Generative Modeling with RNNs

Most slides by Zack Lipton

Critical Review of RNNs:
http://arxiv.org/abs/1506.00019

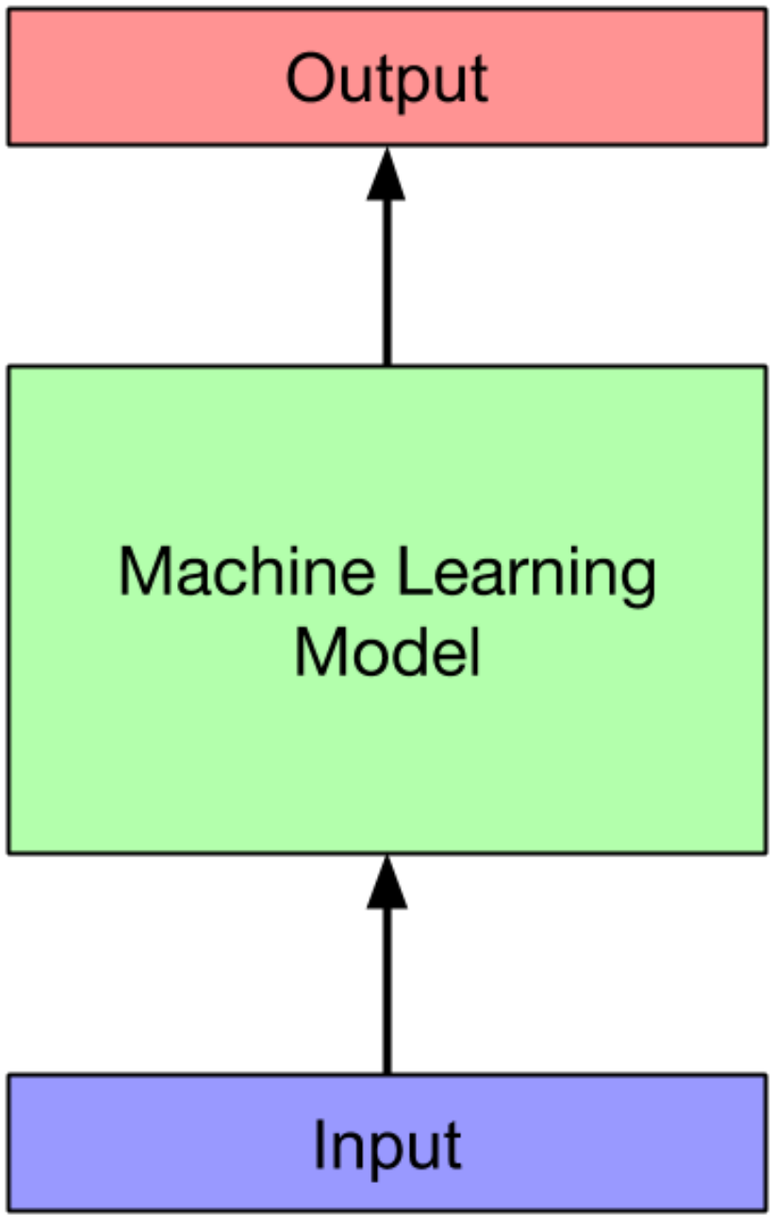
Conditional Generative RNNS:
http://arxiv.org/abs/1511.03683

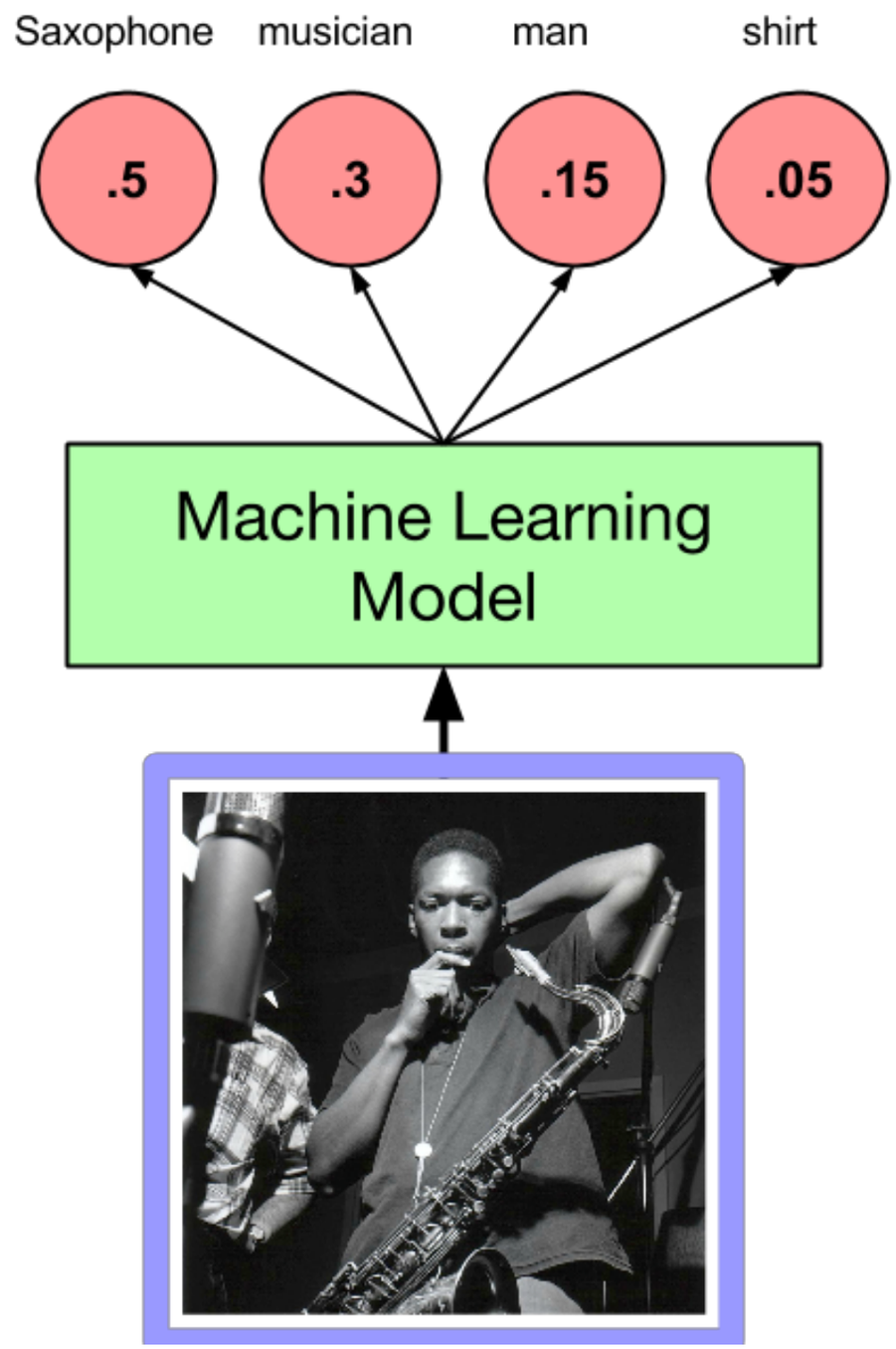
Andrej Karpathy Blogpost on RNNs:
http://karpathy.github.io/2015/05/21/rnn-effectiveness/

