Lecture 10: Recurrent Neural Networks
Administrative

- Project TA matchups out, see Ed for the link
Administrative

- A2 is due next Monday May 2nd, 11:59pm
Administrative

- Discussion section tomorrow 2:30-3:30PT

Object detection & RNNs Review
Last time: Detection and Segmentation

Classification

Semantic Segmentation

Object Detection

Instance Segmentation

- CAT
- GRASS, CAT, TREE, SKY
- DOG, DOG, CAT
- DOG, DOG, CAT

No spatial extent
No objects, just pixels
Multiple Objects

This image is CC0 public domain
Training “Feedforward” Neural Networks

1. **One time set up**: activation functions, preprocessing, weight initialization, regularization, gradient checking

2. **Training dynamics**: babysitting the learning process, parameter updates, hyperparameter optimization

3. **Evaluation**: model ensembles, test-time augmentation, transfer learning
Today: Recurrent Neural Networks
“Vanilla” Neural Network

one to one

Vanilla Neural Networks
Recurrent Neural Networks: Process Sequences

one to one

one to many

e.g. Image Captioning
image -> sequence of words
Recurrent Neural Networks: Process Sequences

e.g. **action prediction**
sequence of video frames -> action class
Recurrent Neural Networks: Process Sequences

E.g. Video Captioning
Sequence of video frames -> caption
Recurrent Neural Networks: Process Sequences

- One to one
- One to many
- Many to one
- Many to many

E.g. Video classification on frame level
Sequential Processing of Non-Sequence Data

Classify images by taking a series of “glimpses”

Figure copyright Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra, 2015. Reproduced with permission.
Sequential Processing of Non-Sequence Data
Generate images one piece at a time!

Gregor et al, "DRAW: A Recurrent Neural Network for Image Generation", ICML 2015
Figure copyright Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra, 2015. Reproduced with permission.
Recurrent Neural Network
Recurrent Neural Network

Key idea: RNNs have an “internal state” that is updated as a sequence is processed.
Unrolled RNN

\[ y_1 \quad y_2 \quad y_3 \quad \ldots \quad y_t \]

\[ x_1 \quad x_2 \quad x_3 \quad \ldots \quad x_t \]
We can process a sequence of vectors $\mathbf{x}$ by applying a recurrence formula at every time step:

$$h_t = f_W(h_{t-1}, x_t)$$

- $h_t$: new state
- $h_{t-1}$: old state
- $x_t$: input vector at some time step
- $f_W$: some function with parameters $W$
RNN output generation

We can process a sequence of vectors $\mathbf{x}$ by applying a **recurrence formula** at every time step:

$$y_t = f_{W_{hy}}(h_t)$$

- $y_t$: output
- $h_t$: new state
- $f_{W_{hy}}$: another function with parameters $W_o$
Recurrent Neural Network

\[ \begin{align*}
RNN & \quad h_0 \quad x_1 \\
RNN & \quad h_1 \quad x_2 \\
RNN & \quad h_2 \quad x_3 \\
RNN & \quad h_3 \quad x_t \\
\end{align*} \]
Recurrent Neural Network

We can process a sequence of vectors $x$ by applying a recurrence formula at every time step:

$$h_t = f_W (h_{t-1}, x_t)$$

Notice: the same function and the same set of parameters are used at every time step.
(Vanilla) Recurrent Neural Network

The state consists of a single "hidden" vector $h_t$:

$$h_t = f_W(h_{t-1}, x_t)$$

$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t)$$

$$y_t = W_{hy}h_t$$

Sometimes called a “Vanilla RNN” or an “Elman RNN” after Prof. Jeffrey Elman.
RNN: Computational Graph
RNN: Computational Graph
RNN: Computational Graph

\[ h_0 \xrightarrow{f_W} h_1 \xrightarrow{f_W} h_2 \xrightarrow{f_W} h_3 \xrightarrow{f_W} \ldots \xrightarrow{f_W} h_T \]
RNN: Computational Graph

Re-use the same weight matrix at every time-step
RNN: Computational Graph: Many to Many
RNN: Computational Graph: Many to Many

\[ h_0 \xrightarrow{f_w} h_1 \xrightarrow{f_w} h_2 \xrightarrow{f_w} h_3 \xrightarrow{\ldots} h_T \]

\[ x_1 \xrightarrow{W} h_0 \]

\[ y_1 \xrightarrow{L_1} y_2 \xrightarrow{L_2} y_3 \xrightarrow{\ldots} y_T \]
RNN: Computational Graph: Many to Many

\[ h_0 \xrightarrow{f_W} h_1 \xrightarrow{f_W} h_2 \xrightarrow{f_W} h_3 \xrightarrow{\ldots} h_T \]

\[ W \]

\[ x_1 \xrightarrow{W} \]

\[ x_2 \xrightarrow{W} \]

\[ x_3 \xrightarrow{W} \]

\[ y_1 \xrightarrow{L_1} \]

\[ y_2 \xrightarrow{L_2} \]

\[ y_3 \xrightarrow{L_3} \]

\[ y_T \xrightarrow{L_T} \]

