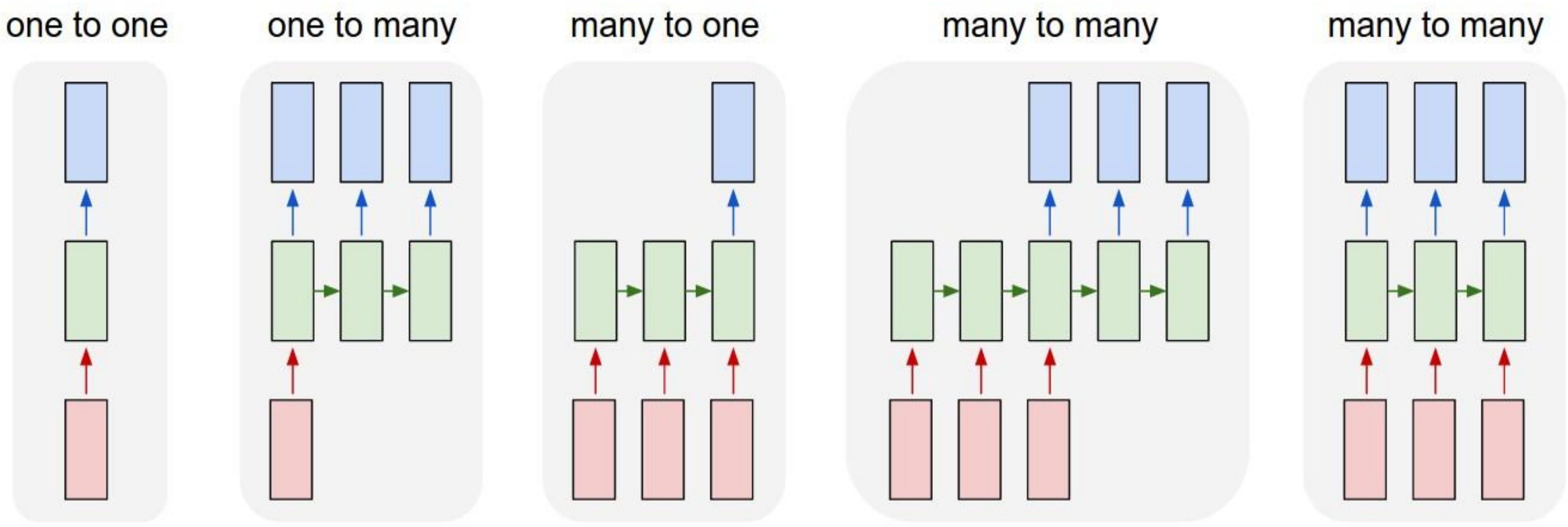
11-785: Recitation 8
RNN Basics

Ojas Bhargave, Clay Yoo

Sequential Data

1. Predicting the next word
    a. “Never gonna give you up. Never gonna let you ____.”
2. Filling in missing word
    a. “Hey, I just met you and this is ____, but here’s my number. So call me, maybe.”
3. Vocab identification from speech
    a. “Heh-low-wer-l-d” -> “Hello World"
4. Machine translation
    a. “Hello World” -> “안녕하세요 세계"
5. Many other tasks…
    a. Speech transcription
    b. Text classification
    c. Text generation
    d. Stock price movement prediction

Data Types and Modeling



(https://i.stack.imgur.com/b4sus.jpg)

RNN Example: Math Proof Generation in Latex

For $\bigoplus_{n=1,\ldots,m}$ where $\mathcal{L}_{m_\bullet} = 0$, hence we can find a closed subset $\mathcal{H}$ in $\mathcal{H}$ and any sets $\mathcal{F}$ on X, U is a closed immersion of S, then $U \to T$ is a separated algebraic space.