Outline

- **Motivation**
- Generative Text Model
- Sequence to Sequence
- Image Captioning
- Supervised Character Model (Beer Reviews)

We Have Great Tools for Fixed-Size Data (vector in, vector out)





We'd Prefer a System that Could output Structured Data

"A group of young people playing a game of Frisbee."
(Karpathy 2014)

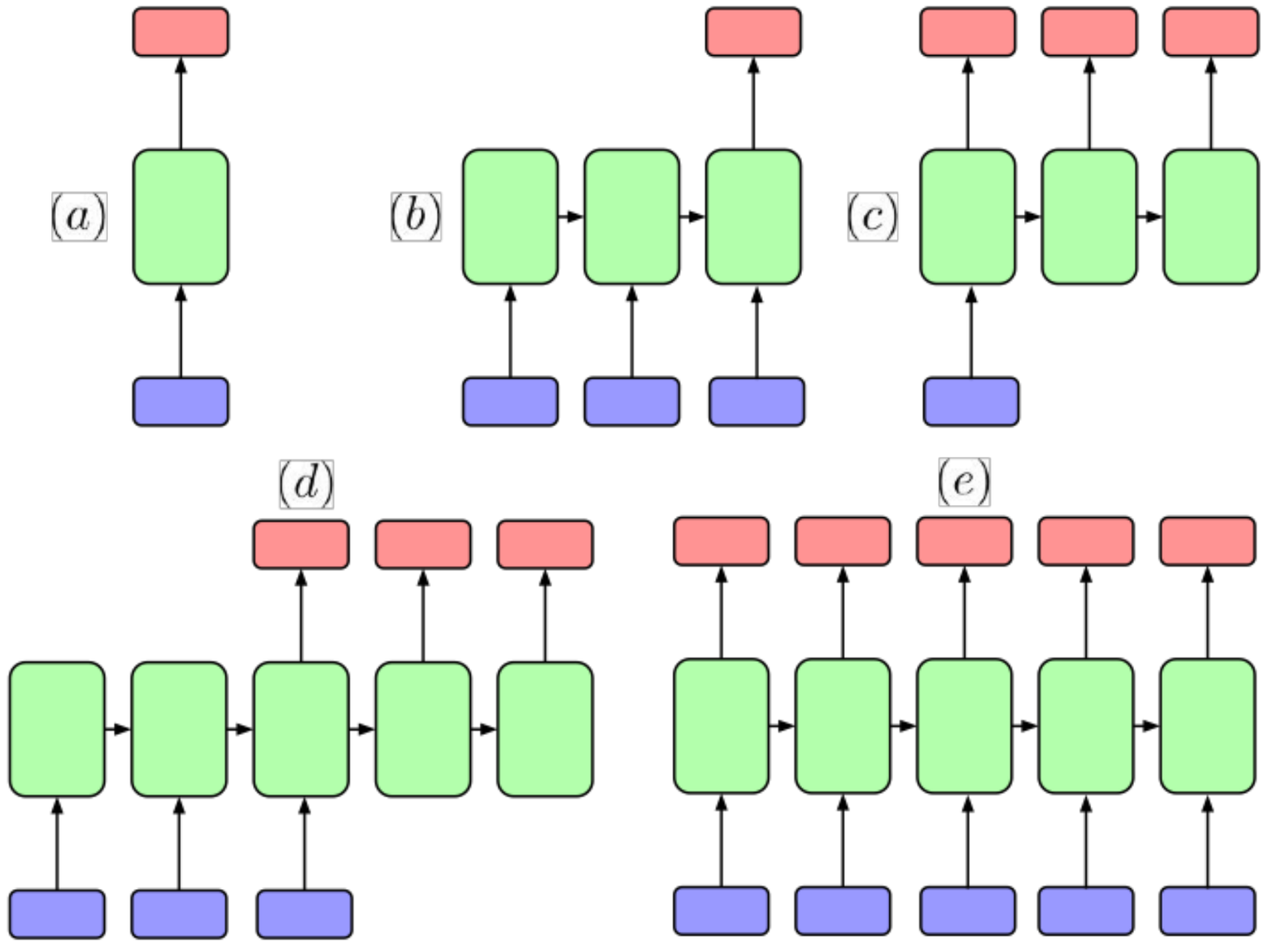
Compose Complex Actions as **Sequences** of Simple Ones

- Sentences might not admit a fixed vector representation but words can

- Break up text input:
  The model can take a word at each step as input

- Break up text output:
  The model can output a word at each time step

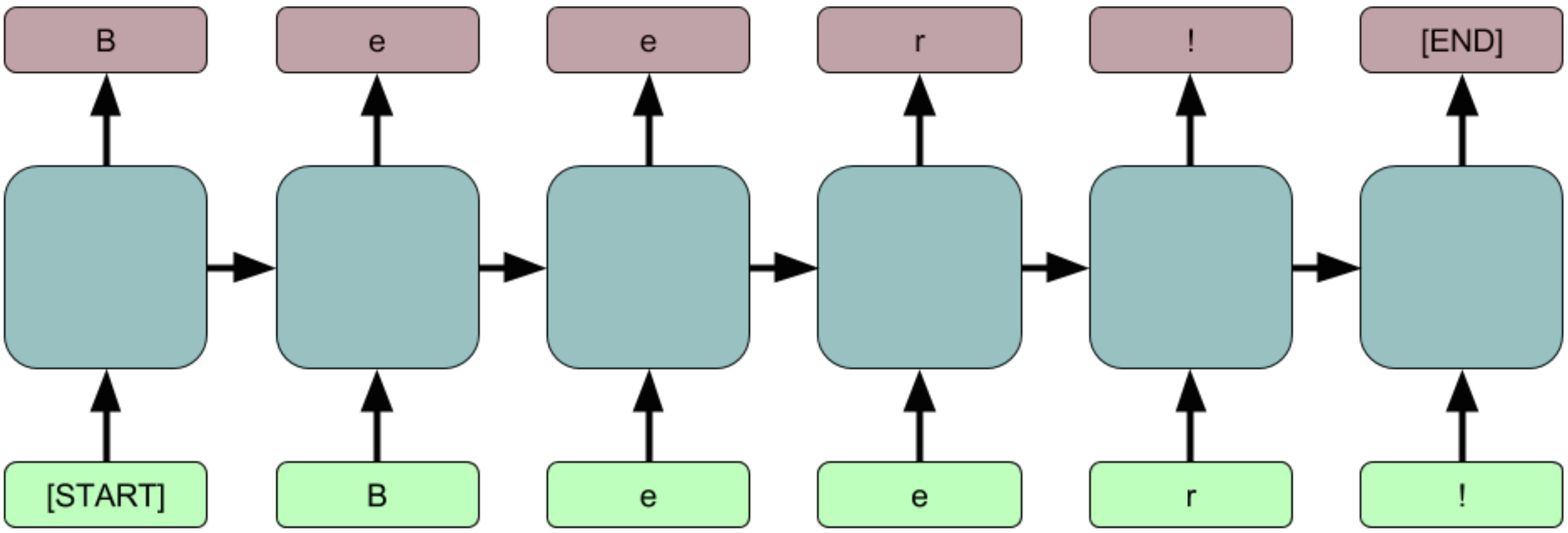
Basic RNN Architectures



Outline

- Motivation
- **Generative Text Model**
- Sequence to Sequence
- Image Captioning
- Supervised Character Model (Beer Reviews)

RNN Generative Model



Shakespeare Generation

PANDARUS:
Alas, I think he shall be come approached and the day
When little srain would be attain'd into being never fed,
And who is but a chain and subjects of his death,
I should not sleep.