L
RNN: Computational Graph: Many to One

\[ h_0 \xrightarrow{f_W} h_1 \xrightarrow{f_W} h_2 \xrightarrow{f_W} h_3 \xrightarrow{f_W} \ldots \xrightarrow{} h_T \]

\[ x_1 \xrightarrow{W} \]

\[ x_2 \xrightarrow{} \]

\[ x_3 \xrightarrow{} \]

\[ y \xrightarrow{} \]
RNN: Computational Graph: Many to One

\[ h_0 \xrightarrow{f_W} h_1 \xrightarrow{f_W} h_2 \xrightarrow{f_W} h_3 \xrightarrow{\ldots} h_T \]

\[ W \]

\[ x_1 \xrightarrow{W} \]

\[ x_2 \]

\[ x_3 \]

\[ y \]
RNN: Computational Graph: One to Many

\[ h_0 \rightarrow f_W \rightarrow h_1 \rightarrow f_W \rightarrow h_2 \rightarrow f_W \rightarrow h_3 \rightarrow \ldots \rightarrow h_T \]

\[ y_1 \rightarrow y_2 \rightarrow y_3 \rightarrow \ldots \rightarrow y_T \]

\[ W \rightarrow x \]
RNN: Computational Graph: One to Many

\[
\begin{align*}
W & \xrightarrow{f_W} h_0 \xrightarrow{f_W} h_1 \xrightarrow{f_W} h_2 \xrightarrow{f_W} h_3 \xrightarrow{f_W} \ldots \xrightarrow{f_W} h_T \\
x & \mapsto y_1 \quad ? \quad ? \quad ? \quad ? \quad ? \\
\end{align*}
\]
RNN: Computational Graph: One to Many

\[ h_0 \rightarrow f_W \rightarrow h_1 \rightarrow f_W \rightarrow h_2 \rightarrow f_W \rightarrow h_3 \rightarrow \ldots \rightarrow h_T \]

\[ y_1 \rightarrow y_2 \rightarrow y_3 \rightarrow y_T \]

\[ W \rightarrow x \rightarrow 0 \rightarrow 0 \rightarrow 0 \]
RNN: Computational Graph: One to Many

```
\begin{array}{c}
  h_0 \rightarrow f_W \rightarrow h_1 \\
  \uparrow & \uparrow & \uparrow \\
  \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \ quad\end{array}
```
Sequence to Sequence: Many-to-one + one-to-many

Many to one: Encode input sequence in a single vector

Sutskever et al, “Sequence to Sequence Learning with Neural Networks”, NIPS 2014
Sequence to Sequence: Many-to-one + one-to-many

Many to one: Encode input sequence in a single vector

One to many: Produce output sequence from single input vector

Sutskever et al, “Sequence to Sequence Learning with Neural Networks”, NIPS 2014
Example:
Character-level Language Model

Vocabulary:
[h,e,l,o]

Example training sequence:
“hello”
Example:
Character-level Language Model

Vocabulary: [h,e,l,o]

Example training sequence: “hello”
Example:
Character-level Language Model

Vocabulary:
[h,e,l,o]

Example training sequence:
“hello”
Example: Character-level Language Model Sampling

Vocabulary: [h,e,l,o]

At test-time sample characters one at a time, feed back to model
**Example: Character-level Language Model Sampling**

Vocabulary: [h,e,l,o]

At test-time sample characters one at a time, feed back to model

---

**Softmax**

```
Sample

Softmax

output layer

hidden layer

input layer

input chars: “h”
```

```
<table>
<thead>
<tr>
<th>Softmax</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>e</td>
<td>0.3</td>
</tr>
<tr>
<td>l</td>
<td>-0.1</td>
</tr>
<tr>
<td>o</td>
<td>0.9</td>
</tr>
</tbody>
</table>

```

```
<table>
<thead>
<tr>
<th>Input</th>
<th>h</th>
<th>e</th>
</tr>
</thead>
<tbody>
<tr>
<td>Layer</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
```

---

**“e”**
Example: Character-level Language Model Sampling

Vocabulary: [h,e,l,o]

At test-time sample characters one at a time, feed back to model.
Example: Character-level Language Model Sampling

Vocabulary: [h,e,l,o]

At test-time sample characters one at a time, feed back to model
Matrix multiply with a one-hot vector just extracts a column from the weight matrix. We often put a separate **embedding** layer between input and hidden layers.
Backpropagation through time

Forward through entire sequence to compute loss, then backward through entire sequence to compute gradient
Truncated Backpropagation through time

Run forward and backward through chunks of the sequence instead of whole sequence
**Truncated Backpropagation through time**

Carry hidden states forward in time forever, but only backpropagate for some smaller number of steps.
Truncated Backpropagation through time
Fei-Fei Li, Jiajun Wu, Ruohan Gao

Lecture 10 - 50
April 28, 2022

min-char-rnn.py gist: 112 lines of Python

https://gist.github.com/karpathy/d4dee566867f8291f086)
THE SONNETS

by William Shakespeare

From fairest creatures we desire increase,
That thereby beauty’s rose might never die,
But as the riper should by time decease,
His tender heir might bear his memory:
But thou, contracted to thine own bright eyes,
Feed’st thy light’s flame with self-substantial fuel,
Making a famine where abundance lies,
Thyself thy foe, to thy sweet self too cruel:
Thou art now the world’s fresh ornament,
And only herald to the gaudy spring,
Within thine own bud hast thou lost thy content,
And tender churl mak’st waste in niggarding:
Pity the world, or else this glutton be,
To eat the world’s due, by the grave and thee.

When forty winters shall besiege thy brow,
And dig deep trenches in thy beauty’s field,
Thy youth’s proud livery so gazed on now,
Will be a tatter’d weed of small worth held:
Then being asked, where all thy beauty lies,
Where all the treasure of thy lusty days;
To say, within thine own deep sunken eyes,
Were an all-seeing shame, and thriftless praise.
How much more praise deserv’d thy beauty’s use,
If thou couldst answer ‘This fair child of mine
Shall sum my count, and make my old excuse,’
Proving his beauty by succession thine!
This were to be new made when thou art old,
And see thy blood warm when thou feel’st it cold.
at first:

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

Aftair fall unsuch that the hall for Prince Velzonski's that me of
her hearly, and behs to so arwage fiving were to it beloge, pavu say falling misfort
how, and Gogition is so overelical and ofter.

