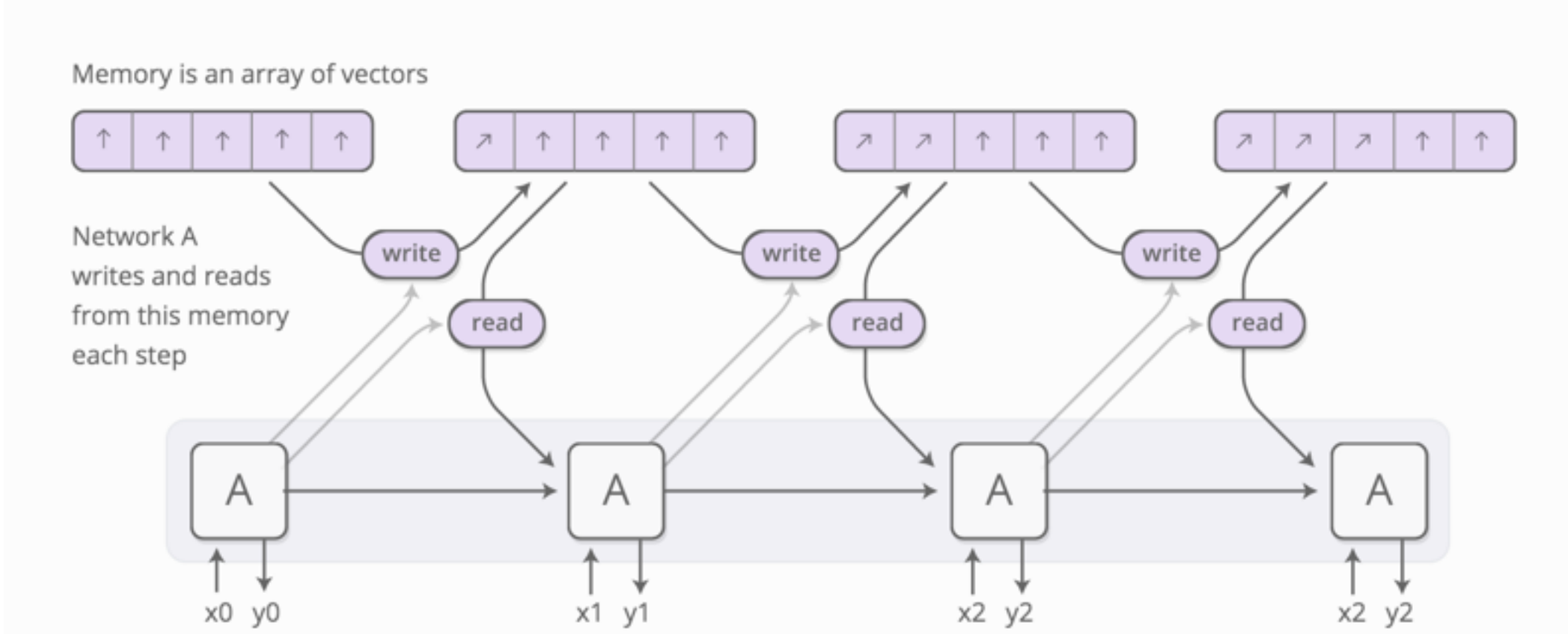
- Using an object detector with an LSTM to track objects
- Model the dynamics of video



[Ning, Zhang, Huang, He, Wang

Neural Turing Machines

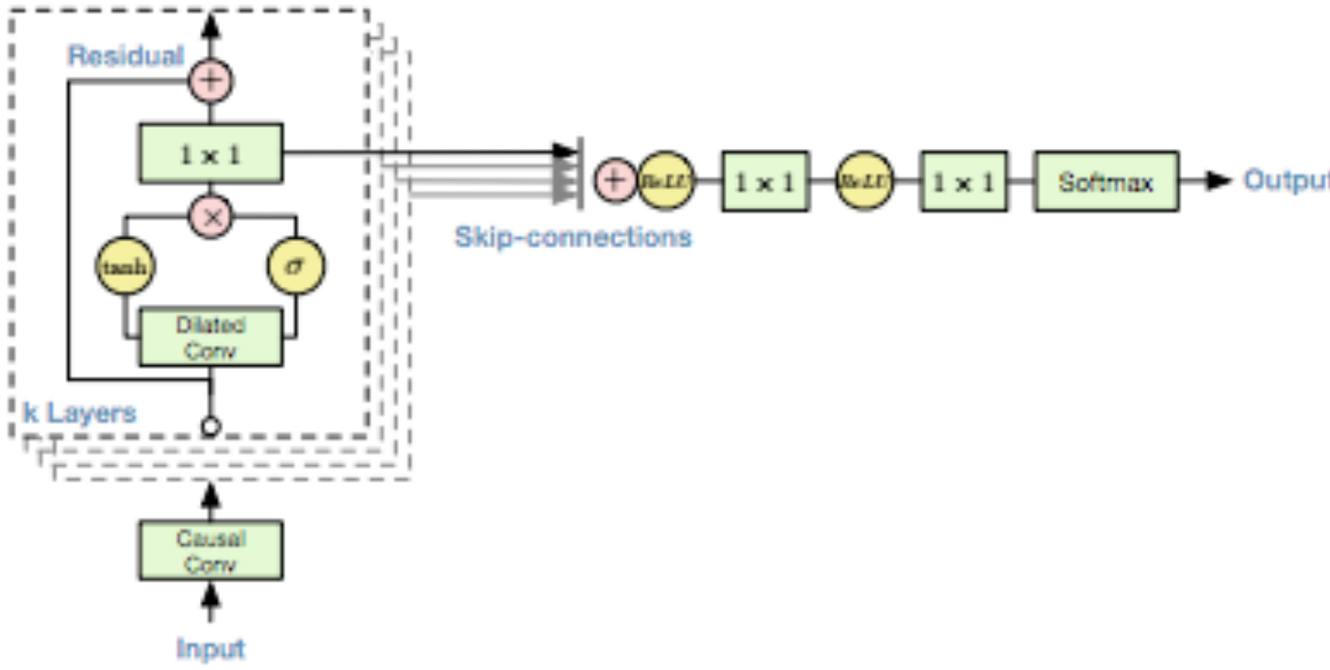
- LSTM with external memory
- Analogous to a Turing Machine



[Chris Olah]

WaveNet

- Using stack of diluted layers
- To generate next sample, it models conditional probability given previous samples



[van den Oord et al.]