Second Senator:
They are away this miseries, produced upon my soul,
Breaking and strongly should be buried, when I perish
The earth and thoughts of many states.

DUKE VINCENTIO:
Well, your wit is in the care of side and that.

Second Lord:
They would be ruled after this chamber, and
my fair nues begun out of the fact, to be conveyed,
Whose noble souls I'll have the heart of the wars.

Clown:
Come, sir, I will make did behold your worship.

VIOLA:
I'll drink it.

Training Data:
All of Shakespeare

Size: 4.4MB

Text generated one character at a time!

Guided Generation

(Sutskever et al., 2011)

Recurrent network with the Stiefel information for logistic regression methods Along with either of the algorithms previously (two or more skew precision) is more similar to the model with the same average mismatched graph. Though this task is to be studied under the reward transform, such as (c) and (C) from the training set, based on target activities for articles a ? 2(6) and (4.3). The PHDPic (PDB) matrix of cav'va using the three relevant information contains for tieming measurements. Moreover, because of the therap tor, the aim is to improve the score to the best patch randomly, but for each initially four data sets. As shown in Figure 11, it is more than 100 steps, we used ?? \to \infty with 1000

Training Data: Machine Learning papers (NIPS and JMLR)

Initialized with "Recurrent"

Again, text generated one letter at a time!

Guided Generation

(Sutskever et al., 2011)

The meaning of life is the tradition of the ancient human reproduction: it is less favorable to the good boy for when to remove her bigger. In the show's agreement unanimously resurfaced. The wild pasteured with consistent street forests were incorporated by the 15th century BE. In 1996 the primary rapford undergoes an effort that the reserve conditioning, written into Jewish cities, sleepers to incorporate the .St Eurasia that activates the population. Mar??a Nationale, Kelli, Zedlat-Dukastoe, Florendon, Ptu's thought is. To adapt in most parts of North America, the dynamic fairy Dan please believes, the free speech are much related to the

Training Data: Wikipedia

Initialized with the bold phrase.

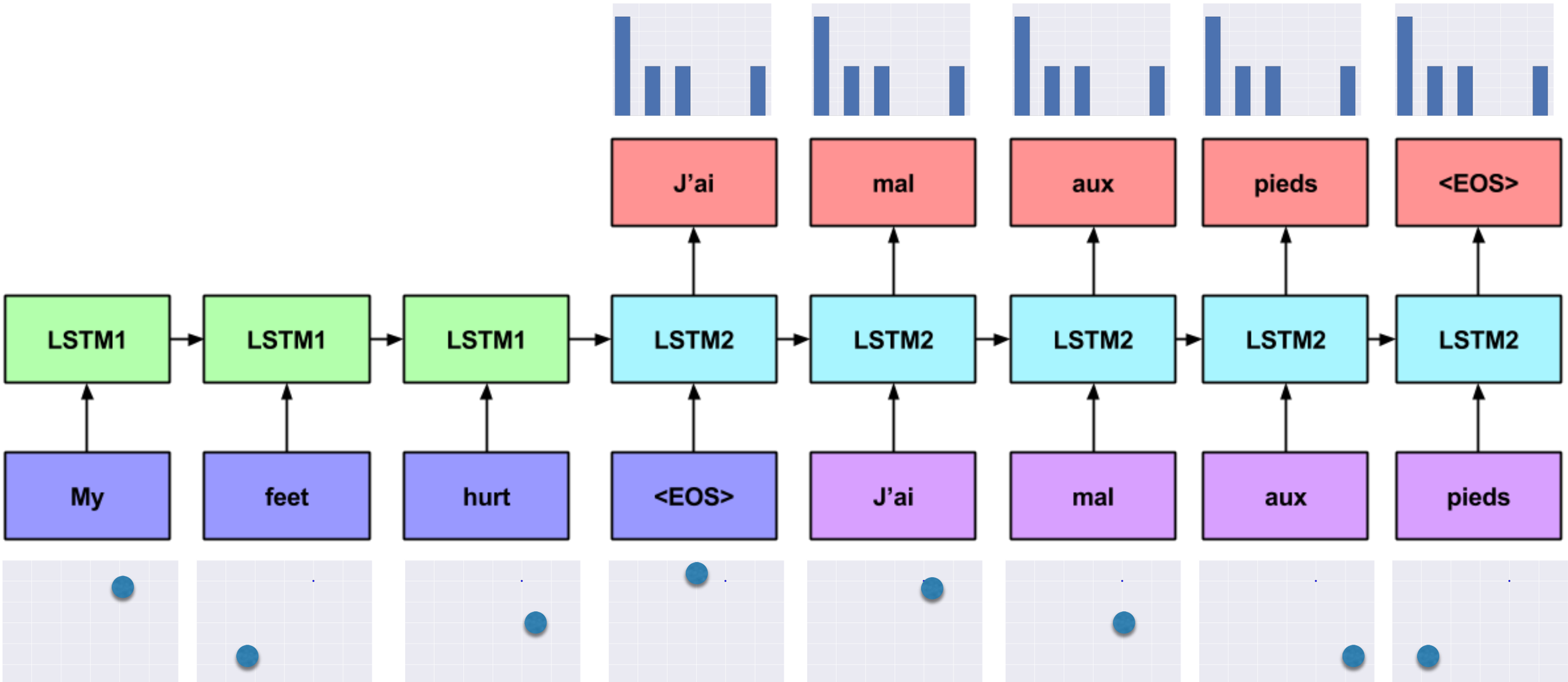
Again, text generated one character at a time!

