RNN:

\[ h_t^l = \tanh W^l \left( \begin{array}{c} h_{t-1}^l \\ h_{t-1}^l \end{array} \right) \]

\( h \in \mathbb{R}^n \quad W^l \ [n \times 2n] \)
LSTM

LSTM: A Search Space Odyssey, Greff et al.
RNN:

\[ h_t^l = \tanh \ W^l \left( \begin{pmatrix} h_t^{l-1} & h_{t-1}^l \end{pmatrix} \right) \]

\( h \in \mathbb{R}^n \quad W^l \ [n \times 2n] \)

LSTM:

\[
\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \text{sigm} \\ \text{sigm} \\ \text{sigm} \\ \text{tanh} \end{pmatrix} \ W^l \left( \begin{pmatrix} h_t^{l-1} & h_{t-1}^l \end{pmatrix} \right)
\]

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

\[ h_t^l = o \odot \tanh(c_t^l) \]
Datasets

Leo Tolstoy’s “War and Peace”

This black-eyed, wide-mouthed girl, not pretty but full of life— with childish bare shoulders which after her run heaved and shook her bodice, with black curls tossed backward, thin bare arms, little legs in lace-frilled drawers, and feet in low slippers— was just at that charming age when a girl is no longer a child, though the child is not yet a young woman. Escaping from her father she ran to hide her flushed face in the lace of her mother’s mantilla—not paying the least attention to her severe remark—and began to laugh. She laughed, and in fragmentary sentences tried to explain about a doll which she produced from the folds of her frock.

"Do you see?... My doll... Mimi... You see..." was all Natasha managed to utter (to her everything seemed funny). She leaned against her mother and burst into such a loud, ringing fit of laughter that even the prim visitor could not help joining in.

"Now then, go away and take your monstroucity with you," said the mother, pushing away her daughter with pretended sternness, and turning to the visitor she added: "She is my youngest girl."

Natasha, raising her face for a moment from her mother’s mantilla, glanced up at her through tears of laughter, and again hid her face.

The visitor, compelled to look on at this family scene, thought it necessary to take some part in it.

"Tell me, my dear," said she to Natasha, "is Mimi a relation of yours? A daughter, I suppose?"

Natasha did not like the visitor’s tone of condescension to childish things. She did not reply, but looked at her seriously.
Character-level language modelling
### Step 1: Performance

<table>
<thead>
<tr>
<th>Model</th>
<th>$n$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>War and Peace Dataset</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n$-gram</td>
<td></td>
<td>2.399</td>
<td>1.928</td>
<td>1.521</td>
<td>1.314</td>
<td>1.232</td>
<td>1.203</td>
<td><strong>1.194</strong></td>
<td>1.194</td>
<td>1.194</td>
<td>1.195</td>
</tr>
<tr>
<td>$n$-NN</td>
<td></td>
<td>2.399</td>
<td>1.931</td>
<td>1.553</td>
<td>1.451</td>
<td>1.339</td>
<td><strong>1.321</strong></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Linux Kernel Dataset</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n$-gram</td>
<td></td>
<td>2.702</td>
<td>1.954</td>
<td>1.440</td>
<td>1.213</td>
<td>1.097</td>
<td>1.027</td>
<td>0.982</td>
<td>0.953</td>
<td>0.933</td>
<td><strong>0.889</strong></td>
</tr>
<tr>
<td>$n$-NN</td>
<td></td>
<td>2.707</td>
<td>1.974</td>
<td>1.505</td>
<td>1.395</td>
<td><strong>1.256</strong></td>
<td>1.376</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
## Step 1: Performance

### War and Peace Dataset

<table>
<thead>
<tr>
<th>Model</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>(n)-gram</td>
<td>2.399</td>
<td>1.928</td>
<td>1.521</td>
<td>1.314</td>
<td>1.232</td>
<td>1.203</td>
<td><strong>1.194</strong></td>
<td>1.194</td>
<td>1.194</td>
<td>1.195</td>
</tr>
<tr>
<td>(n)-NN</td>
<td>2.399</td>
<td>1.931</td>
<td>1.553</td>
<td>1.451</td>
<td>1.339</td>
<td><strong>1.321</strong></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### Linux Kernel Dataset

<table>
<thead>
<tr>
<th>Model</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>(n)-gram</td>
<td>2.702</td>
<td>1.954</td>
<td>1.440</td>
<td>1.213</td>
<td>1.097</td>
<td>1.027</td>
<td>0.982</td>
<td>0.953</td>
<td>0.933</td>
<td><strong>0.889</strong></td>
</tr>
<tr>
<td>(n)-NN</td>
<td>2.707</td>
<td>1.974</td>
<td>1.505</td>
<td>1.395</td>
<td><strong>1.256</strong></td>
<td>1.376</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### LSTM vs RNN vs GRU