DoomBot

- Doom Competition
  - Facebook won 1st place (F1)
  - https://www.youtube.com/watch?v=94EPSjQH38Y

Character-level RNN Language Models

Goal

- Model the probability distribution of the next character in a sequence
    - Given the previous characters

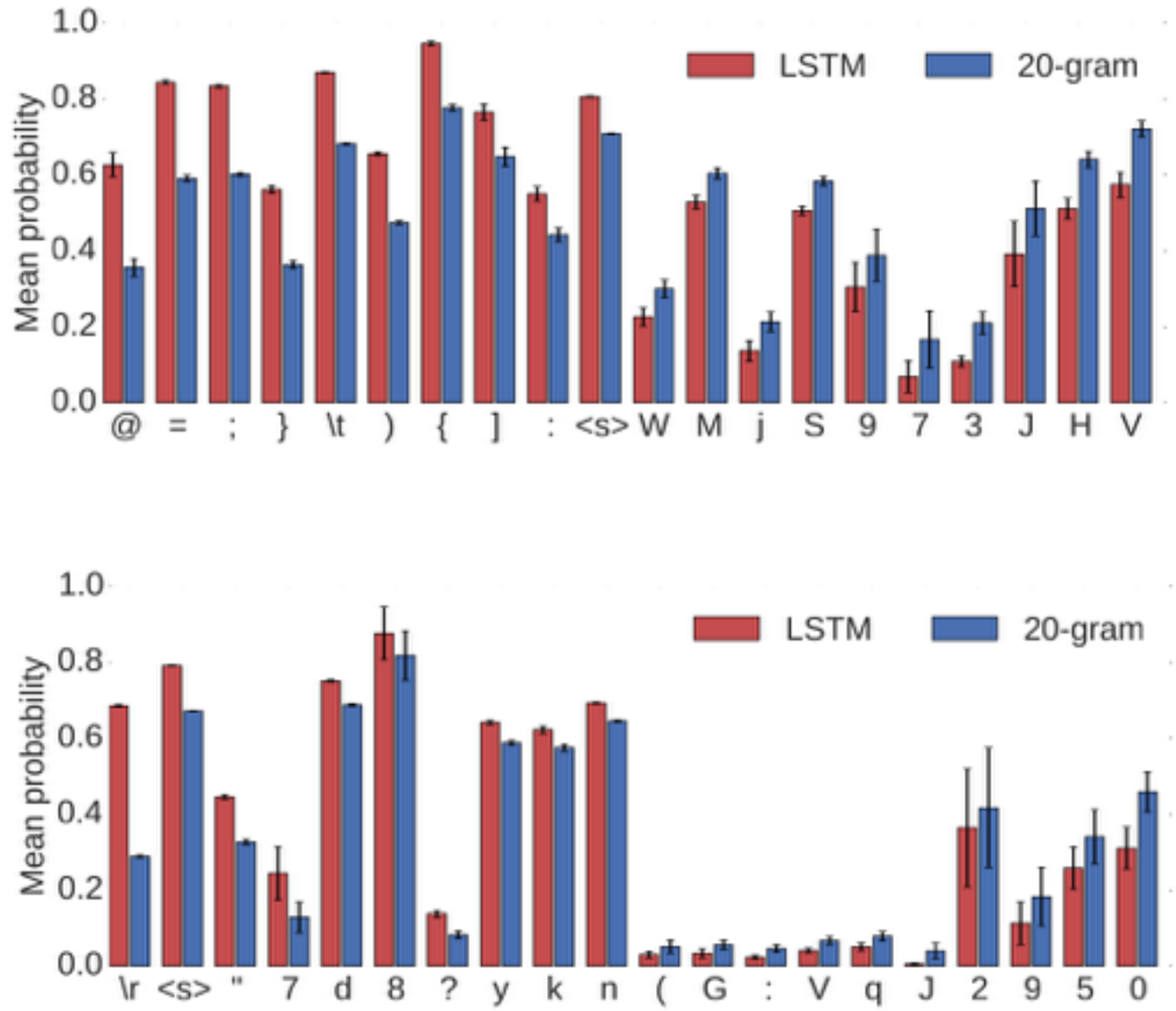
$$P(x_t = k | x_{1:t-1}) = \frac{\exp(w_k h_t)}{\sum_{j=1}^{|V|} \exp(w_j h_t)}$$

[Susanto, Chieu, Lu]

N-grams

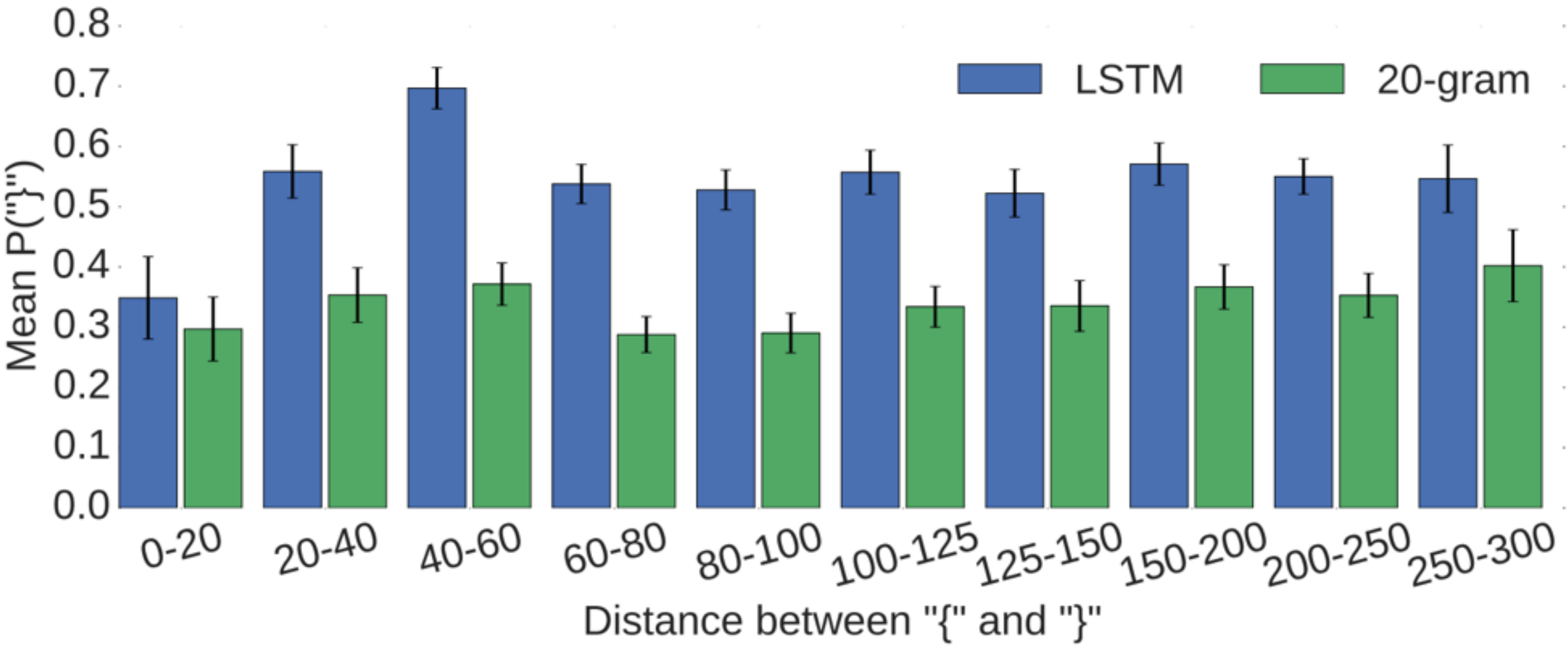
- Group the characters into n characters
  - n=1 unigram
  - n=2 bigram
- Useful for protein sequencing, computational linguistics, etc.