Outline

- Motivation
- Generative Text Model
- **Sequence to Sequence**
- Image Captioning
- Supervised Character Model (Beer Reviews)

Sequence to Sequence

(Sutskever et al. 2014)



Scale

- Sutskever et al. trained Deep LSTMs with four layers.
- Each layer had 1000 cells
- 384M parameters
- Input Vocabulary 160,000 words
- Output Vocabulary: 80,000 words
- Training Algo: Vanilla SGD, attenuated learning rate
- Training time: 10 days with 8 GPUS (4 for Softmax)

Results

- Achieved BLEU (measure of translation quality) comparable to best state of the art systems

- Hybrid approaches and ensembling LSTMs led to scores even better than state of the art systems

- No information about language was explicitly modeled or hardwired (besides the vocabulary itself)

Outline

- Motivation
- Generative Text Model
- Sequence to Sequence
- **Image Captioning**
- Supervised Character Model (Beer Reviews)

Image to Text

(Karpathy et al. 2014), (Mao et al., 2014),
(Vinyals et al., 2014)

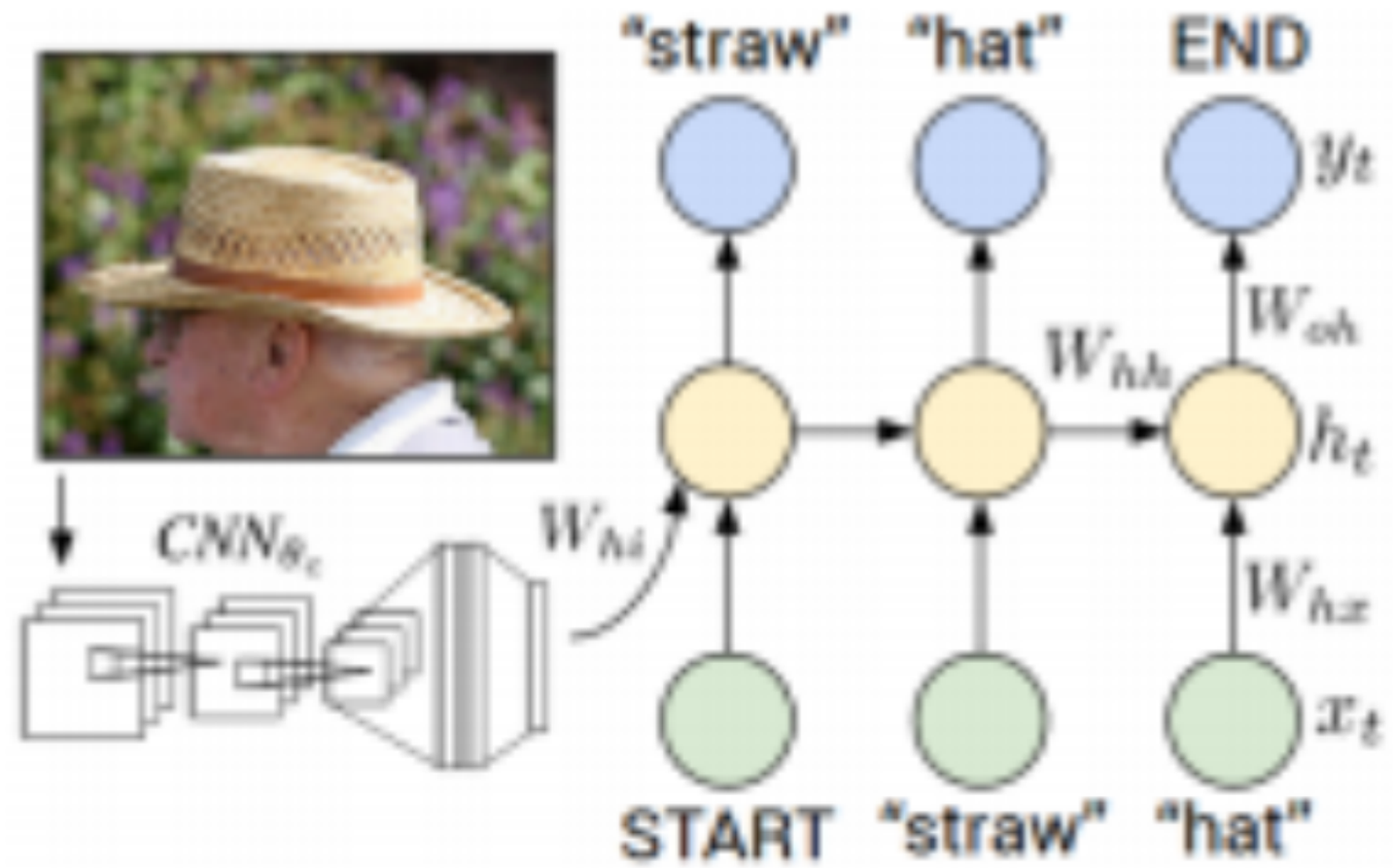


Image Captioning

(slide from Karpathy et al. 2014)

"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."

"boy is doing backflip on wakeboard."

"girl in pink dress is jumping in air."

"black and white dog jumps over bar."

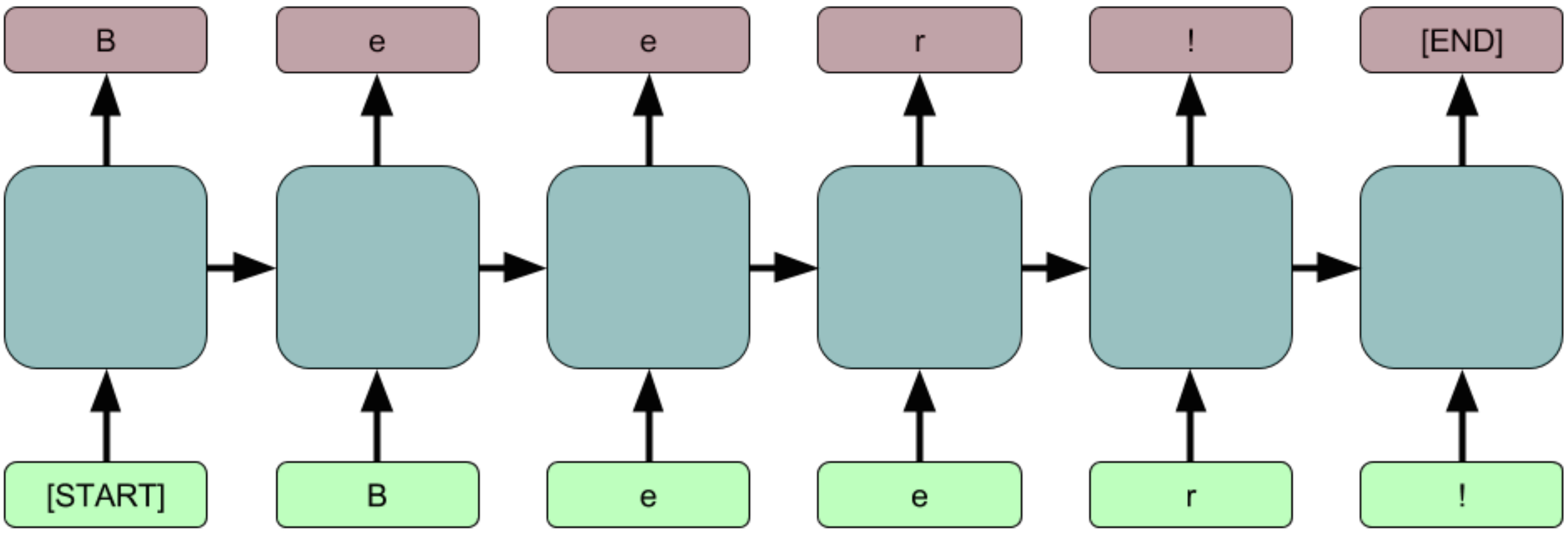
"young girl in pink shirt is swinging on swing."

"man in blue wetsuit is surfing on wave."