"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.
PANDARUS:
Alas, I think he shall be come approached and the day
When little strain 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.
The Stacks Project: open source algebraic geometry textbook

The Stacks Project

Browse chapters

<table>
<thead>
<tr>
<th>Part</th>
<th>Chapter</th>
<th>online</th>
<th>TeX source</th>
<th>view pdf</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preliminaries</td>
<td>1. Introduction</td>
<td>online</td>
<td>tex</td>
<td>pdf</td>
</tr>
<tr>
<td></td>
<td>2. Conventions</td>
<td>online</td>
<td>tex</td>
<td>pdf</td>
</tr>
<tr>
<td></td>
<td>3. Set Theory</td>
<td>online</td>
<td>tex</td>
<td>pdf</td>
</tr>
<tr>
<td></td>
<td>4. Categories</td>
<td>online</td>
<td>tex</td>
<td>pdf</td>
</tr>
<tr>
<td></td>
<td>5. Topology</td>
<td>online</td>
<td>tex</td>
<td>pdf</td>
</tr>
<tr>
<td></td>
<td>6. Sheaves on Spaces</td>
<td>online</td>
<td>tex</td>
<td>pdf</td>
</tr>
<tr>
<td></td>
<td>7. Sites and Sheaves</td>
<td>online</td>
<td>tex</td>
<td>pdf</td>
</tr>
<tr>
<td></td>
<td>8. Stacks</td>
<td>online</td>
<td>tex</td>
<td>pdf</td>
</tr>
<tr>
<td></td>
<td>9. Fields</td>
<td>online</td>
<td>tex</td>
<td>pdf</td>
</tr>
<tr>
<td></td>
<td>10. Commutative Algebra</td>
<td>online</td>
<td>tex</td>
<td>pdf</td>
</tr>
</tbody>
</table>

Parts

1. Preliminaries
2. Schemes
3. Topics in Scheme Theory
4. Algebraic Spaces
5. Topics in Geometry
6. Deformation Theory
7. Algebraic Stacks
8. Miscellany

Statistics

The Stacks project now consists of
- 455910 lines of code
- 14221 tags (56 inactive tags)
- 2366 sections
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 \( \text{Proj}_X(A) = \text{Spec}(B) \) over \( U \) compatible with the complex
\[
\text{Set}(A) = \Gamma(X, \mathcal{O}_X, \mathcal{O}_X).
\]
When in this case of to show that \( Q \to C_{/X} \) 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 \( Z \subset X \) 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

\[
(1) \ f \text{ is locally of finite type. Since } S = \text{Spec}(R) \text{ and } Y = \text{Spec}(R).
\]

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

The following result surjective restrocomposes of this implies that \( F_{x_0} = F_{x_0} = F_{x_i,0} \).

Lemma 0.2. Let \( X \) be a locally Noetherian scheme over \( S \), \( E = F_{X/S} \). Set \( T = J_i \subset T' \). Since \( T' \subset T' \) are nonzero over \( i_0 \leq p \) is a subset of \( J_{i_0,0} \to A_2 \) works.

Lemma 0.3. In Situation ?? and hence we may assume \( q = 0 \).

Proof. We will use the property we see that \( p \) is the next 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 \( \delta_{x+1} \) is a scheme over \( S \).

---

For \( \oplus_{n=1,\ldots,m} \) where \( L_n = 0 \), hence we can find a closed subset \( \mathcal{H} \) in \( \mathcal{H} \) and any sets \( F \) 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
\[
S = \text{Spec}(R) = U \times \times U \times \times U
\]
and the comparely in the fibre product covering we have to prove the lemma generated by \( \prod_{i} \mathcal{H} \to V \). Consider the maps \( M \) along the set of points \( \text{Sch}_{/\mathcal{H}} \) and \( U \to T \) 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 \( \mathcal{H}(G) \) such that \( \text{Spec}(R) \to S \) is smooth or an
\[
U = \bigcup_{i_k} U \times S \times U
\]
which has a nonzero morphism we may assume that \( f \) is of finite presentation over \( S \). We claim that \( \mathcal{O}_{X,S} \) is a scheme where \( x, x', x'' \in S' \) such that \( \mathcal{O}_{X,x} \to \mathcal{O}_{X,x'} \) is separated. By Algebra, Lemma ?? we can define a map of complexes \( \mathcal{O}_{X,S} \to \mathcal{O}_{X,S'} \) and we win.

To prove this we see that \( F_{i,k} \) is a covering of \( X' \), and \( T_{i} \) is an object of \( X_{/S} \) for \( i > 0 \) and \( F_{i,k} \) exists and let \( F_{i,k} \) be a presheaf of \( \mathcal{O}_{X,S} \)-modules on \( C \) as a \( F \)-module.

In particular \( F = U/F \) we have to show that
\[
\bar{M}^* = T^* \otimes_{\text{Spec}(k)} \mathcal{O}_{X,S} \to \mathcal{O}_{X,S}
\]
is a unique morphism of algebraic stacks. Note that
\[
\text{Arrows} = (\text{Sch}/S)^{\text{pp}} \times_{\text{Sch}/S} \text{Spec}(A)
\]
and
\[
V = \Gamma(S, \mathcal{O}_S) \to (U, \text{Spec}(A))
\]
is an open subset of \( X \). Thus \( U \to S \) 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 \( X_{\text{space,etale}} \) which gives an open subspace of \( X \) and \( T \) equal to \( S_{2,\text{etale}} \), see Descent, Lemma ??.