Comparing Against N-Grams



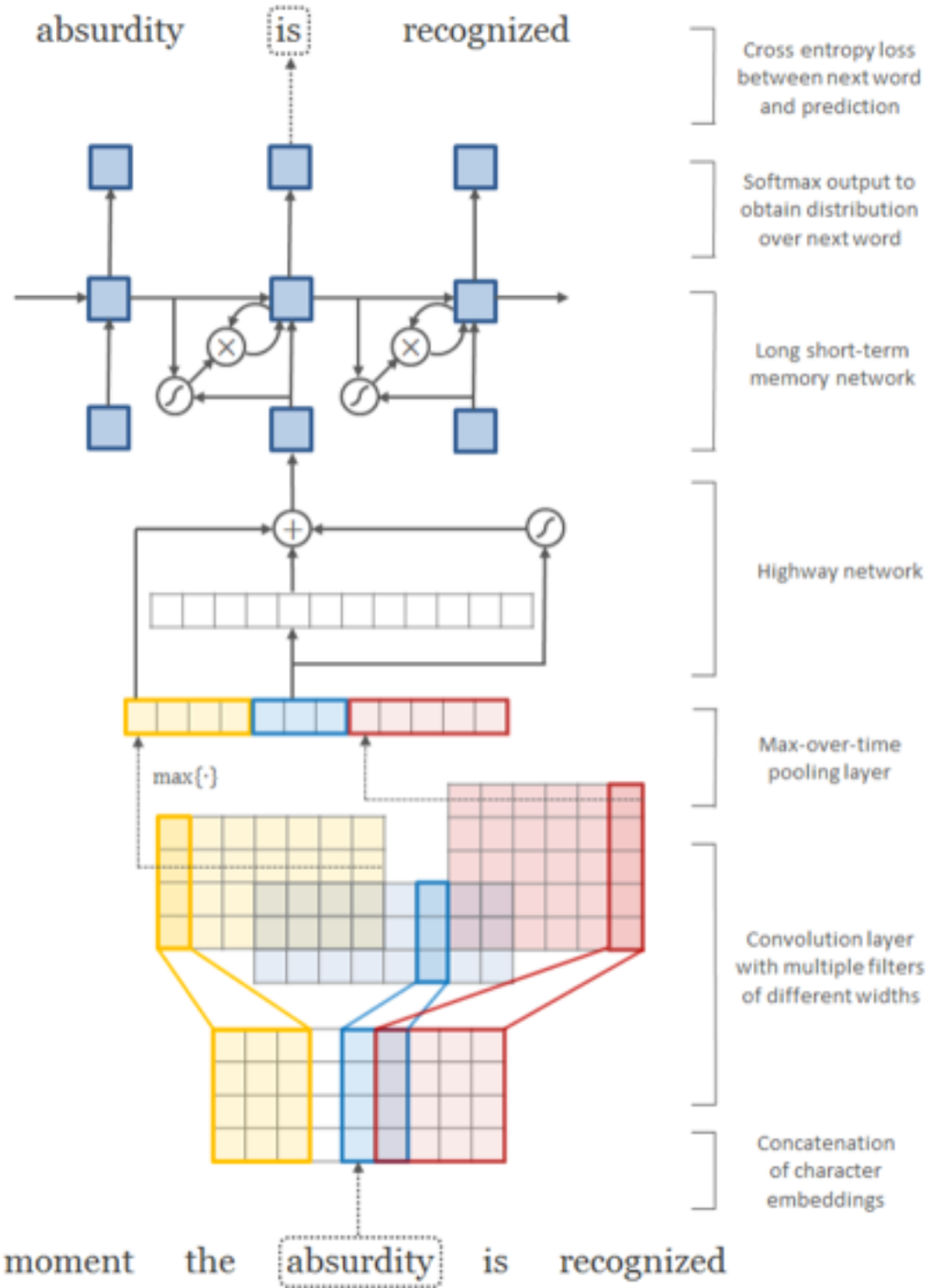
[Karpathy, Johnson, Li]

Remembering for Longer Durations



[Karpathy, Johnson, Li]

Character-Aware Neural Language Models



[Kim, Jernite, Sontag, Rush]

The Effectiveness of an RNN

```
#define REG_PG    vesa_slot_addr_pack
#define PFM_NOCOMP  AFSR(0, load)
#define STACK_DDR(type)     (func)

#define SWAP_ALLOCATE(nr)     (e)
#define emulate_sigs()  arch_get_unaligned_child()
#define access_rw(TST)  asm volatile("movd %%esp, %0, %3" : : "r" (0));   \
  if (__type & DO_READ)

static void stat_PC_SEC __read_mostly offsetof(struct seq_argsqueue, \
          pC>[1]);

static void
os_prefix(unsigned long sys)
{
#ifdef CONFIG_PREEMPT
  PUT_PARAM_RAID(2, sel) = get_state_state();
  set_pid_sum((unsigned long)state, current_state_str(),
           (unsigned long)-1->lr_full; low;
}
```

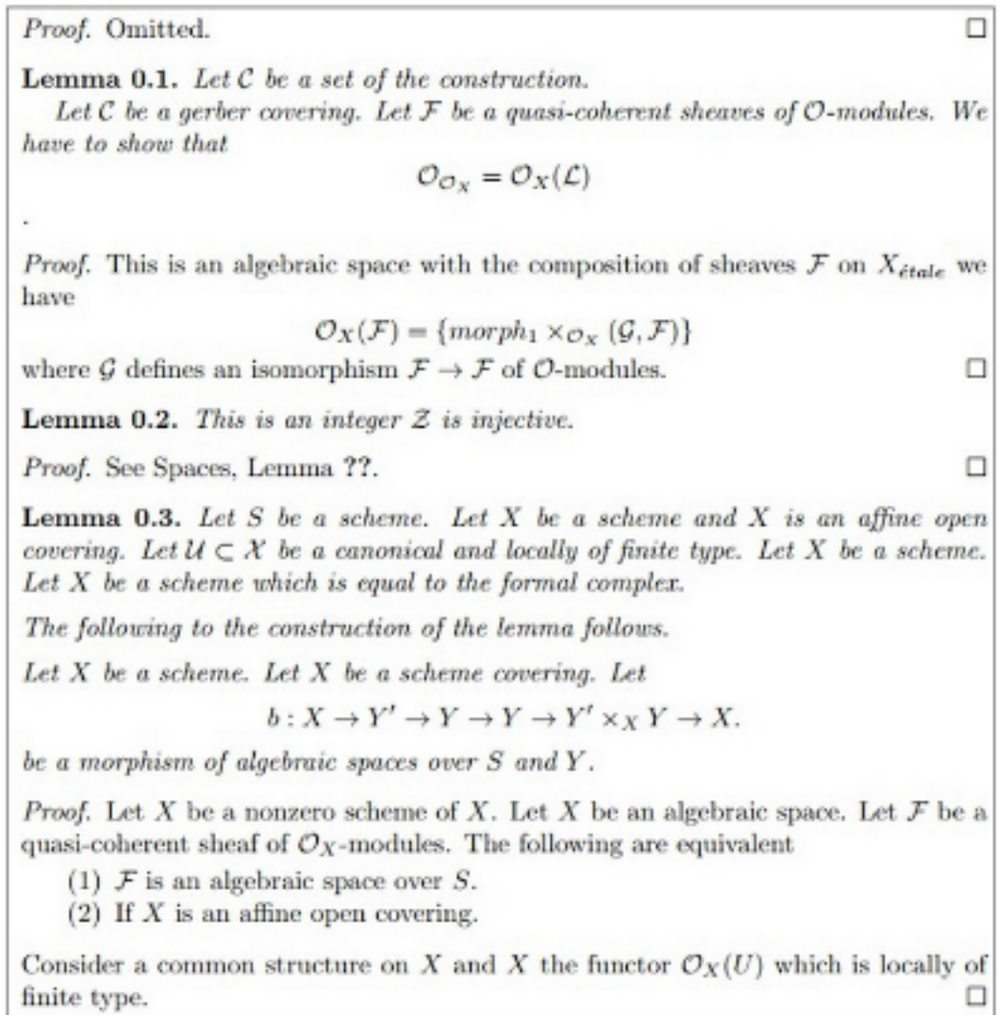
[Andrej Karpathy]