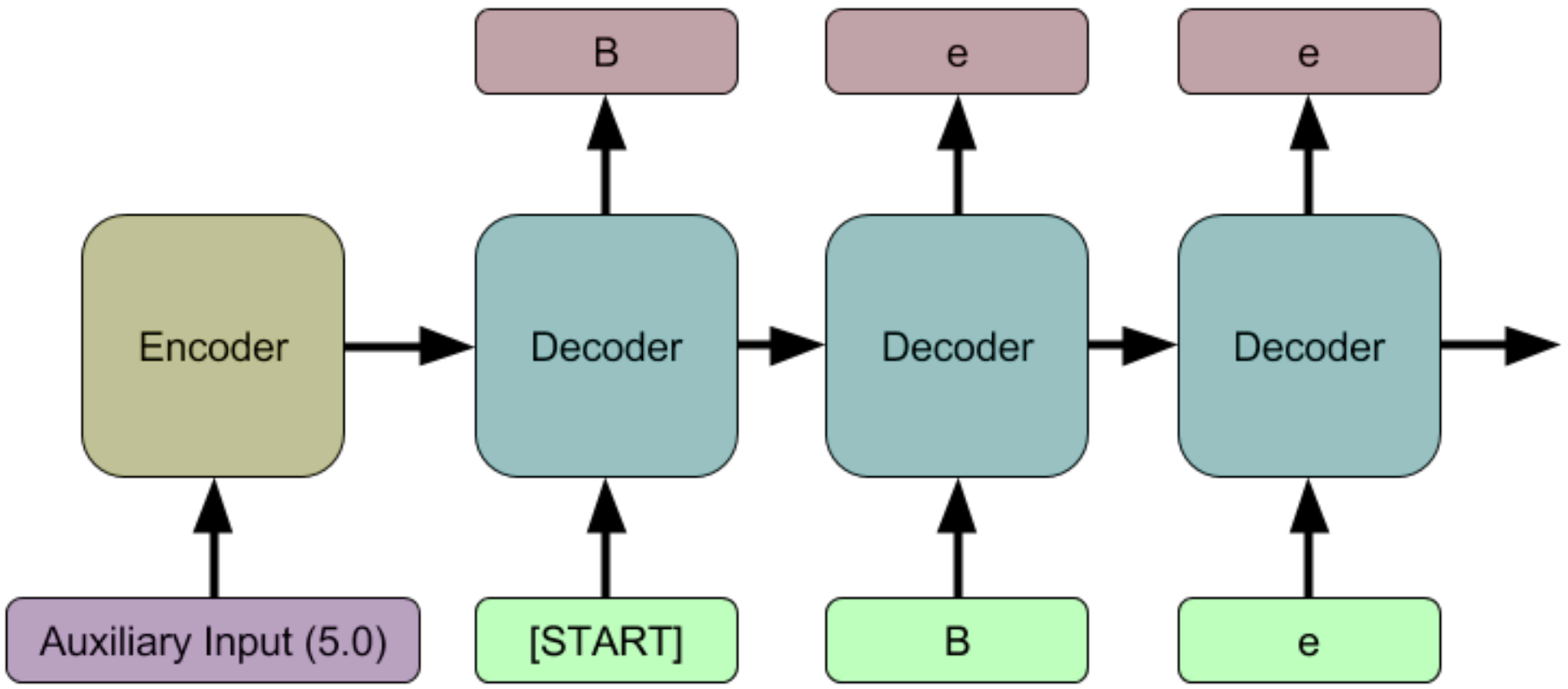
Outline

- Motivation
- Generative Text Model
- Sequence to Sequence
- Image Captioning
- **Supervised Character Model (Beer Review Demo)**