Proof. Proof of (1). It also start we get

$$S = \mathrm{Spec}(R) = U \times_X U \times_X U$$

and the comparicoly in the fibre product covering we have to prove the lemma generated by $\coprod Z \times_U U \to V$. Consider the maps M along the set of points Sch_{fppf} and $U \to U$ is the fibre category of S in U in Section, **??** and the fact that any U affine, see Morphisms, Lemma **??**. Hence we obtain a scheme S and any open subset $W \subset U$ in $Sh(G)$ such that $\mathrm{Spec}(R') \to S$ is smooth or an

$$U = \bigcup U_i \times_{S_i} U_i$$

which has a nonzero morphism we may assume that f_i is of finite presentation over S. We claim that $\mathcal{O}_{X,x}$ is a scheme where $x, x', s'' \in S'$ such that $\mathcal{O}_{X,x'} \to \mathcal{O}'_{X',x'}$ is separated. By Algebra, Lemma **??** we can define a map of complexes $\mathrm{GL}_{S'}(x'/S'')$ and we win. □

To prove study we see that $\mathcal{F}|_U$ is a covering of $\mathcal{X}'$, and $\mathcal{T}_i$ is an object of $\mathcal{F}_{X/S}$ for $i > 0$ and $\mathcal{F}_p$ exists and let $\mathcal{F}_i$ be a presheaf of $\mathcal{O}_X$-modules on $\mathcal{C}$ as a $\mathcal{F}$-module. In particular $\mathcal{F} = U/\mathcal{F}$ we have to show that

$$\widetilde{M}^\bullet = \mathcal{I}^\bullet \otimes_{\mathrm{Spec}(k)} \mathcal{O}_{S,s} - i_X^{-1}\mathcal{F})$$

is a unique morphism of algebraic stacks. Note that

$$\mathrm{Arrows} = (Sch/S)^{opp}_{fppf}, (Sch/S)_{fppf}$$

and

$$V = \Gamma(S, \mathcal{O}) \longmapsto (U, \mathrm{Spec}(A))$$

is an open subset of X. Thus U is affine. This is a continuous map of X is the inverse, the groupoid scheme S.

Proof. See discussion of sheaves of sets. □

The result for prove any open covering follows from the less of Example **??**. It may replace S by $X_{spaces,\acute{e}tale}$ which gives an open subspace of X and T equal to S_{Zar}, see Descent, Lemma **??**. Namely, by Lemma **??** we see that R is geometrically regular over S.

Lemma 0.1. *Assume (3) and (3) by the construction in the description.*

Suppose $X = \lim |X|$ (by the formal open covering X and a single map $\underline{Proj}_X(\mathcal{A}) =$ $\mathrm{Spec}(B)$ over U compatible with the complex

$$Set(\mathcal{A}) = \Gamma(X, \mathcal{O}_{X,\mathcal{O}_X}).$$

When in this case of to show that $\mathcal{Q} \to \mathcal{C}_{Z/X}$ is stable under the following result in the second conditions of (1), and (3). This finishes the proof. By Definition **??** *(without element is when the closed subschemes are catenary. If T is surjective we may assume that T is connected with residue fields of S. Moreover there exists a closed subspace $Z \subset X$ of X where U in X' is proper (some defining as a closed subset of the uniqueness it suffices to check the fact that the following theorem*

(1) *f is locally of finite type. Since $S = \mathrm{Spec}(R)$ and $Y = \mathrm{Spec}(R)$.*

Proof. This is form all sheaves of sheaves on X. But given a scheme U and a surjective étale morphism $U \to X$. Let $U \cap U = \coprod_{i=1,\ldots,n} U_i$ be the scheme X over S at the schemes $X_i \to X$ and $U = \lim_i X_i$. □

The following lemma surjective restrocomposes of this implies that $\mathcal{F}_{x_0} = \mathcal{F}_{x_0} = \mathcal{F}_{\mathcal{X},\ldots,0}$.