The Effectiveness of an RNN

```
Naturalism and decision for the majority of Arab countries' capitalide was grounded
by the Irish language by [[John Clair]], [[An Imperial Japanese Revolt]], associated
with Guangzham's sovereignty. His generals were the powerful ruler of the Portugal
in the [[Protestant Immineners]], which could be said to be directly in Cantonese
Communication, which followed a ceremony and set inspired prison, training. The
emperor travelled back to [[Antioch, Perth, October 25|21]] to note, the Kingdom
of Costa Rica, unsuccessful fashioned the [[Thrales]], [[Cynth's Dajoard]], known
in western [[Scotland]], near Italy to the conquest of India with the conflict.
Copyright was the succession of independence in the slop of Syrian influence that
was a famous German movement based on a more popular servicious, non-doctrinal
and sexual power post. Many governments recognize the military housing of the
[[Civil Liberalization and Infantry Resolution 265 National Party in Hungary]],
that is sympathetic to be to the [[Punjab Resolution]]
(PJS)[http://www.humah.yahoo.com/guardian.
cfm/7754800786d17551963s89.htm Official economics Adjoint for the Nazism, Montgomery
was swear to advance to the resources for those Socialism's rule,
was starting to signing a major tripad of aid exile.]]
```

[Andrej Karpathy]

The Effectiveness of an RNN



Proof. Omitted. □

Lemma 0.1. *Let $\mathcal{C}$ be a set of the construction.*

Let $\mathcal{C}$ be a gerber covering. Let $\mathcal{F}$ be a quasi-coherent sheaves of $\mathcal{O}$-modules. We have to show that

$$\mathcal{O}_{\mathcal{O}_X} = \mathcal{O}_X(\mathcal{L})$$

.

Proof. This is an algebraic space with the composition of sheaves $\mathcal{F}$ on $X_{\acute{e}tale}$ we have

$$\mathcal{O}_X(\mathcal{F}) = \{morph_1 \times_{\mathcal{O}_X} (\mathcal{G}, \mathcal{F})\}$$

where $\mathcal{G}$ defines an isomorphism $\mathcal{F} \to \mathcal{F}$ of $\mathcal{O}$-modules. □

Lemma 0.2. *This is an integer $\mathcal{Z}$ is injective.*

Proof. See Spaces, Lemma ??. □

Lemma 0.3. *Let S be a scheme. Let X be a scheme and X is an affine open covering. Let $\mathcal{U} \subset \mathcal{X}$ be a canonical and locally of finite type. Let X be a scheme. Let X be a scheme which is equal to the formal complex.*

The following to the construction of the lemma follows.

Let X be a scheme. Let X be a scheme covering. Let

$$b : X \to Y' \to Y \to Y \to Y' \times_X Y \to X.$$

be a morphism of algebraic spaces over S and Y.

Proof. Let X be a nonzero scheme of X. Let X be an algebraic space. Let $\mathcal{F}$ be a quasi-coherent sheaf of $\mathcal{O}_X$-modules. The following are equivalent

(1) $\mathcal{F}$ is an algebraic space over S.
(2) If X is an affine open covering.

Consider a common structure on X and X the functor $\mathcal{O}_X(U)$ which is locally of finite type. □

This since $\mathcal{F} \in \mathcal{F}$ and $x \in \mathcal{G}$ the diagram

$$\begin{array}{ccc} \mathcal{S} & \longrightarrow & \\ \downarrow & & \\ \xi & \longrightarrow & \mathcal{O}_{X'} \\ & \searrow^{\text{gor}_s} & \uparrow \\ & & = a' \longrightarrow \\ & & \updownarrow \\ & & = a' \longrightarrow a \end{array} \qquad \begin{array}{c} \\ \\ \\ \\ X \\ \downarrow \\ \text{Mor}_{Sets} \quad \text{d}(\mathcal{O}_{X_{fppf}}, \mathcal{G}) \end{array}$$

$\text{Spec}(K_\psi)$

is a limit. Then $\mathcal{G}$ is a finite type and assume S is a flat and $\mathcal{F}$ and $\mathcal{G}$ is a finite type f_*. This is of finite type diagrams, and

- the composition of $\mathcal{G}$ is a regular sequence,
- $\mathcal{O}_{X'}$ is a sheaf of rings.

□

Proof. We have see that $X = \text{Spec}(R)$ and $\mathcal{F}$ is a finite type representable by algebraic space. The property $\mathcal{F}$ is a finite morphism of algebraic stacks. Then the cohomology of X is an open neighbourhood of U. □

Proof. This is clear that $\mathcal{G}$ is a finite presentation, see Lemmas ??.
A *reduced above* we conclude that U is an open covering of $\mathcal{C}$. The functor $\mathcal{F}$ is a "field

$$\mathcal{O}_{X,s} \longrightarrow \mathcal{F}_{\overline{s}} \quad \text{-1}(\mathcal{O}_{X_{\acute{e}tale}}) \longrightarrow \mathcal{O}_{X_\lambda}^{-1} \mathcal{O}_{X_\lambda}(\mathcal{O}_{X_\lambda}^{\overline{s}})$$

is an isomorphism of covering of $\mathcal{O}_{X_i}$. If $\mathcal{F}$ is the unique element of $\mathcal{F}$ such that X is an isomorphism.
The property $\mathcal{F}$ is a disjoint union of Proposition ?? and we can filtered set of presentations of a scheme $\mathcal{O}_X$-algebra with $\mathcal{F}$ are opens of finite type over S.
If $\mathcal{F}$ is a scheme theoretic image points. □

If $\mathcal{F}$ is a finite direct sum $\mathcal{O}_{X_\lambda}$ is a closed immersion, see Lemma ??. This is a sequence of $\mathcal{F}$ is a similar morphism.