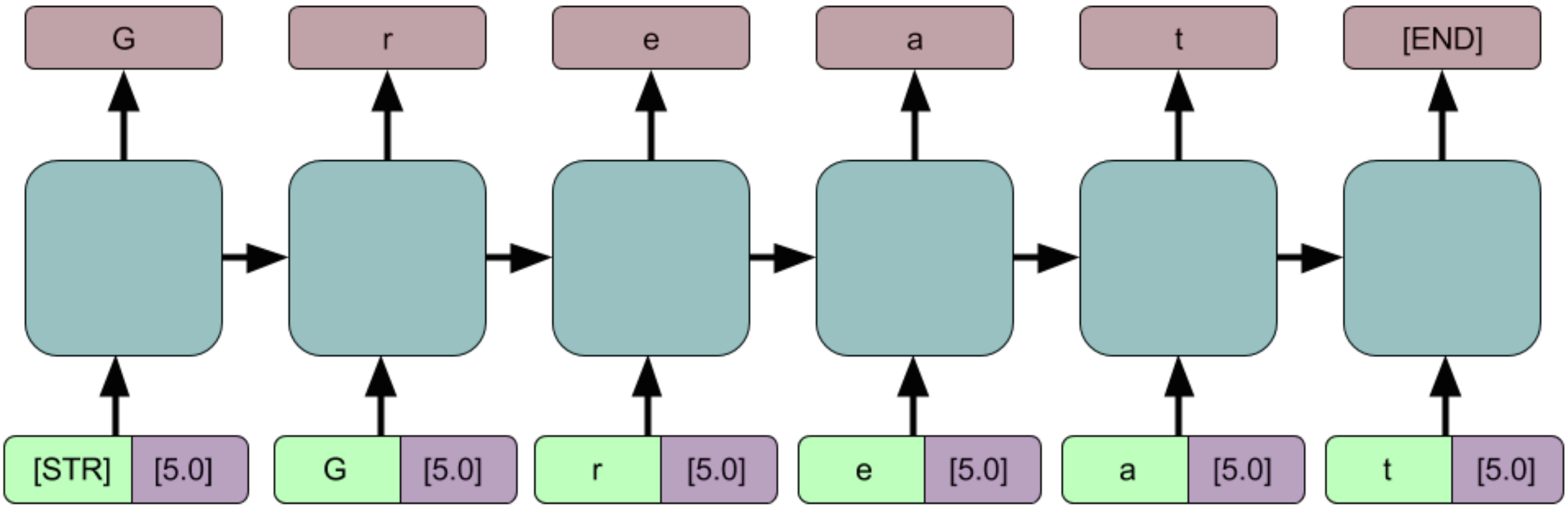
Recall Unsupervised Character Model



Past Supervised Approaches relied upon Encoder-Decoder Model



Bridging Long Time Intervals with Concatenated Inputs



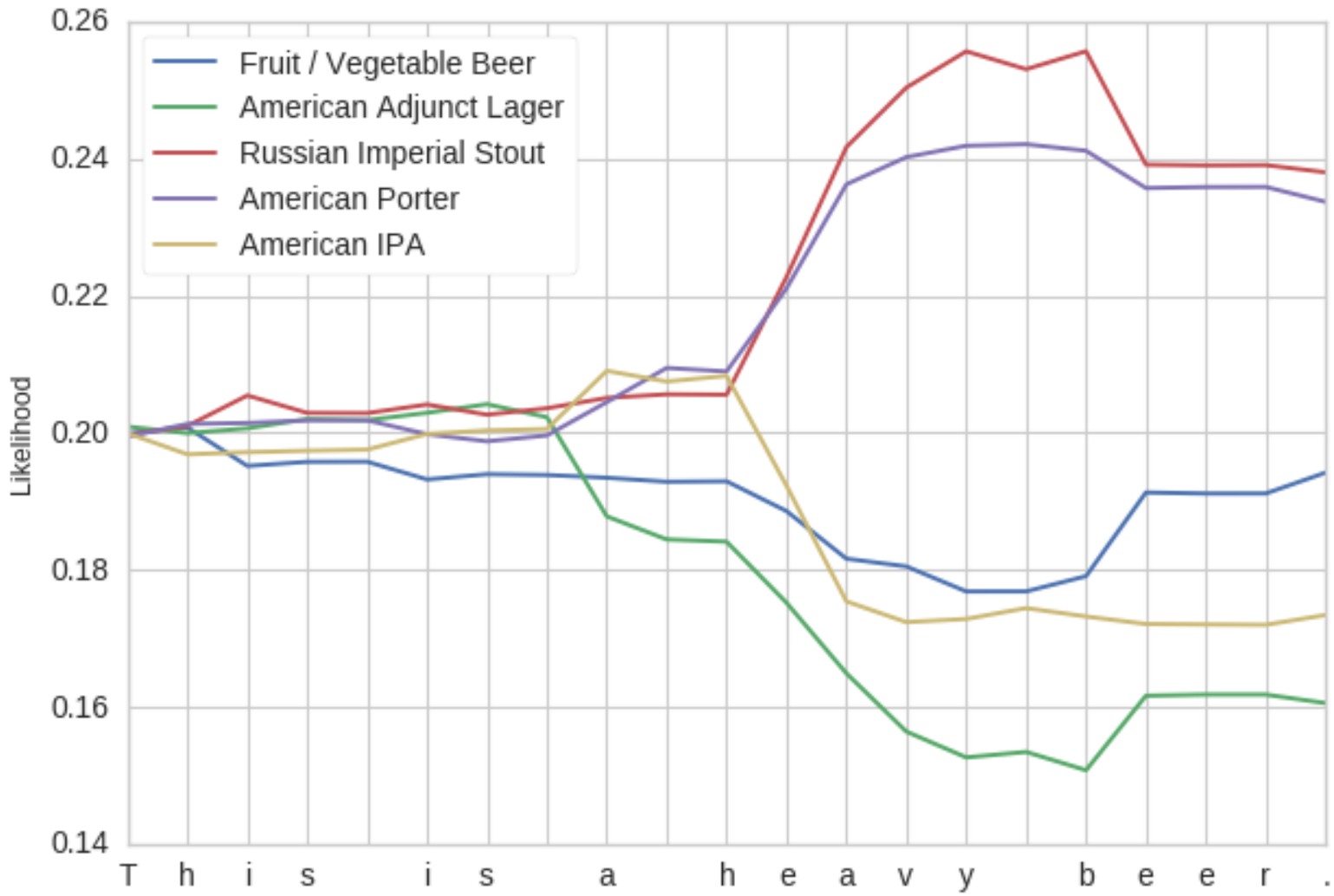
Example

A.5 FRUIT/VEGETABLE BEER:

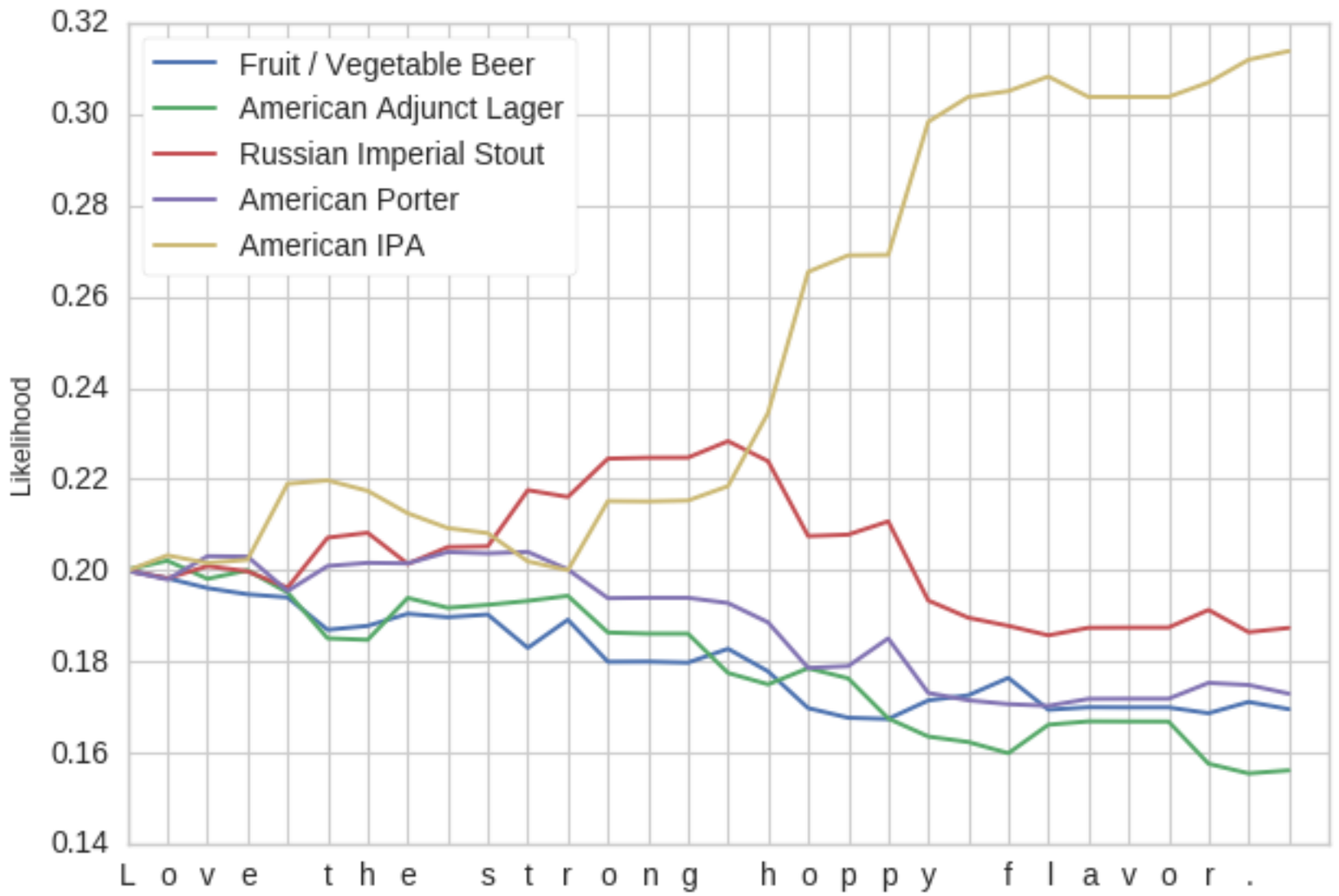
<STR>On tap at the brewpub. A nice dark red color with a nice head that left a lot of lace on the glass. Aroma is of raspberries and chocolate. Not much depth to speak of despite consisting of raspberries. The bourbon is pretty subtle as well. I really don't know that I find a flavor this beer tastes like. I would prefer a little more carbonization to come through. It's pretty drinkable, but I wouldn't mind if this beer was available. <EOS>

