Visualizing and Understanding Recurrent Neural Networks

Andrej Karpathy, Justin Johnson*, Fei-Fei Li*

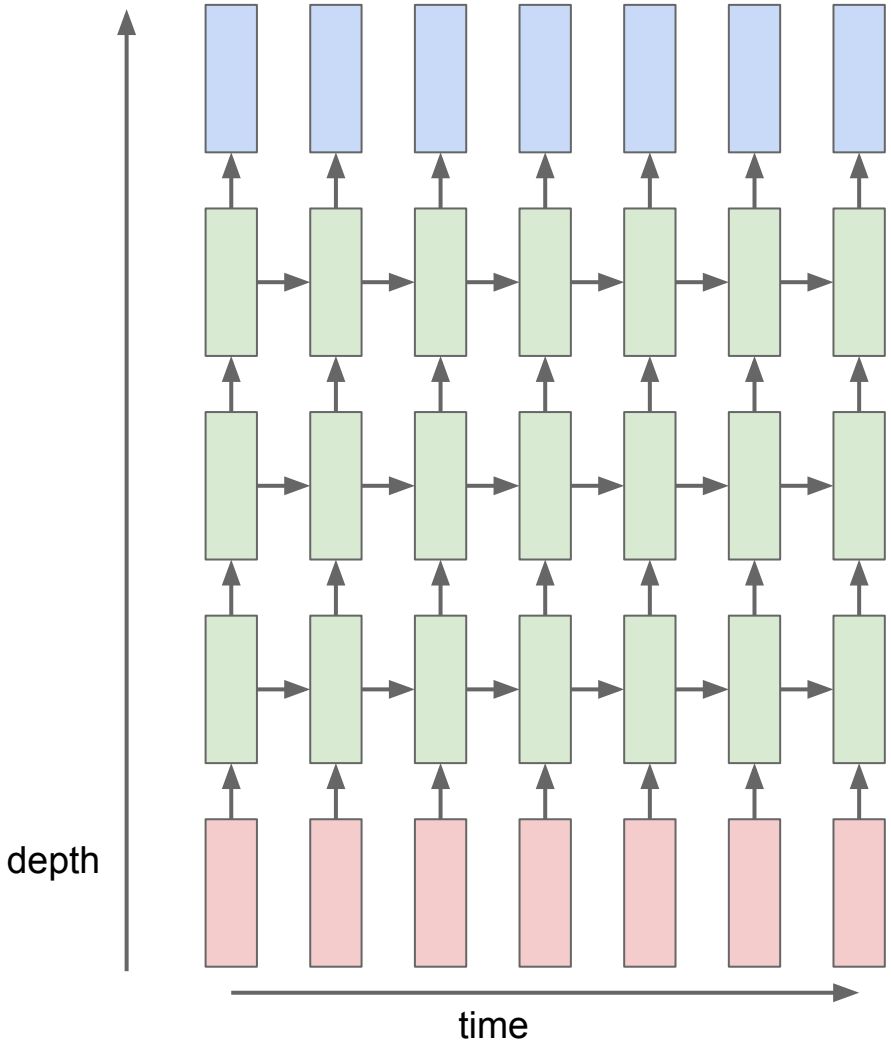
Presented by Fanyi Xiao

Slides from Andrej Karpathy

RNN:

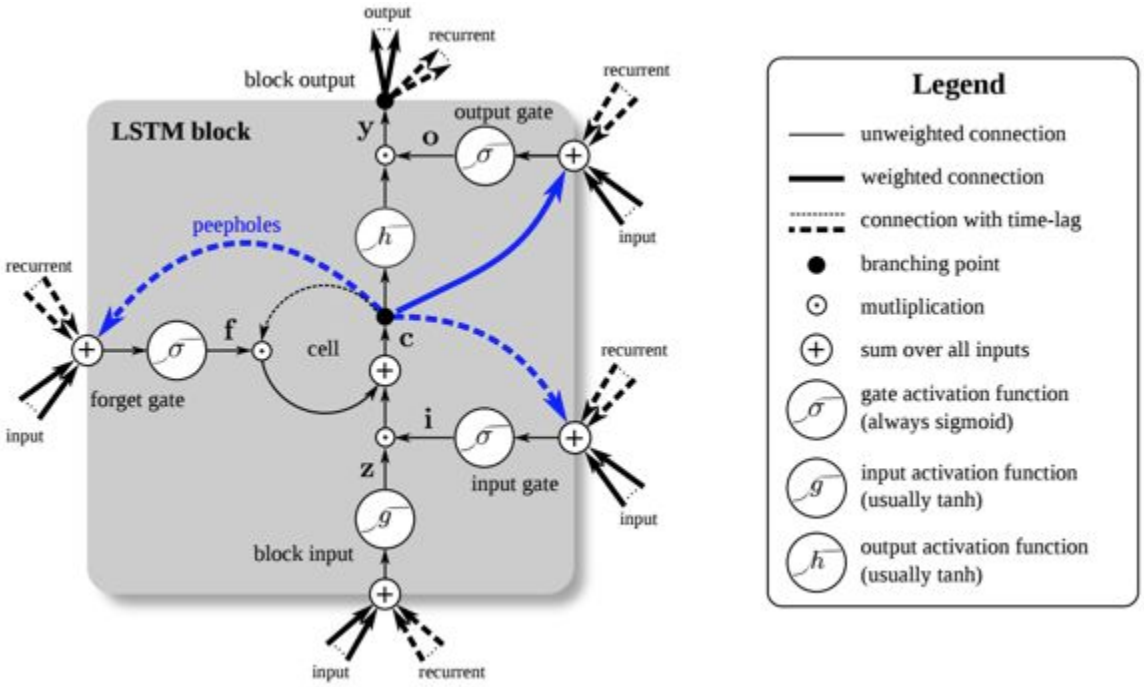
$$h_t^l = \tanh W^l \begin{pmatrix} h_t^{l-1} \\ h_{t-1}^l \end{pmatrix}$$

$h \in \mathbb{R}^n$ $\qquad W^l \ [n \times 2n]$



Slides from Andrej Karpathy

LSTM



:S

LSTM: A Search Space Odyssey, Greff et al.

Slides from Andrej Karpathy

RNN:

$$h_t^l = \tanh W^l \begin{pmatrix} h_t^{l-1} \\ h_{t-1}^l \end{pmatrix}$$

$h \in \mathbb{R}^n$ $\qquad$ $W^l \ [n \times 2n]$

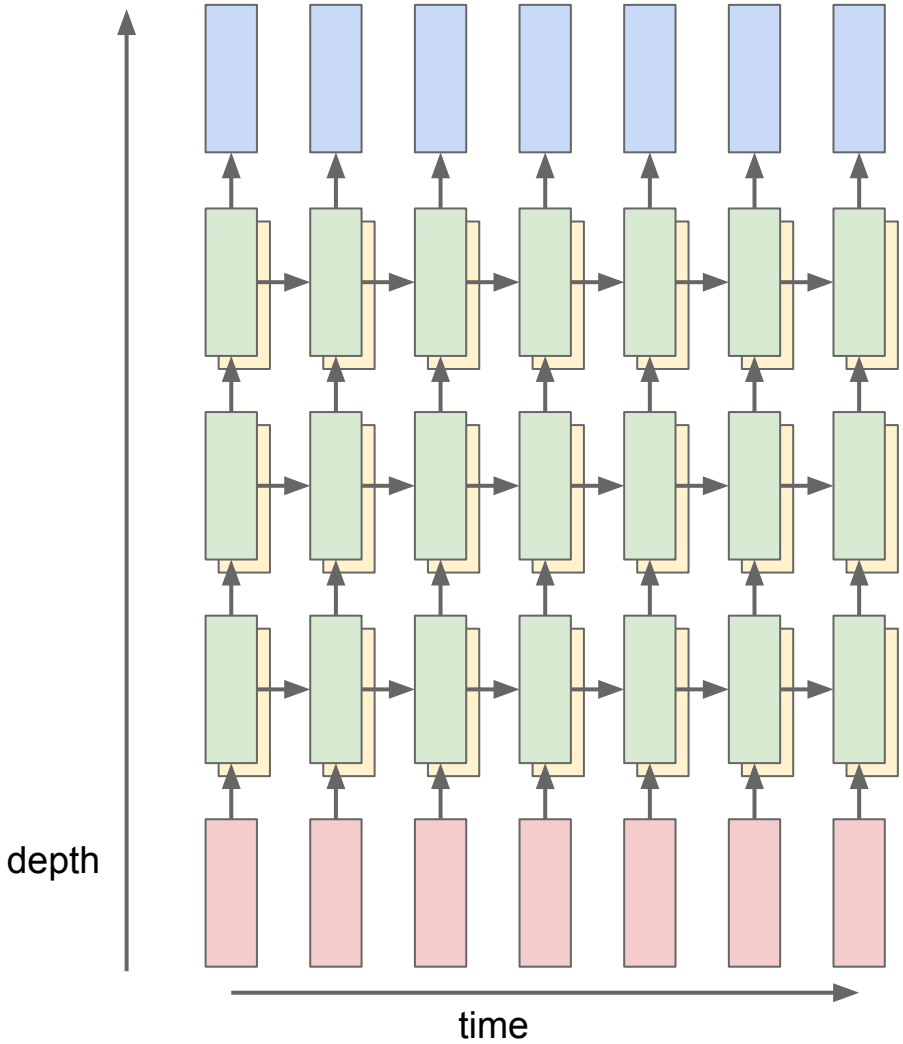
LSTM:

$W^l \ [4n \times 2n]$

$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \text{sigm} \\ \text{sigm} \\ \text{sigm} \\ \tanh \end{pmatrix} W^l \begin{pmatrix} h_t^{l-1} \\ h_{t-1}^l \end{pmatrix}$$

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

$$h_t^l = o \odot \tanh(c_t^l)$$



Slides from Andrej Karpathy

Datasets

Leo Tolstoy's
"War and Peace"

Linus Torvald's
"Linux Kernel"

structure

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 monstrosity 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.

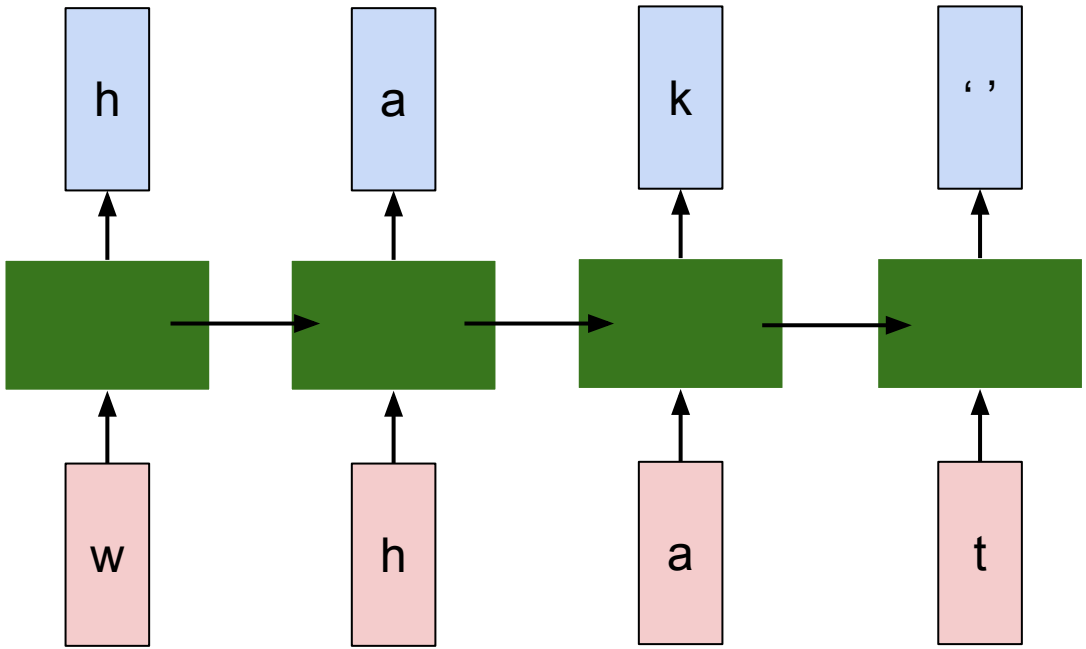
The visitor, compelled to look on at this family scene, thought it necessary to take some part in it.