[Andrej Karpathy]

The Effectiveness of an RNN

Trained on *War & Peace*

Iteration: 100

```
tyntd-iafhatawiaoihrdemot  lytdws  e ,tfti, astai f ogoh eoase rrranbyne 'nhthnee e
plia tklrgd t o idoe ns,smtt   h ne etie h,hregtrs nigtike,aoaenns lng
```

Iteration: 300

```
"Tmont thithey" fomesscerliund
Keushey. Thom here
sheulke, anmerenith ol sivh I lalterthend Bleipile shuwy fil on aseterlome
coaniogennc Phe lism thond hon at. MeiDimorotion in ther thize."
```

Iteration: 2000

```
"Why do what that day," replied Natasha, and wishing to himself the fact the
princess, Princess Mary was easier, fed in had oftened him.
Pierre aking his soul came to the packs and drove up his father-in-law women.
```

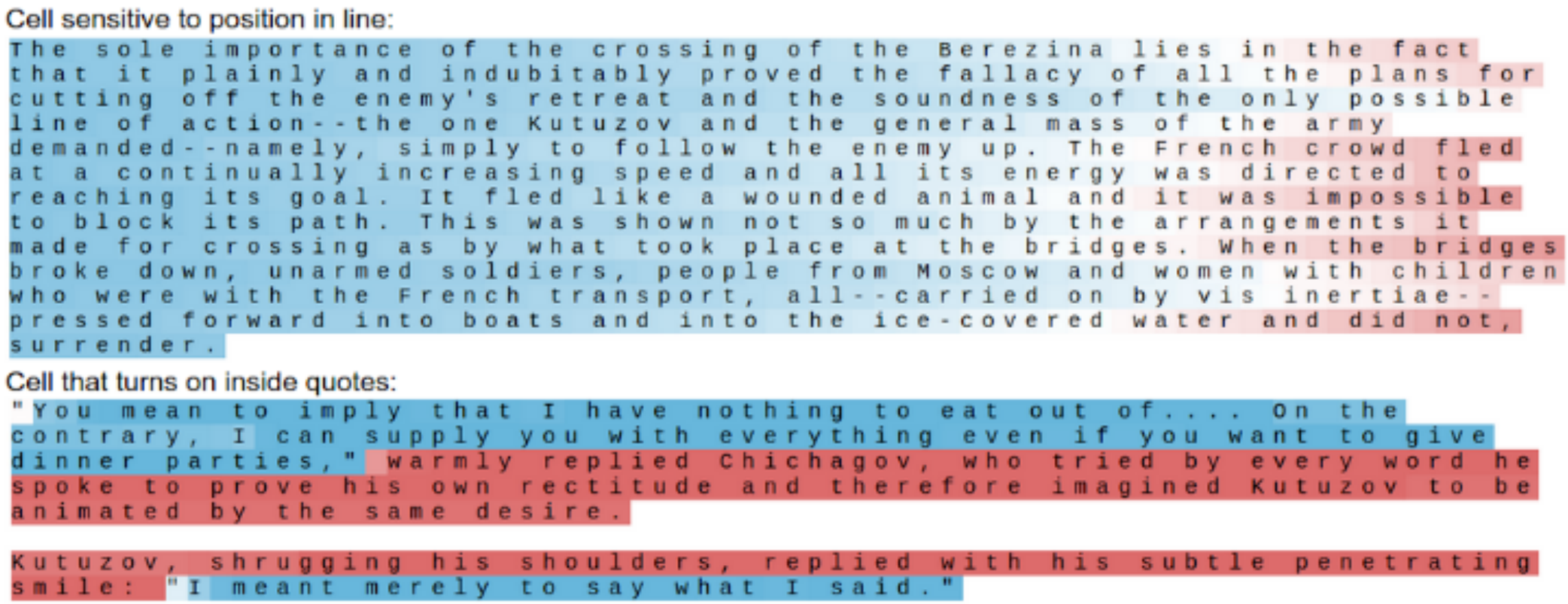
[Andrej Karpathy]

Visualize the Neurons of an RNN

```
static int __dequeue_signal(struct sigpending *pending, sigset_t *mask,
        siginfo_t *info)
{
  int sig = next_signal(pending, mask);
  if (sig) {
    if (current->notifier) {
      if (sigismember(current->notifier_mask, sig)) {
        if (!(current->notifier)(current->notifier_data)) {
          clear_thread_flag(TIF_SIGPENDING);
          return 0;
        }
      }
    }
    collect_signal(sig, pending, info);
  }
  return sig;
}
```

[Andrej Karpathy]

Visualize the Neurons of an RNN

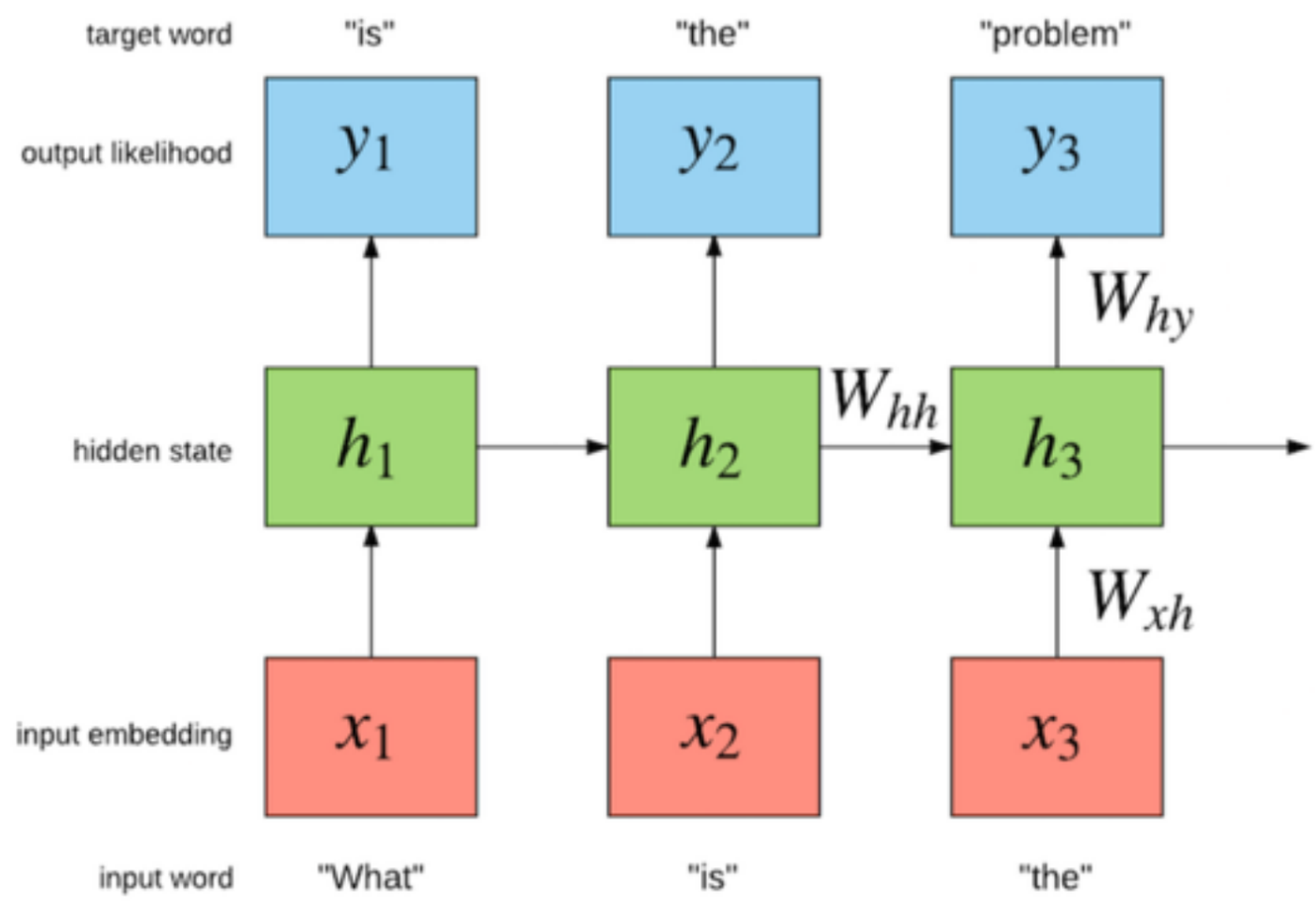


[Andrej Karpathy]