Lemma 0.2. *Let X be a locally Noetherian scheme over S, $E = \mathcal{F}_{X/S}$. Set $\mathcal{I} = \mathcal{J}_1 \subset \mathcal{I}'_n$. Since $\mathcal{I}^n \subset \mathcal{I}^n$ are nonzero over $i_0 \leq \mathfrak{p}$ is a subset of $\mathcal{J}_{n,0} \circ \overline{A}_2$ works.*

Lemma 0.3. *In Situation* **??**. *Hence we may assume $\mathfrak{q}' = 0$.*

Proof. We will use the property we see that $\mathfrak{p}$ is the mext functor (**??**). On the other hand, by Lemma **??** we see that

$$D(\mathcal{O}_{X'}) = \mathcal{O}_X(D)$$

where K is an F-algebra where δ_{n+1} is a scheme over S. □

(https://karpathy.github.io/2015/05/21/rnn-effectiveness/)

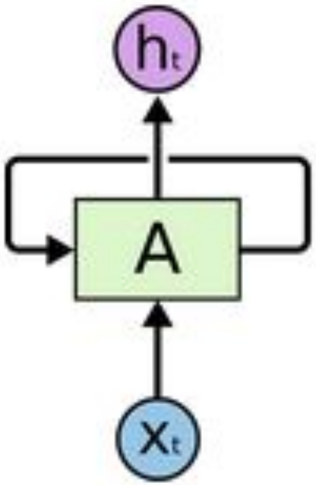
RNN Example: Linux Source Code Generation

```
/*
 * Increment the size file of the new incorrect UI_FILTER group information
 * of the size generatively.
 */
static int indicate_policy(void)
{
  int error;
  if (fd == MARN_EPT) {
    /*
     * The kernel blank will coeld it to userspace.
     */
    if (ss->segment < mem_total)
      unblock_graph_and_set_blocked();
    else
      ret = 1;
    goto bail;
  }
  segaddr = in_SB(in.addr);
  selector = seg / 16;
  setup_works = true;
  for (i = 0; i < blocks; i++) {
    seq = buf[i++];
    bpf = bd->bd.next + i * search;
    if (fd) {
      current = blocked;
    }
  }
  rw->name = "Getjbbregs";
  bprm_self_clearl(&iv->version);
  regs->new = blocks[(BPF_STATS << info->historidac)] | PFMR_CLOBATHINC_SECONDS << 12;
  return segtable;
}
```

(https://karpathy.github.io/2015/05/21/rnn-effectiveness/)