Namely, by Lemma ?? we see that \( R \) is geometrically regular over \( S \).
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

$$O_{\mathcal{C}} = O_{\mathcal{X}}(\mathcal{L})$$

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

$$O_{\mathcal{X}}(\mathcal{F}) = \{\text{morph}_1 \times_{O_{\mathcal{X}}} (\mathcal{G}, \mathcal{F})\}$$

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

**Lemma 0.2.** This is an integer $\mathbb{Z}$ 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 $U \subset X$ 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

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

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.
Linux kernel source tree

- 520,037 commits
- 1 branch
- 420 releases
- 5,039 contributors

Branch: master

- torvalds authored 9 hours ago
  - latest commit 4b1706927d

- Documentation
  - Merge git://git.kernel.org/pub/scm/linux/kernel/git/nab/target-pending
    - 6 days ago

- arch
  - Merge branch 'x86-urgent-for-linux' of git://git.kernel.org/pub/scm/…
    - a day ago

- block
  - block: discard bdi_unregister() in favour of bdi_destroy()
    - 9 days ago

- crypto
  - Merge git://git.kernel.org/pub/scm/linux/kernel/git/herbert/crypto-2.6
    - 10 days ago

- drivers
  - Merge branch 'drm-fixes' of git://people.freedesktop.org/~airlied/linux
    - 9 hours ago

- firmware
  - firmware/hex2fw.c: restore missing default in switch statement
    - 2 months ago

- fs
  - vfs: read_file_handle only once in handle_to_path
    - 4 days ago

- include
  - Merge branch 'perf-urgent-for-linux' of git://git.kernel.org/pub/scm/…
    - a day ago

- init
  - init: fix regression by supporting devices with major:minor:offset fo…
    - a month ago

HTTPS clone URL
https://github.com
static void do_command(struct seq_file *m, void *v)
{
    int column = 32 << (cmd[2] & 0x80);
    if (state)
        cmd = (int)(int_state ^ (in_8(&ch->ch_flags) & Cmd) ? 2 : 1);
    else
        seq = 1;
    for (i = 0; i < 16; i++) {
        if (k & (1 << i))
            pipe = (in_use & UMXTTHREAD_UNCCA) +
                ((count & 0x00000000fffffff8) & 0x000000f) << 8;
        if (count == 0)
            sub(pid, ppc_md.kexec_handle, 0x20000000);
        pipe_set_bytes(i, 0);
    }
    /* Free our user pages pointer to place camera if all dash */
    subsystem_info = &of_changes[PAGE_SIZE];
    rek_controls(offset, idx, &offset);
    /* Now we want to deliberately put it to device */
    control_check_polarity(&context, val, 0);
    for (i = 0; i < COUNTER; i++)
        seq_puts(s, "policy ");
}
/*
 * Copyright (c) 2006-2010, Intel Mobile Communications. All rights reserved.
 *
 * This program is free software; you can redistribute it and/or modify it
 * under the terms of the GNU General Public License version 2 as published by
 * the Free Software Foundation.
 *
 * This program is distributed in the hope that it will be useful,
 * but WITHOUT ANY WARRANTY; without even the implied warranty of
 * MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the
 * GNU General Public License for more details.
 *
 * You should have received a copy of the GNU General Public License
 * along with this program; if not, write to the Free Software Foundation,
 * Inc., 675 Mass Ave, Cambridge, MA 02139, USA.
 */

#include <linux/kexec.h>
#include <linux/errno.h>
#include <linux/io.h>
#include <linux/platform_device.h>
#include <linux/multi.h>
#include <linux/ckevent.h>

#include <asm/io.h>
#include <asm/prom.h>
#include <asm/e820.h>
#include <asm/system_info.h>
#include <asm/setev.h>
#include <asm/pgproto.h>
```c
#include <asm/io.h>
#include <asm/prom.h>
#include <asm/e820.h>
#include <asm/system_info.h>
#include <asm/setev.h>
#include <asm/pgproto.h>

#define REG_PG     vesa_slot_addr_pack
#define PFM_NOPCOM  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)
{
  #ifndef 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;
  
```
Add this image of a rocketship:
https://i1.sndcdn.com/artworks-j8xjG7zc1wmTeO7b-06I83w-t500x500.jpg
In a shocking finding, scientists discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

The scientist named the population, after their distinctive horn, Ovid’s Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.
Searching for interpretable cells
Searching for interpretable cells
Searching for interpretable cells

"You mean to imply that I have nothing to eat out of.... On the contrary, I can supply you with everything even if you want to give dinner parties," warmly replied Chichagov, who tried by every word he spoke to prove his own rectitude and therefore imagined Kutuzov to be animated by the same desire.

Kutuzov, shrugging his shoulders, replied with his subtle penetrating smile: "I meant merely to say what I said."

quote detection cell
Searching for interpretable cells

Cell sensitive to position in line:

The sole importance of the crossing of the Berezina lies in the fact that it plainly and indubitably proved the fallacy of all the plans for cutting off the enemy's retreat and the soundness of the only possible line of action--the one Kutuzov and the general mass of the army demanded--namely, simply to follow the enemy up. The French crowd fled at a continually increasing speed and all its energy was directed to reaching its goal. It fled like a wounded animal and it was impossible to block its path. This was shown not so much by the arrangements it made for crossing as by what took place at the bridges. When the bridges broke down, unarmed soldiers, people from Moscow and women with children who were with the French transport, all--carried on by vis inertiae--pressed forward into boats and into the ice-covered water and did not surrender.