Word-level RNN Language Models

Goals

- Model the probability distribution of the next word in a sequence
  - Given the previous words



[Nicholas Leonard]

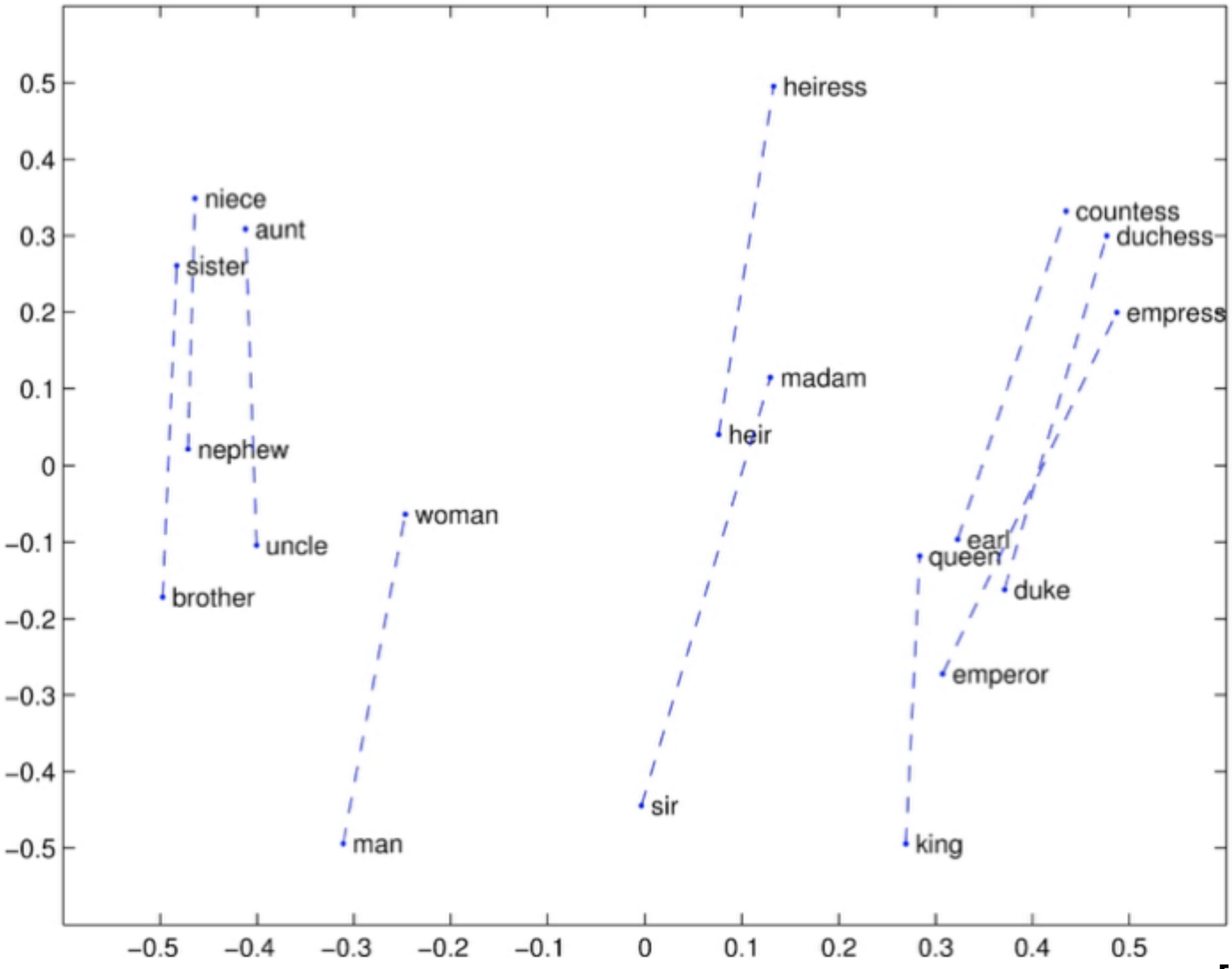
Global Vectors for Word Representation (GloVe)

- Provide semantic information/context for words
- Unsupervised method for learning word representations

$$J(\theta) = \frac{1}{2} \sum_{i,j=1}^{W} f(P_{ij})(u_i^T v_j - \log P_{ij})^2$$

[Richard Socher]

Glove Visualization

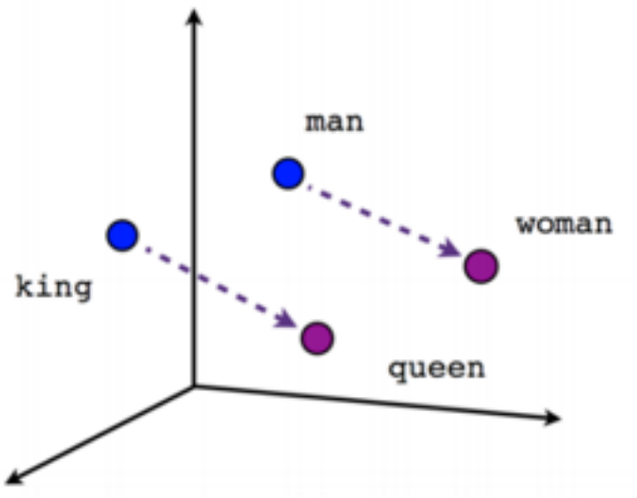


[Richard Socher]

Word2Vec

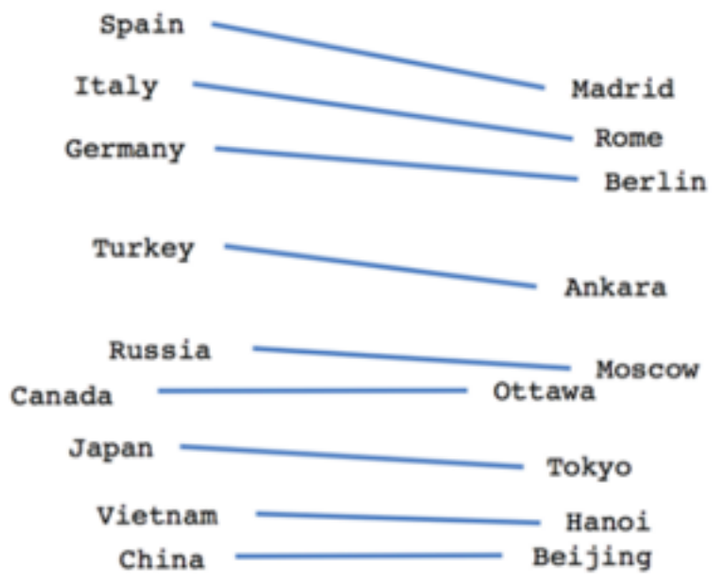
- Learn word embeddings
- Shallow, two-layer neural network
- Trained to reconstruct linguistic context between words
- Produces a vector space for the words

Word2Vec Visualization



Male-Female

Verb tense



Country-Capital

[Tensorflow]

Question Time

- What is the main difference between word2vec and GloVe?