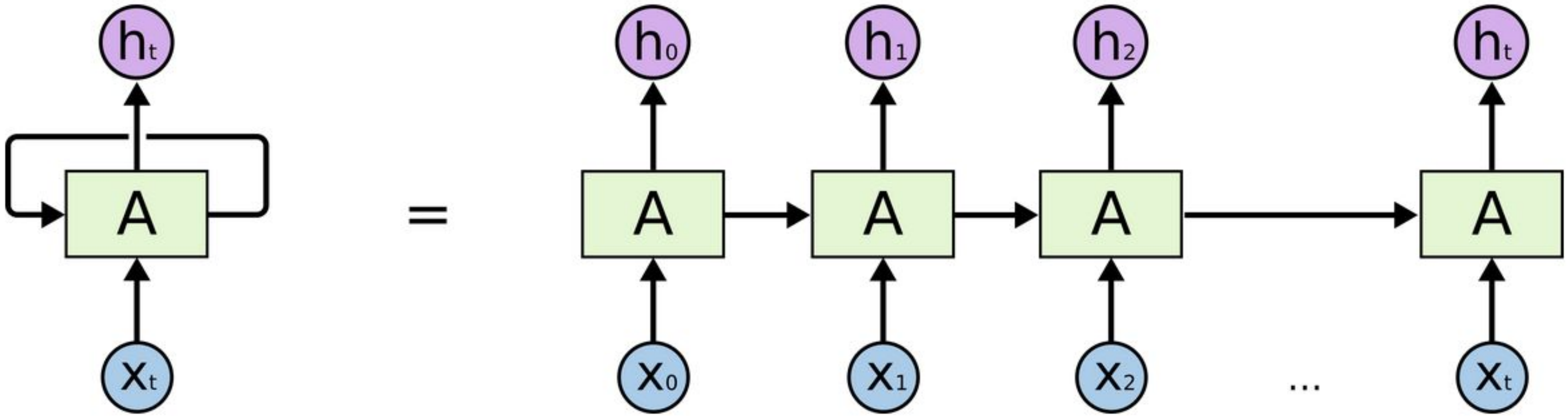
Recurrent Neural Networks

- Looping network
- Parameter sharing across timesteps
- Derivatives aggregated across all time steps
- “Backpropagation through time (BPTT)”



(http://colah.github.io/posts/2015-08-Understanding-LSTMs/)

RNN Unrolled

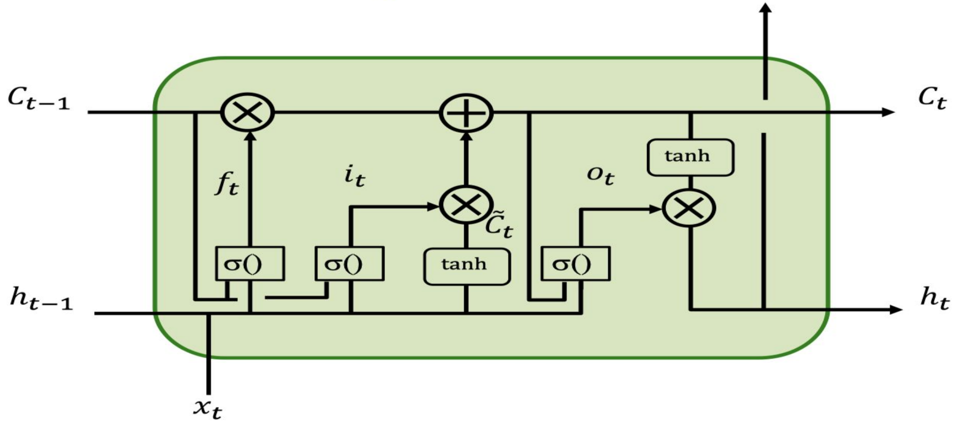


An unrolled recurrent neural network.

(http://colah.github.io/posts/2015-08-Understanding-LSTMs/)

LSTM Cell from Lecture

LSTM computation: Forward



- Forward rules:

Gates

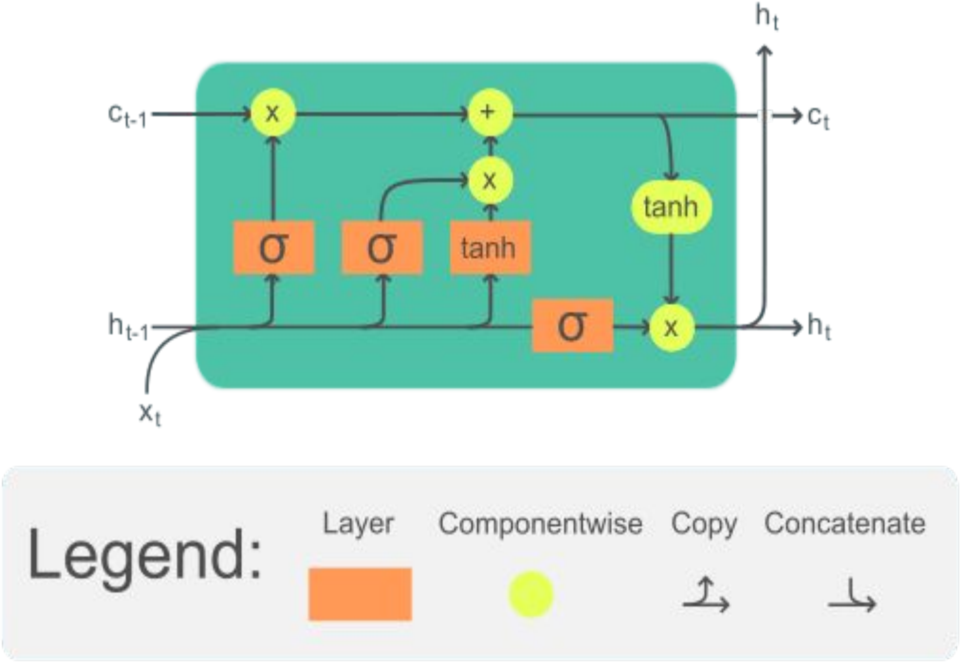
$$f_t = \sigma\left(W_f \cdot [\boldsymbol{C_{t-1}}, h_{t-1}, x_t] \; + \; b_f\right)$$
$$i_t = \sigma\left(W_i \cdot [\boldsymbol{C_{t-1}}, h_{t-1}, x_t] \; + \; b_i\right)$$
$$o_t = \sigma\left(W_o \cdot [\boldsymbol{C_t}, h_{t-1}, x_t] \; + \; b_o\right)$$