IF the above link doesn't work, cut and paste:
http://deepx.ucsd.edu/#/home/beermind

Character-based Classification



“Love the Strong Hoppy Flavor”

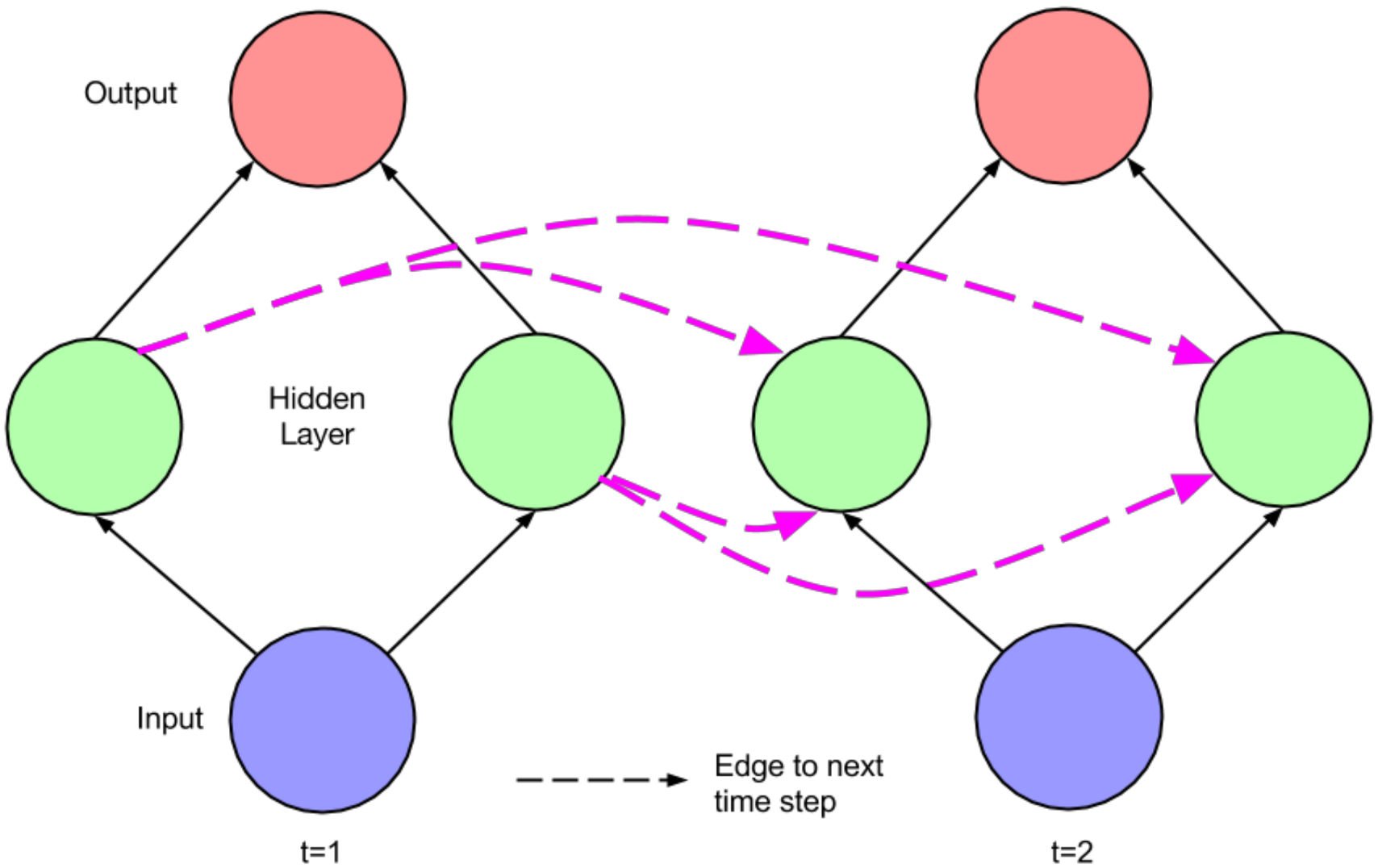


Summary

- We can use recurrent networks as *generators of data.*
- A standard way to do this is to train it on a large corpus, and then let it “hallucinate” text by sampling from its output, and using that as the next input
- We can use recurrent networks to transform sequences, leading to state of the art translation models
- We can “bolt together” a convnet and a language model:
  - the convnet gets a representation of an image,
  - the language model generates the image caption.
- Using biasing input, we can exert some control over the generation: Beer Reviews

Supplementary Slides (LSTM Basics)

Recurrent Net (Unfolded)