"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.

```
static ssize_t
sched_feat_write(struct file *filp, const char __user *ubuf,
                 size_t cnt, loff_t *ppos)
{
        char buf[64];
        char *cmp;
        int i;
        struct inode *inode;

        if (cnt > 63)
                cnt = 63;

        if (copy_from_user(&buf, ubuf, cnt))
                return -EFAULT;

        buf[cnt] = 0;
        cmp = strstrip(buf);

        /* Ensure the static_key remains in a consistent state */
        inode = file_inode(filp);
        mutex_lock(&inode->i_mutex);
        i = sched_feat_set(cmp);
        mutex_unlock(&inode->i_mutex);
        if (i == __SCHED_FEAT_NR)
                return -EINVAL;

        *ppos += cnt;

        return cnt;
}
```

Slides from Andrej Karpathy



Character-level language modelling

Slides from Andrej Karpathy

Step 1: Performance

Model \ n	1	2	3	4	5	6	7	8	9	20
War and Peace Dataset										
n-gram	2.399	1.928	1.521	1.314	1.232	1.203	**1.194**	1.194	1.194	1.195
n-NN	2.399	1.931	1.553	1.451	1.339	**1.321**	-	-	-	-
Linux Kernel Dataset										
n-gram	2.702	1.954	1.440	1.213	1.097	1.027	0.982	0.953	0.933	**0.889**
n-NN	2.707	1.974	1.505	1.395	**1.256**	1.376	-	-	-	-

Slides from Andrej Karpathy

Step 1: Performance

Model \ n	1	2	3	4	5	6	7	8	9	20
War and Peace Dataset										
n-gram	2.399	1.928	1.521	1.314	1.232	1.203	**1.194**	1.194	1.194	1.195
n-NN	2.399	1.931	1.553	1.451	1.339	**1.321**	-	-	-	-
Linux Kernel Dataset										
n-gram	2.702	1.954	1.440	1.213	1.097	1.027	0.982	0.953	0.933	**0.889**
n-NN	2.707	1.974	1.505	1.395	**1.256**	1.376	-	-	-	-

	LSTM			**RNN**			**GRU**		
Layers	1	2	3	1	2	3	1	2	3
Size	War and Peace Dataset								
64	1.449	1.442	1.540	1.446	1.401	1.396	1.398	**1.373**	1.472
128	1.277	**1.227**	1.279	1.417	1.286	1.277	1.230	**1.226**	1.253
256	1.189	**1.137**	1.141	1.342	1.256	1.239	1.198	1.164	**1.138**
512	1.161	1.092	1.082	-	-	-	1.170	1.201	**1.077**
	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**

Slides from Andrej Karpathy

Step 1: Performance

Model \ n	1	2	3	4	5	6	7	8	9	20
War and Peace Dataset										
n-gram	2.399	1.928	1.521	1.314	1.232	1.203	**1.194**	1.194	1.194	1.195
n-NN	2.399	1.931	1.553	1.451	1.339	**1.321**	-	-	-	-
Linux Kernel Dataset										
n-gram	2.702	1.954	1.440	1.213	1.097	1.027	0.982	0.953	0.933	**0.889**
n-NN	2.707	1.974	1.505	1.395	**1.256**	1.376	-	-	-	-

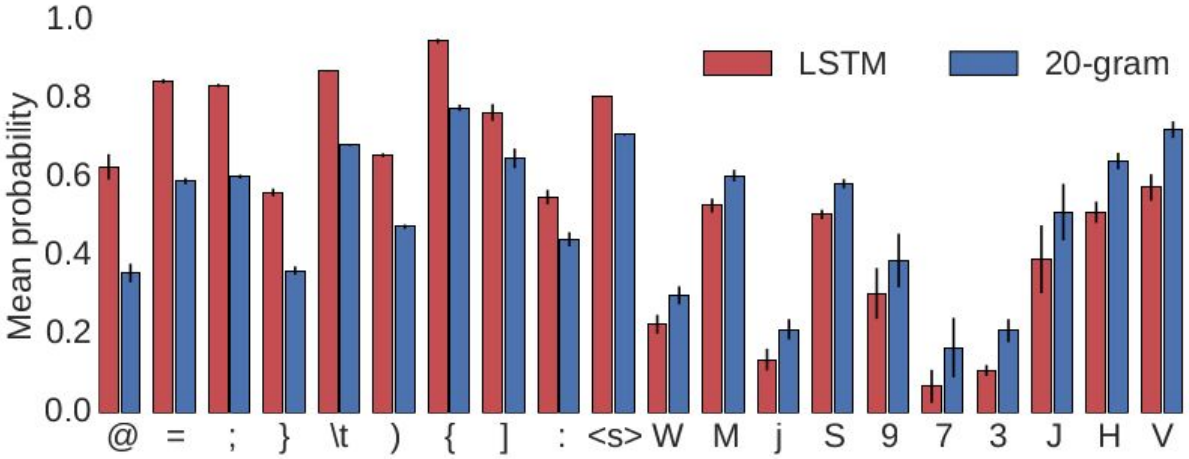
	LSTM			**RNN**			**GRU**		
Layers	1	2	3	1	2	3	1	2	3
Size	War and Peace Dataset								
64	1.449	1.442	1.540	1.446	1.401	1.396	1.398	**1.373**	1.472
128	1.277	**1.227**	1.279	1.417	1.286	1.277	1.230	**1.226**	1.253
256	1.189	**1.137**	1.141	1.342	1.256	1.239	1.198	1.164	**1.138**
512	1.161	1.092	1.082	-	-	-	1.170	1.201	1.077
	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**