Variables

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$
$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$
$$h_t = o_t * \tanh\left(C_t\right)$$

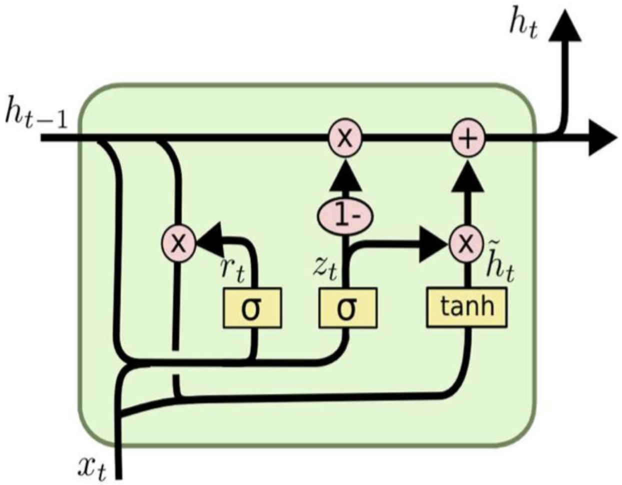
(Gers and Schmidhuber 2000: *Recurrent Nets that Time and Count*)

LSTM Cell from Wikipedia



$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f)$$
$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i)$$
$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o)$$
$$\tilde{c}_t = \sigma_c(W_c x_t + U_c h_{t-1} + b_c)$$
$$c_t = f_t \circ c_{t-1} + i_t \circ \tilde{c}_t$$
$$h_t = o_t \circ \sigma_h(c_t)$$