<table>
<thead>
<tr>
<th>Size</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>War and Peace Dataset</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>64</td>
<td>1.449</td>
<td>1.442</td>
<td>1.540</td>
<td>1.446</td>
<td>1.401</td>
<td>1.396</td>
<td>1.398</td>
<td><strong>1.373</strong></td>
<td>1.472</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>128</td>
<td>1.277</td>
<td><strong>1.227</strong></td>
<td>1.279</td>
<td>1.417</td>
<td>1.286</td>
<td>1.277</td>
<td>1.230</td>
<td><strong>1.226</strong></td>
<td>1.253</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>256</td>
<td>1.189</td>
<td><strong>1.137</strong></td>
<td>1.141</td>
<td>1.342</td>
<td>1.256</td>
<td>1.239</td>
<td>1.198</td>
<td>1.164</td>
<td><strong>1.138</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>512</td>
<td>1.161</td>
<td>1.092</td>
<td>1.082</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.170</td>
<td>1.201</td>
<td><strong>1.077</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Linux Kernel Dataset |
| 64     | 1.355 | **1.331** | 1.366 | 1.407 | 1.371 | 1.383 | 1.335 | 1.298 | 1.357 |       |       |       |
| 128    | 1.149 | **1.128** | 1.177 | 1.241 | 1.120 | 1.220 | 1.154 | 1.125 | 1.150 |       |       |       |
| 256    | 1.026 | **0.972** | 0.998 | 1.171 | 1.116 | 1.116 | 1.039 | 0.991 | 1.026 |       |       |       |
| 512    | 0.952 | 0.840 | 0.846 | -     | -     | -     | 0.943 | 0.861 | **0.829** |       |       |       |
## Step 1: Performance

<table>
<thead>
<tr>
<th>Model</th>
<th>$n$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>War and Peace Dataset</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n$-gram</td>
<td>1</td>
<td>2.399</td>
<td>1.928</td>
<td>1.521</td>
<td>1.314</td>
<td>1.232</td>
<td>1.203</td>
<td><strong>1.194</strong></td>
<td>1.194</td>
<td>1.194</td>
<td>1.195</td>
</tr>
<tr>
<td>$n$-NN</td>
<td>1</td>
<td>2.399</td>
<td>1.931</td>
<td>1.553</td>
<td>1.451</td>
<td>1.339</td>
<td><strong>1.321</strong></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Linux Kernel Dataset</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n$-gram</td>
<td>2</td>
<td>2.702</td>
<td>1.954</td>
<td>1.440</td>
<td>1.213</td>
<td>1.097</td>
<td>1.027</td>
<td>0.982</td>
<td>0.953</td>
<td>0.933</td>
<td>0.889</td>
</tr>
<tr>
<td>$n$-NN</td>
<td>2</td>
<td>2.707</td>
<td>1.974</td>
<td>1.505</td>
<td>1.395</td>
<td><strong>1.256</strong></td>
<td>1.376</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### LSTM, RNN, GRU

<table>
<thead>
<tr>
<th>Layers</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>War and Peace Dataset</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Size</td>
<td>64</td>
<td>1.449</td>
<td>1.442</td>
<td>1.540</td>
<td>1.446</td>
<td>1.401</td>
<td>1.396</td>
<td>1.398</td>
<td><strong>1.373</strong></td>
</tr>
<tr>
<td></td>
<td>128</td>
<td>1.277</td>
<td><strong>1.227</strong></td>
<td>1.279</td>
<td>1.417</td>
<td>1.286</td>
<td>1.277</td>
<td>1.230</td>
<td><strong>1.226</strong></td>
</tr>
<tr>
<td></td>
<td>256</td>
<td>1.189</td>
<td><strong>1.137</strong></td>
<td>1.141</td>
<td>1.342</td>
<td>1.256</td>
<td>1.239</td>
<td>1.198</td>
<td>1.164</td>
</tr>
<tr>
<td></td>
<td>512</td>
<td>1.161</td>
<td>1.092</td>
<td>1.082</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.170</td>
<td>1.201</td>
</tr>
<tr>
<td><strong>Linux Kernel Dataset</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Size</td>
<td>64</td>
<td>1.355</td>
<td><strong>1.331</strong></td>
<td>1.366</td>
<td>1.407</td>
<td>1.371</td>
<td>1.383</td>
<td>1.335</td>
<td>1.298</td>
</tr>
<tr>
<td></td>
<td>128</td>
<td>1.149</td>
<td><strong>1.128</strong></td>
<td>1.177</td>
<td>1.241</td>
<td>1.120</td>
<td>1.220</td>
<td>1.154</td>
<td>1.125</td>
</tr>
<tr>
<td></td>
<td>256</td>
<td>1.026</td>
<td><strong>0.972</strong></td>
<td>0.998</td>
<td>1.171</td>
<td>1.116</td>
<td>1.116</td>
<td>1.039</td>
<td>0.991</td>
</tr>
<tr>
<td></td>
<td>512</td>
<td>0.952</td>
<td>0.840</td>
<td>0.846</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.943</td>
<td>0.861</td>
</tr>
</tbody>
</table>
Linux
Linux