line length tracking cell

Karpathy, Johnson, and Fei-Fei: Visualizing and Understanding Recurrent Networks, ICLR Workshop 2016
Figures copyright Karpathy, Johnson, and Fei-Fei, 2015; reproduced with permission
Searching for interpretable cells

```c
static int __dequeue_signal(struct sigpending *pending, sigset_t *mask, siginfo_t *info)
{
    int sig = next_signal(pending, mask);
    if (sig != -1) {
        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;
}
```

if statement cell

Karpathy, Johnson, and Fei-Fei: Visualizing and Understanding Recurrent Networks, ICLR Workshop 2016
Figures copyright Karpathy, Johnson, and Fei-Fei, 2015; reproduced with permission
Searching for interpretable cells

```
/* Duplicate LSM field information. The lsm_rule is opaque, so
 * re-initialized. */
static inline int audit_dupe_lsm_field(struct audit_field *df,
            struct audit_field *sf)
{
  int ret = 0;
  char *lsm_str;
  /* our own copy of lsm_str */
  lsm_str = kstrdup(sf->lsm_str, GFP_KERNEL);
  if (unlikely(!lsm_str))
    return -ENOMEM;
  df->lsm_str = lsm_str;
  /* our own (refreshed) copy of lsm_rule */
  ret = security_audit_rule_init(df->type, df->op, df->lsm_str,
      (void **)&df->lsm_rule);
  /* keep currently invalid fields around in case they
   * become valid after a policy reload. */
  if (ret == -EINVAL) {
    pr_warn("audit rule for LSM \
"\%s\" is invalid\n",  
   df->lsm_str);
    ret = 0;
  }
  return ret;
}
```

quote/comment cell

Karpathy, Johnson, and Fei-Fei: Visualizing and Understanding Recurrent Networks, ICLR Workshop 2016
Figures copyright Karpathy, Johnson, and Fei-Fei, 2015; reproduced with permission
Searching for interpretable cells

```c
#ifdef CONFIG_AUDITSYSSCALL
static inline int audit_match_class_bits(int class, u32 *mask)
{
    int i;
    if (classes[class]) {
        for (i = 0; i < AUDIT_BITMASK_SIZE; i++)
            if (mask[i] & classes[class][i])
                return 0;
    }
    return 1;
}
```

code depth cell

Karpathy, Johnson, and Fei-Fei: Visualizing and Understanding Recurrent Networks, ICLR Workshop 2016
Figures copyright Karpathy, Johnson, and Fei-Fei, 2015; reproduced with permission
RNN tradeoffs

RNN Advantages:
- Can process any length input
- Computation for step $t$ can (in theory) use information from many steps back
- Model size doesn’t increase for longer input
- Same weights applied on every timestep, so there is symmetry in how inputs are processed.

RNN Disadvantages:
- Recurrent computation is slow
- In practice, difficult to access information from many steps back
Image Captioning

Explain Images with Multimodal Recurrent Neural Networks, Mao et al.
Deep Visual-Semantic Alignments for Generating Image Descriptions, Karpathy and Fei-Fei
Show and Tell: A Neural Image Caption Generator, Vinyals et al.
Long-term Recurrent Convolutional Networks for Visual Recognition and Description, Donahue et al.
Learning a Recurrent Visual Representation for Image Caption Generation, Chen and Zitnick
Rec current Neural Network

Convolutional Neural Network
test image
test image
before:
\[ h = \tanh(W_{xh} \cdot x + W_{hh} \cdot h) \]

now:
\[ h = \tanh(W_{xh} \cdot x + W_{hh} \cdot h + W_{ih} \cdot v) \]
Fei-Fei Li, Jiajun Wu, Ruohan Gao

Lecture 10 - April 28, 2022

Image processing model:
- conv-64
- maxpool
- conv-128
- maxpool
- conv-256
- maxpool
- conv-512
- maxpool
- conv-512
- maxpool
- conv-512
- maxpool
- FC-4096
- FC-4096

Sample output:
- x0
- straw

Test image:
- Input image
- Output: y0

Sample text:

`<START>`
test image

sample!

Fei-Fei Li, Jiajun Wu, Ruohan Gao
test image

sample

<END> token

=> finish.
Image Captioning: Example Results

A cat sitting on a suitcase on the floor
A cat is sitting on a tree branch
A dog is running in the grass with a frisbee
A white teddy bear sitting in the grass

Two people walking on the beach with surfboards
A tennis player in action on the court
Two giraffes standing in a grassy field
A man riding a dirt bike on a dirt track
Image Captioning: Failure Cases

A woman is holding a cat in her hand

A woman standing on a beach holding a surfboard

A person holding a computer mouse on a desk

A bird is perched on a tree branch

A man in a baseball uniform throwing a ball

Captions generated using neuraltalk2
All images are CC0 Public domain: fur coat, handstand, spider web, baseball
Visual Question Answering (VQA)

A: A bald eagle.
A: A sparrow.
A: A humming bird.
A: A raven.

Q: What endangered animal is featured on the truck?

A: Onto 24 ¾ Rd.
A: Onto 25 ¾ Rd.
A: Onto 23 ¾ Rd.
A: Onto Main Street.

Q: Where will the driver go if turning right?

A: During a wedding.
A: During a bar mitzvah.
A: During a funeral.
A: During a Sunday church service.

Q: When was the picture taken?

A: Two women.
A: A child.
A: An old man.
A: A husband and a wife.

Q: Who is under the umbrella?


Figure from Zhu et al, copyright IEEE 2016. Reproduced for educational purposes.
Visual Question Answering (VQA)

“How many horses are in this image?”

Visual Dialog: Conversations about images

Das et al., "Visual Dialog", CVPR 2017
Figures from Das et al, copyright IEEE 2017. Reproduced with permission.
Agent encodes instructions in language and uses an RNN to generate a series of movements as the visual input changes after each move.

**Instruction**

Turn right and head towards the kitchen. Then turn left, pass a table and enter the hallway. Walk down the hallway and turn into the entry way to your right without doors. Stop in front of the toilet.