GRU Cell



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

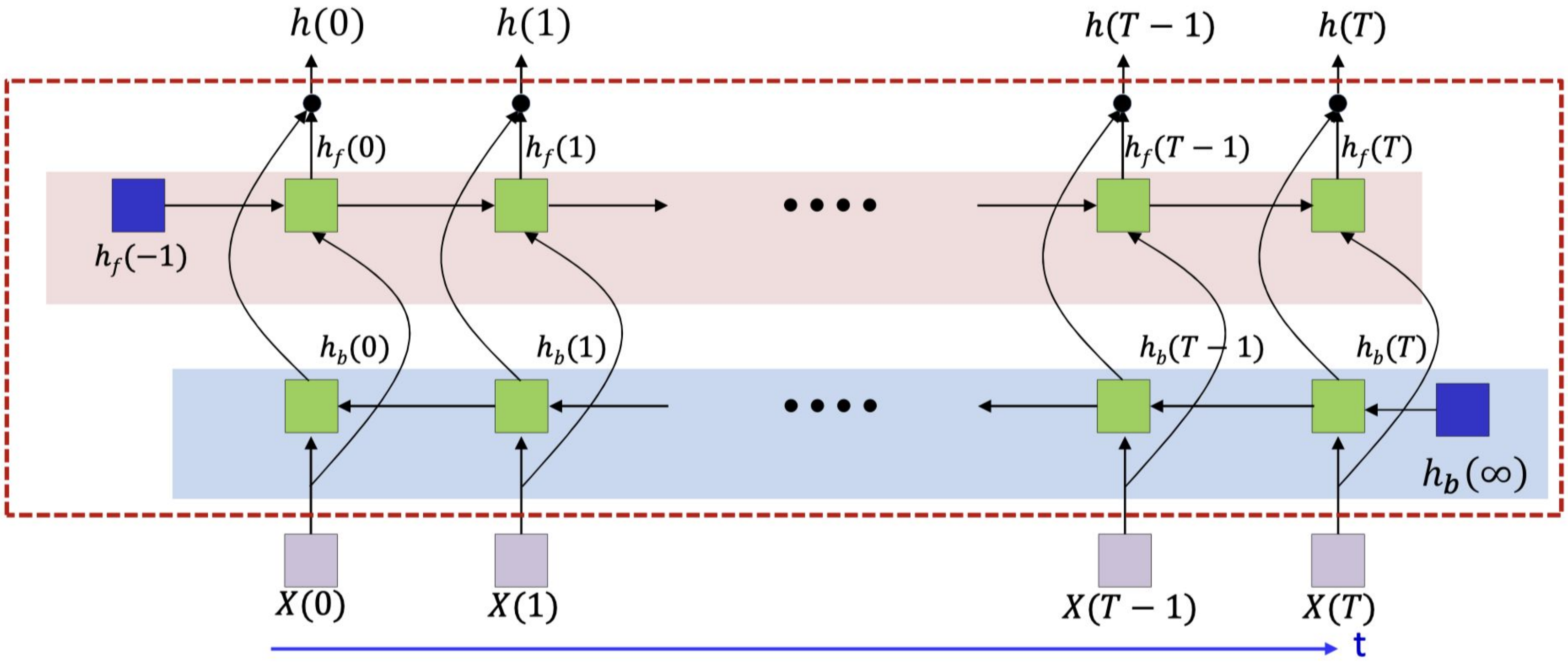
$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

GRUs can't count! (Weiss et al. 2018: *On the Practical Computational Power of Finite Precision RNNs for Language Recognition*)

Bidirectional RNN



Caution in PyTorch Implementation

```python
import torch

lstm = torch.nn.LSTM(input_size = 1, hidden_size = 4, num_layers = 1)
for name, param in lstm.named_parameters():
    print(name, param.shape)
```

```
weight_ih_l0 torch.Size([16, 1])
weight_hh_l0 torch.Size([16, 4])
bias_ih_l0 torch.Size([16])
bias_hh_l0 torch.Size([16])
```

Questions:

1. What are weight_ih and weight_hh?
2. How to interpret the dimensions?
3. Which version of LSTM is this?
4. How should you use initialization (e.g. Xavier, Kaiming)?

Caution in PyTorch Implementation

```python
import torch

lstm = torch.nn.LSTM(input_size = 1, hidden_size = 4, num_layers = 1)
for name, param in lstm.named_parameters():
    print(name, param.shape)
```