War and Peace
Closing brace ("{" ) case study
Searching for interpretable cells

\[
\begin{pmatrix}
  i \\
  f \\
  o \\
  g
\end{pmatrix}
= \begin{pmatrix}
  \text{sigm} \\
  \text{sigm} \\
  \text{sigm} \\
  \tanh
\end{pmatrix}
W^l \begin{pmatrix}
  h^{l-1}_t \\
  h^l_t \\
  h^l_{t-1}
\end{pmatrix}
\]

\[
c^l_t = f \odot c^l_{t-1} + i \odot g
\]

\[
h^l_t = o \odot \tanh(c^l_t)
\]
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

"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

Interesting note: backpropagation horizon was only 100 time steps
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.
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.
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)
    {
        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
Searching for interpretable cells
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_BITMAP_SIZE; i++)
            if (mask[i] & classes[class][i])
                return 0;
    }
    return 1;
}
#endif
```

code depth cell
Searching for interpretable cells

something interesting cell
(not quite sure what)
Learning Dynamics

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

"Kite vouch!" he repeated by her
door. "But I would be done and quarts, feeling, then, son is people...."

"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 asking his soul came to the packs and drove up his father-in-law women.
Learning Dynamics

Training iterations

![Graph showing mean KL divergence across LSTM epochs for 1-NN to 4-NN](image)

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

We counter. He stutn co des. His stanted out one ofer that co
to gearang reay Jotrets and with fre colt oft paitt thin wall.

Afthair fall unsuch that the hall for Phe
her hearly, and behs to so arwage fivin
how, and Gogition is so overetical and

"Kite vouch!" he repeated by her
door. "But I would be done and quarts, feeling, then, son is

"Why do what that day," replied Natasha
princess, Princess Mary was easier, fe
Pierre aking his soul came to the packs
Learning Dynamics

Sequence to Sequence Learning with Neural Networks,
Sutskever et al.
Error Analysis
n-gram oracle

Optimistic estimate of how many errors could be eliminated by better modeling the last n characters:

- remove error if correctly classified by ANY n-gram model, for n = 1 .. 9

18% of errors
dynamic memory oracle

“Jon yelled at Mary but Mary couldn’t hear him”
dynamic memory oracle

Remove errors for words that just occurred within the last n characters. (n = 100, 500, 1000, 5000)

Up to 500 memory

circular, memorandum, or report, skillfully, pointedly, and elegantly. Bilibin's services were valued not only for what he wrote, but also for his skill in dealing and conversing with those in the highest spheres. Bilibin liked conversation as he liked work, only when it could be made elegantly witty. In society he always awaited an opportunity to say something striking and took part in a conversation only when that was possible. His conversation was always sprinkled with wittily original,

6% of errors
rare words oracle

Remove errors for words that occur very infrequently in the training data \((n = 0\ldots5)\).

Less than 3 training examples of word

Nicholas and Sonya, the niece. Sonya was a slender little brunette with a tender look in her eyes which were veiled by long lashes, thick black plaits coiling twice round her head, and a tawny tint in her complexion and especially in the color of her slender but graceful and muscular arms and neck. By the grace of her movements, by the softness and flexibility of her small limbs, and by a certain coyness and reserve of

9% of errors
difficult next letter oracles

After space, quote, new line

"No, impossible!" said Prince Andrew, laughing and pressing Pierre's hand to show that there was no need to ask the question. He wished to

Anna Pavlovna smiled and promised to take Pierre in hand. She knew his father to be a connection of Prince Vasili's. The elderly lady who had been sitting with the old aunt rose hurriedly and overtook Prince Vasili

37% of errors
Bigger model makes 44K less errors: 36K of these are n-gram errors! (81%) 5K boost 3K distributed among rest

=> Scaling the model up gets rid of mainly n-gram errors and leaves the other error types almost untouched by comparison!
Conclusions

- LSTMs are powerful models and do learn interesting, interpretable, long-term interactions

- Limitations:
  - n-gram failures: fixable with scaling up the model
  - rare word failures: scale up data / transfer learning
  - dynamic memory errors: ??? (memory nets?)
  - word-level errors: hierarchies? clockwork RNN? not clear
Thank you!