Figures from Wang et al, copyright IEEE 2017. Reproduced with permission.
Visual Question Answering: Dataset Bias

What is the dog playing with?
Frisbee

Image
Question
Answer
Model

Yes or No

<table>
<thead>
<tr>
<th>Models</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image + Question + Answer</td>
<td>67.1</td>
</tr>
<tr>
<td>without question</td>
<td>63.8</td>
</tr>
<tr>
<td>without image</td>
<td>58.5</td>
</tr>
<tr>
<td>without image or question</td>
<td>52.9</td>
</tr>
</tbody>
</table>

Multilayer RNNs
Long Short Term Memory (LSTM)

Vanilla RNN

\[ h_t = \tanh \left( W \left( h_{t-1}, x_t \right) \right) \]

LSTM

\[
\begin{pmatrix}
  i \\
  f \\
  o \\
  g
\end{pmatrix} =
\begin{pmatrix}
  \sigma \\
  \sigma \\
  \sigma \\
  \tanh
\end{pmatrix}
W
\begin{pmatrix}
  h_{t-1} \\
  x_t
\end{pmatrix}
\]

\[
c_t = f \odot c_{t-1} + i \odot g
\]

\[
h_t = o \odot \tanh(c_t)
\]

Hochreiter and Schmidhuber, “Long Short Term Memory”, Neural Computation 1997
Vanilla RNN Gradient Flow

\[ h_t = \tanh(W_{hh} h_{t-1} + W_{xh} x_t) \]

\[ = \tanh \left( (W_{hh} \ W_{hx}) \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix} \right) \]

\[ = \tanh \left( W \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix} \right) \]
Vanilla RNN Gradient Flow

Backpropagation from $h_t$ to $h_{t-1}$ multiplies by $W$ (actually $W_{hh}^T$)

$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t)$$

$$= \tanh \left( (W_{hh} \ W_{hx}) \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix} \right)$$

$$= \tanh \left( W \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix} \right)$$
Vanilla RNN Gradient Flow

Backpropagation from $h_t$ to $h_{t-1}$ multiplies by $W$ (actually $W_{hh}^T$)

$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t)$$
$$= \tanh \left((W_{hh} \quad W_{hx}) \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix}\right)$$
$$= \tanh \left(W \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix}\right)$$

$$\frac{\partial h_t}{\partial h_{t-1}} = \tanh'(W_{hh}h_{t-1} + W_{xh}x_t)W_{hh}$$

Bengio et al, "Learning long-term dependencies with gradient descent is difficult", IEEE Transactions on Neural Networks, 1994
Vanilla RNN Gradient Flow

\[ \frac{\partial L}{\partial W} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial W} \]

Bengio et al., “Learning long-term dependencies with gradient descent is difficult”, IEEE Transactions on Neural Networks, 1994
Vanilla RNN Gradient Flow
Gradients over multiple time steps:

\[
\frac{\partial L_T}{\partial W} = \frac{\partial L_T}{\partial h_T} \frac{\partial h_t}{\partial h_{t-1}} \cdots \frac{\partial h_1}{\partial W}
\]

Bengio et al, “Learning long-term dependencies with gradient descent is difficult”, IEEE Transactions on Neural Networks, 1994
Vanilla RNN Gradient Flow

Gradients over multiple time steps:

\[ \frac{\partial L}{\partial W} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial W} \]

\[ \frac{\partial L_T}{\partial W} = \frac{\partial L_T}{\partial h_T} \frac{\partial h_T}{\partial h_{T-1}} \cdots \frac{\partial h_1}{\partial W} = \frac{\partial L_T}{\partial h_T} \left( \prod_{t=2}^{T} \frac{\partial h_t}{\partial h_{t-1}} \right) \frac{\partial h_1}{\partial W} \]

Bengio et al, “Learning long-term dependencies with gradient descent is difficult”, IEEE Transactions on Neural Networks, 1994
Vanilla RNN Gradient Flow

Gradients over multiple time steps:

\[ \frac{\partial L}{\partial W} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial W} \]

\[ \frac{\partial h_t}{\partial h_{t-1}} = \tanh'(W_{hh}h_{t-1} + W_{xh}x_t)W_{hh} \]

\[ \frac{\partial L_T}{\partial W} = \frac{\partial L_T}{\partial h_T} \frac{\partial h_T}{\partial h_{t-1}} \cdots \frac{\partial h_1}{\partial W} = \frac{\partial L_T}{\partial h_T} \left( \prod_{t=2}^{T} \frac{\partial h_t}{\partial h_{t-1}} \right) \frac{\partial h_1}{\partial W} \]

Bengio et al, “Learning long-term dependencies with gradient descent is difficult”, IEEE Transactions on Neural Networks, 1994
Vanilla RNN Gradient Flow

Gradients over multiple time steps:

\[
\frac{\partial L}{\partial W} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial W}
\]

Almost always < 1
Vanishing gradients

\[
\frac{\partial L_T}{\partial W} = \frac{\partial L_T}{\partial h_T} \left( \prod_{t=2}^{T} \frac{\partial h_t}{\partial h_T} \right) \left( \text{tanh}' \left( W_{hh} h_{t-1} + W_{xh} x_t \right) \right) W_{hh}^{T-1} \frac{\partial h_1}{\partial h_T}
\]

Bengio et al, “Learning long-term dependencies with gradient descent is difficult”, IEEE Transactions on Neural Networks, 1994
Vanilla RNN Gradient Flow

Gradients over multiple time steps:

\[
\frac{\partial L}{\partial W} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial W}
\]

What if we assumed no non-linearity?