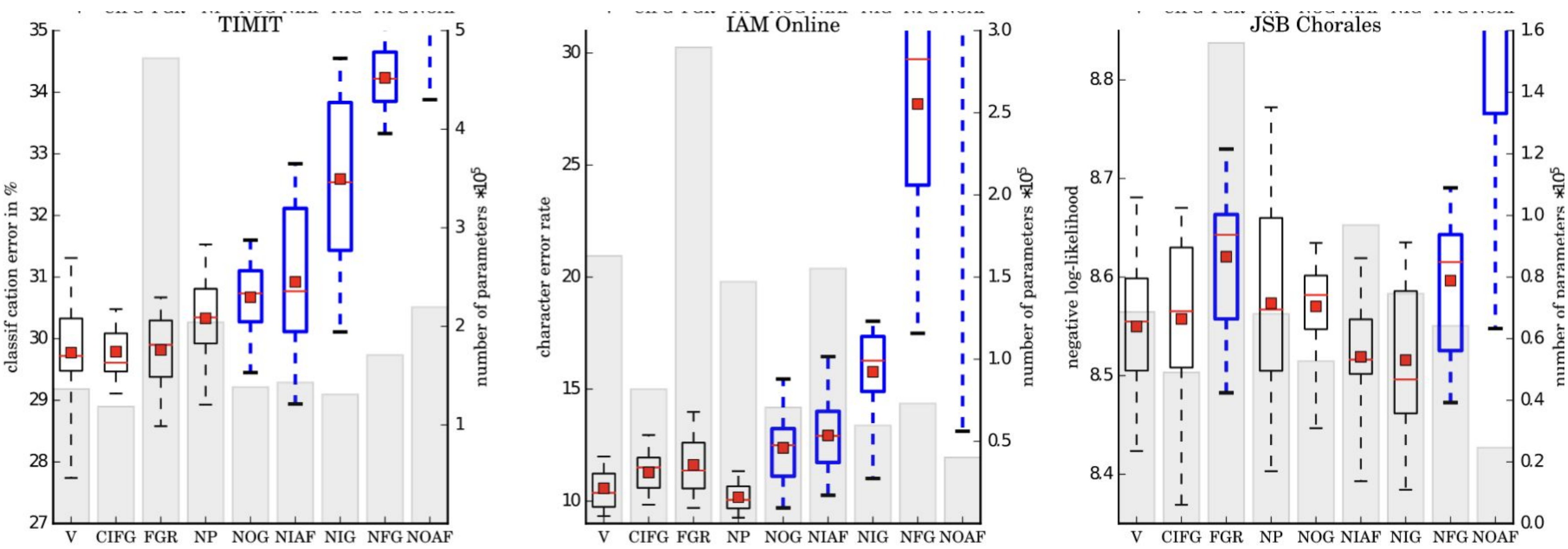
```
weight_ih_l0 torch.Size([16, 1])
weight_hh_l0 torch.Size([16, 4])
bias_ih_l0 torch.Size([16])
bias_hh_l0 torch.Size([16])
```

Questions:

1. What are weight_ih and weight_hh? Input weights and hidden weights
2. How to interpret the dimensions? Input, forget, cell, and output weights stacked (reference)
3. Which version of LSTM is this? Wikipedia version (no peephole connection)
4. How should you use initialization (e.g. Xavier, Kaiming)? For any initialization using fan_out, we initialize each one of four (three if GRU) matrices separately

Performance per LSTM Component

(Greff et al. 2017: *LSTM: A Search Space Odyssey*)



CIFG: GRU, NP: No peepholes, FGR: Full gate recurrence, NOG: No output gate, NIG: No input gate, **NFG: No forget gate**, NIAF: No input activation function, **NOAF: No output activation function**)