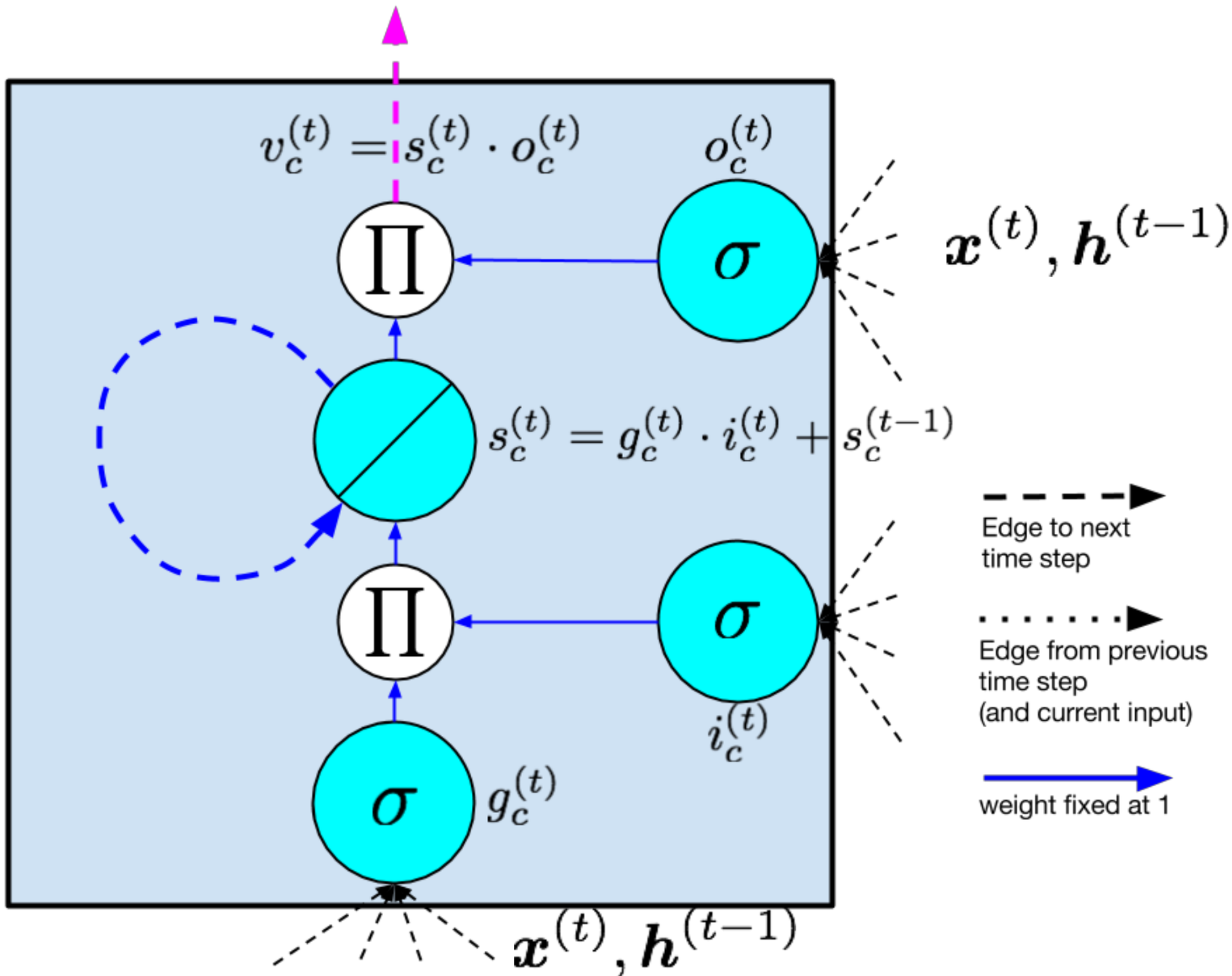


$$h^{(t)} = \sigma(W_{hx}\boldsymbol{x}^{(t)} + W_{hh}\boldsymbol{h}^{(t-1)} + \boldsymbol{b}_h)$$

$$\hat{\boldsymbol{y}}^{(t)} = \text{softmax}(W_{yh}\boldsymbol{h}^{(t)} + \boldsymbol{b}_y)$$

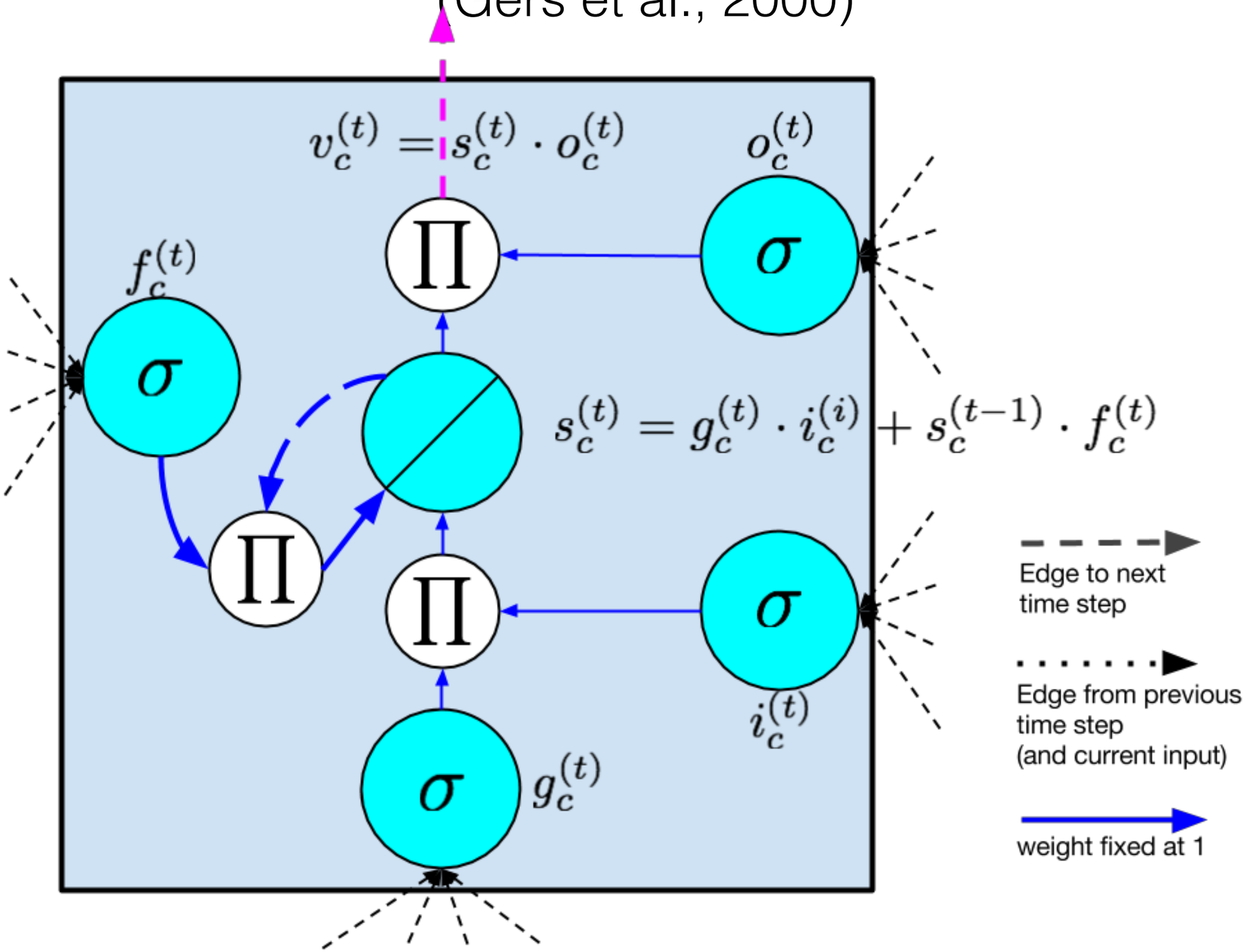
LSTM Memory Cell

(Hochreiter & Schmidhuber, 1997)



Memory Cell with Forget Gate

(Gers et al., 2000)



LSTM Forward Pass

$$\boldsymbol{g}^{(t)} = \phi(W_{gx}\boldsymbol{x}^{(t)} + W_{ih}\boldsymbol{h}^{(t-1)} + \boldsymbol{b}_g)$$

$$\boldsymbol{i}^{(t)} = \sigma(W_{ix}\boldsymbol{x}^{(t)} + W_{ih}\boldsymbol{h}^{(t-1)} + \boldsymbol{b}_i)$$

$$\boldsymbol{f}^{(t)} = \sigma(W_{fx}\boldsymbol{x}^{(t)} + W_{fh}\boldsymbol{h}^{(t-1)} + \boldsymbol{b}_f)$$

$$\boldsymbol{o}^{(t)} = \sigma(W_{ox}\boldsymbol{x}^{(t)} + W_{oh}\boldsymbol{h}^{(t-1)} + \boldsymbol{b}_o)$$

$$\boldsymbol{s}^{(t)} = \boldsymbol{g}^{(t)} \odot \boldsymbol{i}^{(i)} + \boldsymbol{s}^{(t-1)} \odot \boldsymbol{f}^{(t)}$$

$$\boldsymbol{h}^{(t)} = \boldsymbol{s}^{(t)} \odot \boldsymbol{o}^{(t)}$$

LSTM (full network)