Bengio et al, “Learning long-term dependencies with gradient descent is difficult”, IEEE Transactions on Neural Networks, 1994
Vanilla RNN Gradient Flow

Gradients over multiple time steps:

\[
\frac{\partial L}{\partial W} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial W}
\]

What if we assumed no non-linearity?

Largest singular value > 1: Exploding gradients

Largest singular value < 1: Vanishing gradients

Bengio et al, “Learning long-term dependencies with gradient descent is difficult”, IEEE Transactions on Neural Networks, 1994
Vanilla RNN Gradient Flow

Gradients over multiple time steps:

What if we assumed no non-linearity?

\[ \frac{\partial L}{\partial W} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial W} \]

Largest singular value > 1: Exploding gradients

\[ \frac{\partial L_T}{\partial W} = \frac{\partial L_T}{\partial h_T} W^{T-1} \frac{\partial h_1}{\partial W} \]

Largest singular value < 1: Vanishing gradients

Bengio et al, “Learning long-term dependencies with gradient descent is difficult”, IEEE Transactions on Neural Networks, 1994

Gradient clipping:
Scale gradient if its norm is too big

```
grad_norm = np.sum(grad * grad)
if grad_norm > threshold:
    grad *= (threshold / grad_norm)
```
Vanilla RNN Gradient Flow

Gradients over multiple time steps:

\[
\frac{\partial L}{\partial W} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial W}
\]

\[
\frac{\partial L_T}{\partial W} = \frac{\partial L_T}{\partial h_T} W_{T-1}^{-1} \frac{\partial h_1}{\partial W}
\]

What if we assumed no non-linearity?

- Largest singular value > 1: Exploding gradients
- Largest singular value < 1: Vanishing gradients

Bengio et al, “Learning long-term dependencies with gradient descent is difficult”, IEEE Transactions on Neural Networks, 1994

Fei-Fei Li, Jiajun Wu, Ruohan Gao
Lecture 10 - 103
April 28, 2022
Long Short Term Memory (LSTM)

**Vanilla RNN**

\[ h_t = \tanh \left( W \left( h_{t-1} \right) \right) \]

**LSTM**

\[
\begin{pmatrix}
  i \\
  f \\
  o \\
  g
\end{pmatrix}
= \begin{pmatrix}
  \sigma \\
  \sigma \\
  \sigma \\
  \tanh
\end{pmatrix}
W
\begin{pmatrix}
  h_{t-1} \\
  x_t
\end{pmatrix}
\]

\[
c_t = f \odot c_{t-1} + i \odot g
\]

\[
h_t = o \odot \tanh(c_t)
\]

Hochreiter and Schmidhuber, “Long Short Term Memory”, Neural Computation 1997
Long Short Term Memory (LSTM)

Vanilla RNN

\[ h_t = \tanh \left( W \left( h_{t-1}, x_t \right) \right) \]

LSTM

\[
\begin{pmatrix}
i \\
f \\
o \\
g
\end{pmatrix} =
\begin{pmatrix}
\sigma \\
\sigma \\
\sigma \\
\tanh
\end{pmatrix} W \begin{pmatrix} h_{t-1} \\
x_t \end{pmatrix}
\]

\[ c_t = f \odot c_{t-1} + i \odot g \]

\[ h_t = o \odot \tanh(c_t) \]

Hochreiter and Schmidhuber, “Long Short Term Memory”, Neural Computation 1997
Long Short Term Memory (LSTM)

[Hochreiter et al., 1997]
Long Short Term Memory (LSTM)  
[Hochreiter et al., 1997]

\[\begin{align*}
    i &= \sigma (W_i x + W_i h) \\
    f &= \sigma (W_f x + W_f h) \\
    o &= \sigma (W_o x + W_o h) \\
    g &= \tanh (W_g x + W_g h)
\end{align*}\]

\[c_t = f \odot c_{t-1} + i \odot g\]

\[h_t = o \odot \tanh(c_t)\]
Long Short Term Memory (LSTM)  
[Hochreiter et al., 1997]

- **i**: Input gate, whether to write to cell
- **f**: Forget gate, Whether to erase cell
- **o**: Output gate, How much to reveal cell
- **g**: Gate gate (?), How much to write to cell

\[
\begin{pmatrix}
i \\
f \\
o \\
g
\end{pmatrix} = \begin{pmatrix}
\sigma \\
\sigma \\
\sigma \\
tanh
\end{pmatrix} \cdot W \begin{pmatrix}
h_{t-1} \\
x_t
\end{pmatrix}
\]

\[
c_t = f \odot c_{t-1} + i \odot g
\]

\[
h_t = o \odot \tanh(c_t)
\]
Long Short Term Memory (LSTM) 
[Hochreiter et al., 1997]

\[
\begin{align*}
\text{i: Input gate, whether to write to cell} & \\
\text{f: Forget gate, Whether to erase cell} & \\
\text{g: Gate gate (?), How much to write to cell} &
\end{align*}
\]

\[
\begin{align*}
\frac{4h \times 2h}{4h} \rightarrow \frac{4h}{4h} \rightarrow \frac{4 \times h}{}
\end{align*}
\]

\[
\begin{align*}
\text{vector from} & \\
\text{below (x)} & \\
\text{vector from} & \\
\text{before (h)} &
\end{align*}
\]

\[
\begin{align*}
\text{sigmoid} & \\
\text{sigmoid} & \\
\text{sigmoid} & \\
\text{tanh} &
\end{align*}
\]

\[
\begin{align*}
i & \\
f & \\
o & \\
g &
\end{align*}
\]