Performance per LSTM Component

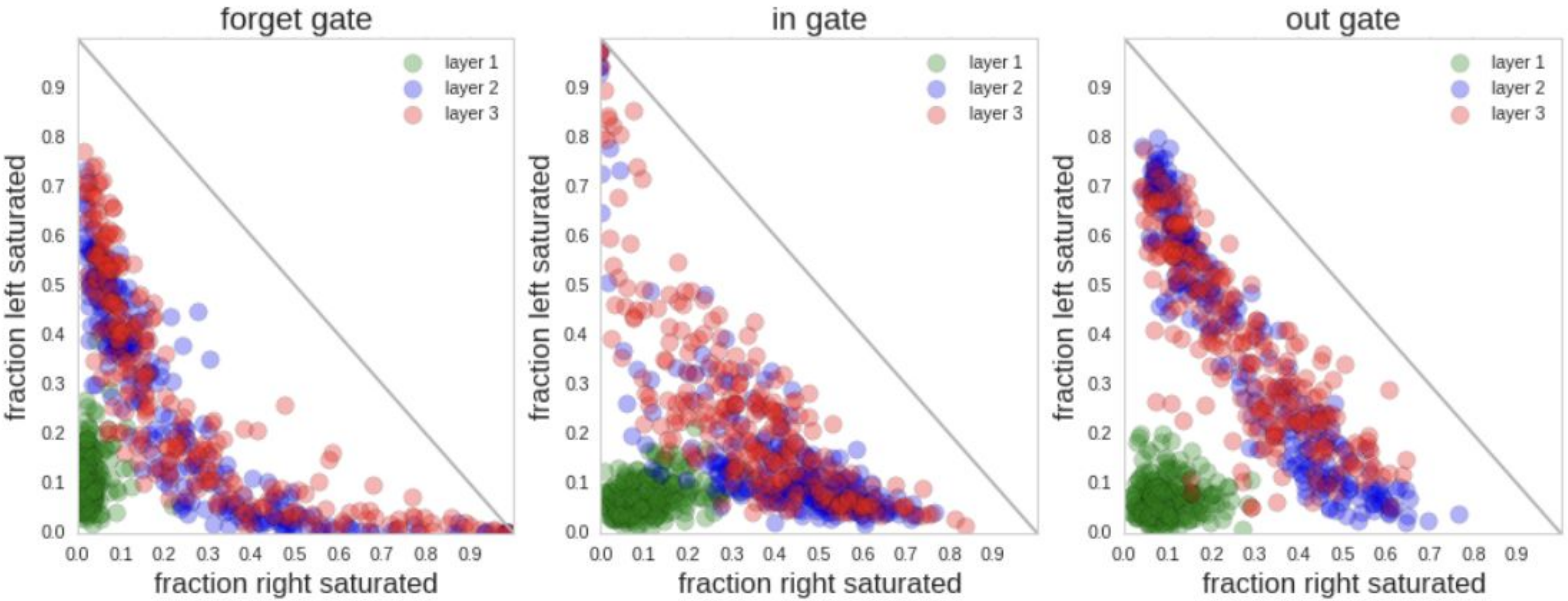
Arch.	Arith.	XML	PTB
Tanh	0.29493	0.32050	0.08782
LSTM	0.89228	0.42470	0.08912
LSTM-f	0.29292	0.23356	0.08808
LSTM-i	0.75109	0.41371	0.08662
LSTM-o	0.86747	0.42117	0.08933
LSTM-b	0.90163	0.44434	0.08952
GRU	0.89565	0.45963	0.09069
MUT1	**0.92135**	**0.47483**	0.08968
MUT2	0.89735	**0.47324**	0.09036
MUT3	0.90728	0.46478	**0.09161**

(Jozefowicz et al. 2015: *An Empirical Exploration of Recurrent Network Architectures*)

Interpretability of LSTM Cells

(Karpathy et al. 2015: *Visualizing and Understanding Recurrent Networks*)

Interpretability of LSTM Cells



(Karpathy et al. 2015: *Visualizing and Understanding Recurrent Networks*)

Interpretability of LSTM Cells

Once in a while you get amazed over how BAD a film can be, and how in the world anybody could raise money to make this kind of crap. There is absolutely No talent included in this film - from a crappy script, to a crappy story to crappy acting. Amazing...

Team Spirit is maybe made by the best intentions, but it misses the warmth of "All Stars" (1997) by Jean van de Velde. Most scenes are identic, just not that funny and not that well done. The actors repeat the same lines as in "All Stars" but without much feeling.

(Radford et al. 2017: *Learning to Generate Reviews and Discovering Sentiment*)