Word2vec with RNNs

Method	Adjectives	Nouns	Verbs	All
LSA-80	9.2	11.1	17.4	12.8
LSA-320	11.3	18.1	20.7	16.5
LSA-640	9.6	10.1	13.8	11.3
RNN-80	9.3	5.2	30.4	16.2
RNN-320	18.2	19.0	45.0	28.5
RNN-640	21.0	25.2	54.8	34.7
RNN-1600	**23.9**	**29.2**	**62.2**	**39.6**

Table 2: Results for identifying syntactic regularities for different word representations. Percent correct.

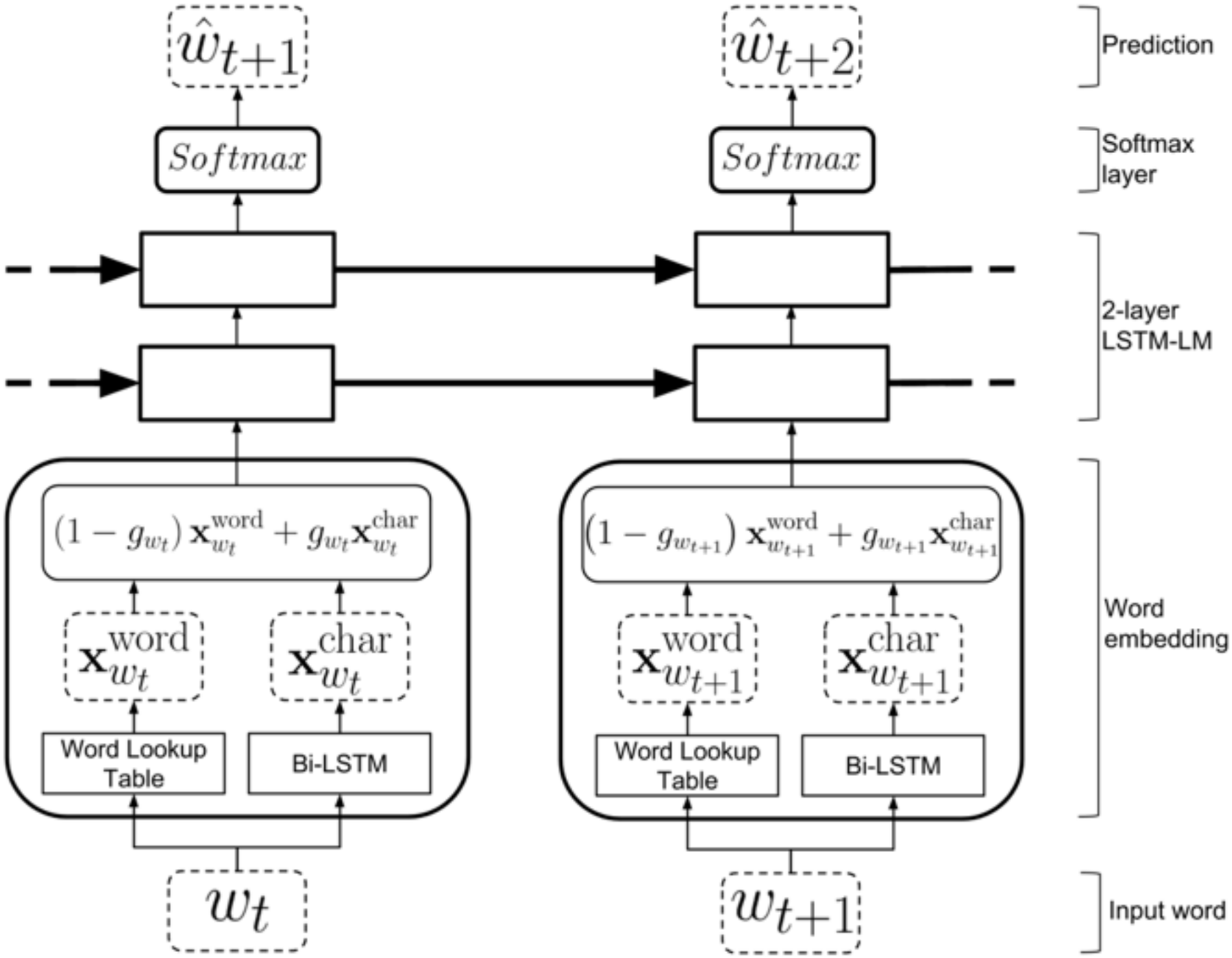
[Mikolov, Yih, Zweig]

Word RNN trained on Shakespeare

```
LEONTES:
Why, my Irish time?
And argue in the lord; the man mad, must be deserved a spirit as drown the warlike Pray him, how seven in.

KING would be made that, methoughts I may married a Lord dishonour
Than thou that be mine kites and sinew for his honour
In reason prettily the sudden night upon all shalt bid him thus again. times than one from mine unaccustom'

LARTIUS:
O,'tis aediles, fight!
Farewell, it himself have saw.

SLY:
Now gods have their VINCENTIO:
Whipt fearing but first I know you you, hinder truths.

ANGELO:
This are entitle up my dearest state but deliver'd.

DUKE look dissolved: seemeth brands
That He being and
full of toad, they knew me to joy.
```

[Sung Kim]

Gated Word RNN



[Miyamoto, Cho]