\[
\begin{align*}
\begin{pmatrix}
i \\
f \\
o \\
g
\end{pmatrix} & = \\
\begin{pmatrix}
\sigma \\
\sigma \\
\sigma \\
\tanh
\end{pmatrix} W \begin{pmatrix}
h_{t-1} \\
x_t
\end{pmatrix}
\end{align*}
\]

\[
\begin{align*}
c_t & = f \odot c_{t-1} + i \odot g \\
h_t & = o \odot \tanh(c_t)
\end{align*}
\]
Long Short Term Memory (LSTM)  
[Hochreiter et al., 1997]

\[
\begin{align*}
\text{Input gate} & : i \\
\text{Forget gate} & : f \\
\text{Output gate} & : o \\
\text{Gate gate} & : g
\end{align*}
\]

\[
\begin{pmatrix}
i \\
f \\
o \\
g
\end{pmatrix} =
\begin{pmatrix}
\sigma \\
\sigma \\
\sigma \\
tanh
\end{pmatrix}
W
\begin{pmatrix}
h_{t-1} \\
x_t
\end{pmatrix}
\]

\[
c_t = f \odot c_{t-1} + i \odot g
\]

\[
h_t = o \odot \tanh(c_t)
\]
Long Short Term Memory (LSTM)

[Hochreiter et al., 1997]

\[
\begin{align*}
W & \quad \text{stack} \\
h_{t-1} & \quad x_t \\
\circ & \quad \text{stack} \\
f & \quad g \\
i & \quad o \\
\sigma & \quad \sigma \\
\text{tanh} & \quad \text{tanh} \\
& \quad h_t
\end{align*}
\]

\[
\begin{pmatrix}
i \\
f \\
o \\
g
\end{pmatrix} =
\begin{pmatrix}
\sigma \\
\sigma \\
\text{tanh}
\end{pmatrix}
W
\begin{pmatrix}
h_{t-1} \\
x_t
\end{pmatrix}
\]

\[
c_t = f \odot c_{t-1} + i \odot g \\
h_t = o \odot \text{tanh}(c_t)
\]
Long Short Term Memory (LSTM): Gradient Flow

[ Hochreiter et al., 1997 ]

$$c_t = f \odot c_{t-1} + i \odot g$$

$$h_t = o \odot \tanh(c_t)$$
Long Short Term Memory (LSTM): Gradient Flow
[Hochreiter et al., 1997]

Uninterrupted gradient flow!

Notice that the gradient contains the f gate’s vector of activations
- allows better control of gradients values, using suitable parameter updates of the forget gate.

Also notice that are added through the f, i, g, and o gates
- better balancing of gradient values
Do LSTMs solve the vanishing gradient problem?

The LSTM architecture makes it easier for the RNN to preserve information over many timesteps

- e.g. if the $f = 1$ and the $i = 0$, then the information of that cell is preserved indefinitely.
- By contrast, it’s harder for vanilla RNN to learn a recurrent weight matrix $W_h$ that preserves info in hidden state

LSTM doesn’t guarantee that there is no vanishing/exploding gradient, but it does provide an easier way for the model to learn long-distance dependencies.
Long Short Term Memory (LSTM): Gradient Flow

[Hochreiter et al., 1997]

Uninterrupted gradient flow!

Similar to ResNet!
Long Short Term Memory (LSTM): Gradient Flow

[Hochreiter et al., 1997]

Uninterrupted gradient flow!

In between:

Highway Networks

Srivastava et al, "Highway Networks",
ICML DL Workshop 2015

In between:

\[ g = T(x, W_T) \]
\[ y = g \odot H(x, W_H) + (1 - g) \odot x \]
Other RNN Variants

**GRU** ([Learning phrase representations using rnn encoder-decoder for statistical machine translation, Cho et al. 2014])

\[
\begin{align*}
    r_t &= \sigma(W_{xr}x_t + W_{hr}h_{t-1} + b_r) \\
    z_t &= \sigma(W_{xz}x_t + W_{hz}h_{t-1} + b_z) \\
    \tilde{h}_t &= \tanh(W_{xh}x_t + W_{hh}(r_t \odot h_{t-1}) + b_h) \\
    h_t &= z_t \odot h_{t-1} + (1 - z_t) \odot \tilde{h}_t
\end{align*}
\]

[[LSTM: A Search Space Odyssey, Greff et al., 2015]]

[An Empirical Exploration of Recurrent Network Architectures, Jozefowicz et al., 2015]

\[
\begin{align*}
    z_t &= \text{sigmoid}(W_{xz}x_t + W_{hz}h_{t-1} + b_z) \\
    r_t &= \text{sigmoid}(W_{xr}x_t + W_{hr}h_t + b_r) \\
    h_{t-1} &= \text{tanh}(W_{hh}(r_t \odot h_t) + h_t) \odot z \\
    h_t &= h_{t-1} \odot (1 - z) \\
    h_{t+1} &= \tanh(W_{hh}(r_t \odot h_t) + W_{xh}x_t + b_h) \odot z \\
    h_t &= h_{t+1} \odot (1 - z)
\end{align*}
\]
Neural Architecture Search for RNN architectures

LSTM cell

Cell they found

Zoph et Le, "Neural Architecture Search with Reinforcement Learning", ICLR 2017
Figures copyright Zoph et al., 2017. Reproduced with permission.
Summary

- RNNs allow a lot of flexibility in architecture design
- Vanilla RNNs are simple but don’t work very well
- Common to use LSTM or GRU: their additive interactions improve gradient flow
- Backward flow of gradients in RNN can explode or vanish. Exploding is controlled with gradient clipping. Vanishing is controlled with additive interactions (LSTM)
- Better/simpler architectures are a hot topic of current research, as well as new paradigms for reasoning over sequences
- Better understanding (both theoretical and empirical) is needed.
Next time: Attention and Transformers