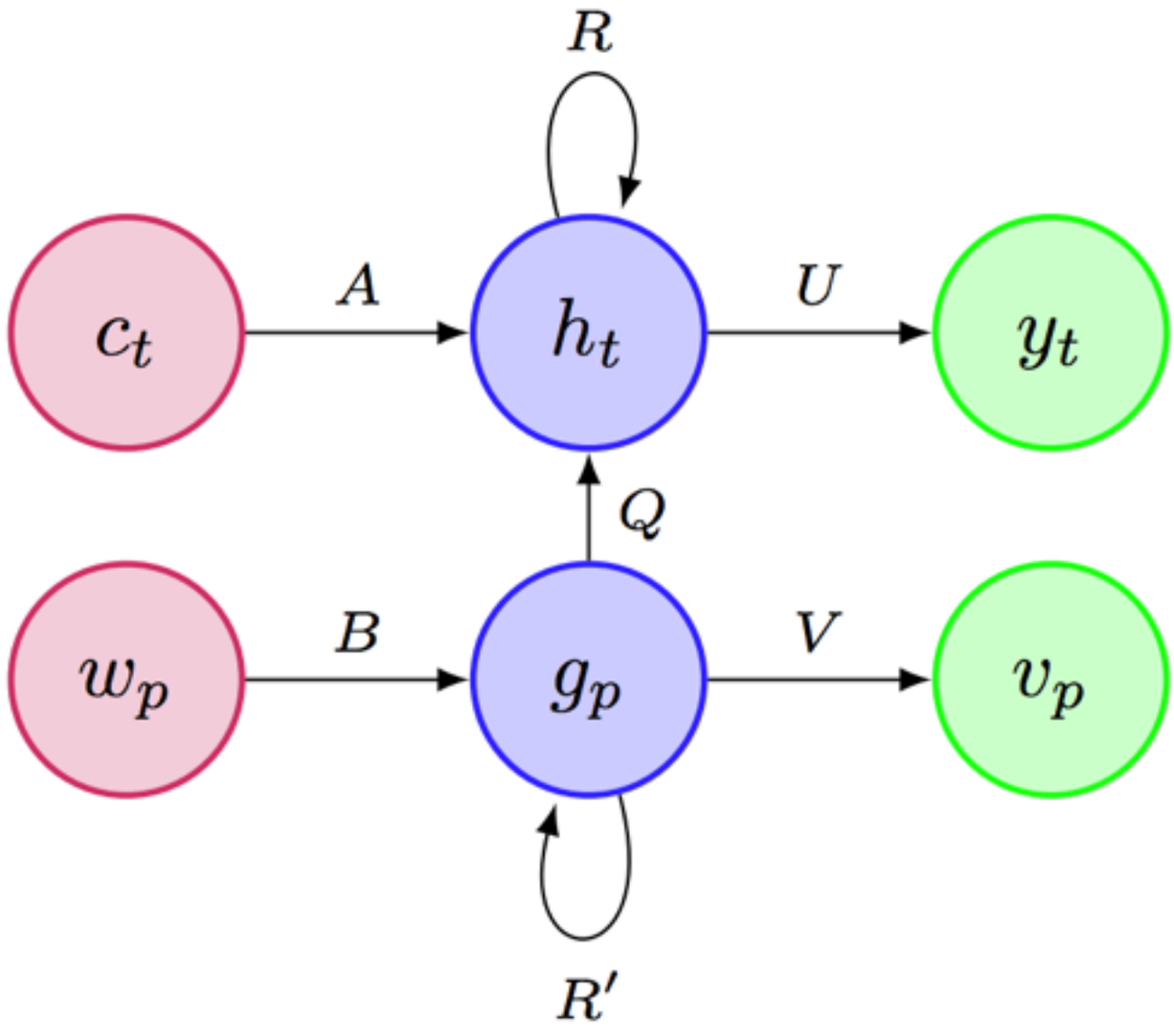
Gated Word RNN Results

Model	PTB		BBC		IMDB	
	Validation	**Test**	**Validation**	**Test**	**Validation**	**Test**
Gated Word & Char, adaptive	117.49	113.87	**78.56**	**87.16**	71.99	72.29
Gated Word & Char, adaptive (Pre-train)	117.03	112.90	80.37	87.51	71.16	71.49
Gated Word & Char, $g = 0.25$	119.45	115.55	79.67	88.04	71.81	72.14
Gated Word & Char, $g = 0.25$ (Pre-train)	**117.01**	**113.52**	80.07	87.99	**70.60**	**70.87**
Gated Word & Char, $g = 0.5$	126.01	121.99	89.27	94.91	106.78	107.33
Gated Word & Char, $g = 0.5$ (Pre-train)	117.54	113.03	82.09	88.61	109.69	110.28
Gated Word & Char, $g = 0.75$	135.58	135.00	105.54	111.47	115.58	116.02
Gated Word & Char, $g = 0.75$ (Pre-train)	179.69	172.85	132.96	136.01	106.31	106.86
Word Only	118.03	115.65	84.47	90.90	72.42	72.75
Character Only	132.45	126.80	88.03	97.71	98.10	98.59
Word & Character	125.05	121.09	88.77	95.44	77.94	78.29
Word & Character (Pre-train)	122.31	118.85	84.27	91.24	80.60	81.01
Non-regularized LSTM (Zaremba, 2014)	120.7	114.5	-	-	-	-

Table 1: Validation and test perplexities on Penn Treebank (PTB), BBC, IMDB Movie Reviews datasets.

[Miyamoto, Cho]

Combining Character & Word Level



[Bojanowski, Joulin, Mikolov]

Question Time

- In which situation(s) can you see character-level RNN more suitable than a word-level RNN?

Character vs Word Level Models

Character vs Word-Level Models

	EN-Wikipedia				EN-WSJ			
	Acc.	P	R	F_1	Acc.	P	R	F_1
					Word-based Approach			
LM ($N = 3$)	94.94	89.34	84.61	86.91	95.59	91.56	78.79	84.70
LM ($N = 5$)	94.93	89.42	84.41	86.84	95.62	91.72	78.79	84.77
CRF-WORD	96.60	94.96	87.16	90.89	97.64	93.12	90.41	91.75
Chelba and Acero (2006)			n/a		97.10	-	-	-
					Character-based Approach			
CRF-CHAR	96.99	94.60	89.27	91.86	97.00	94.17	84.46	89.05
LSTM-SMALL	96.95	93.05	90.59	91.80	97.83	93.99	90.92	***92.43***
LSTM-LARGE	97.41	93.72	92.67	***93.19***	97.72	93.41	90.56	91.96
GRU-SMALL	96.46	92.10	89.10	90.58	97.36	92.28	88.60	90.40
GRU-LARGE	96.95	92.75	90.93	91.83	97.27	90.86	90.20	90.52

[Kim, Jernite, Sontag, Rush]

Word Representations of Character & Word Models

	In Vocabulary					Out-of-Vocabulary		
	while	*his*	*you*	*richard*	*trading*	*computer-aided*	*misinformed*	*looooook*
LSTM-Word	*although*	*your*	*conservatives*	*jonathan*	*advertised*	–	–	–
	letting	*her*	*we*	*robert*	*advertising*	–	–	–
	though	*my*	*guys*	*neil*	*turnover*	–	–	–
	minute	*their*	*i*	*nancy*	*turnover*	–	–	–
LSTM-Char (before highway)	*chile*	*this*	*your*	*hard*	*heading*	*computer-guided*	*informed*	*look*
	whole	*hhs*	*young*	*rich*	*training*	*computerized*	*performed*	*cook*
	meanwhile	*is*	*four*	*richer*	*reading*	*disk-drive*	*transformed*	*looks*
	white	*has*	*youth*	*richter*	*leading*	*computer*	*inform*	*shook*
LSTM-Char (after highway)	*meanwhile*	*hhs*	*we*	*eduard*	*trade*	*computer-guided*	*informed*	*look*
	whole	*this*	*your*	*gerard*	*training*	*computer-driven*	*performed*	*looks*
	though	*their*	*doug*	*edward*	*traded*	*computerized*	*outperformed*	*looked*
	nevertheless	*your*	*i*	*carl*	*trader*	*computer*	*transformed*	*looking*

Table 6: Nearest neighbor words (based on cosine similarity) of word representations from the large word-level and character-level (before and after highway layers) models trained on the PTB. Last three words are OOV words, and therefore they do not have representations in the word-level model.

[Kim, Jernite, Sontag, Rush]