3GB models

11MB models

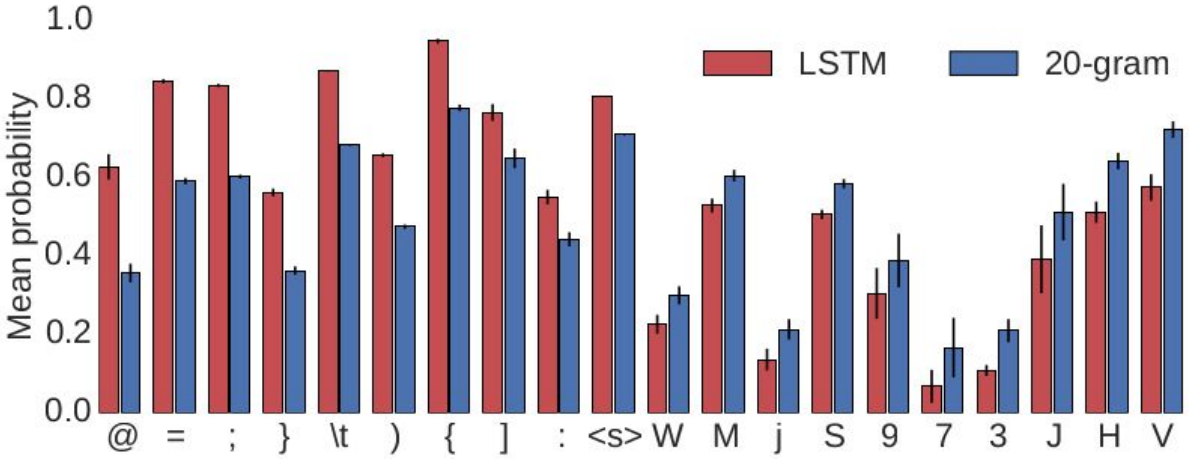
Slides from Andrej Karpathy

Linux

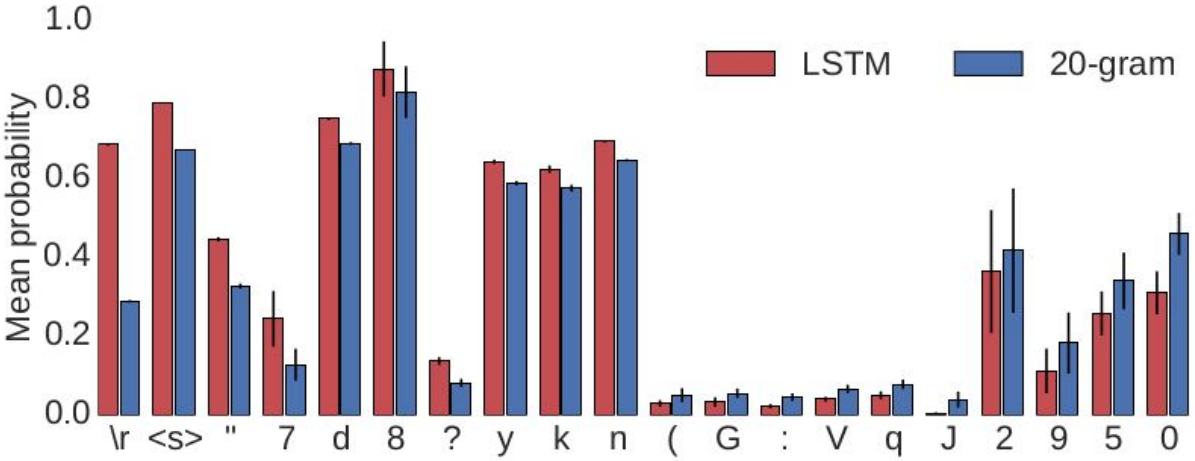


Slides from Andrej Karpathy

Linux

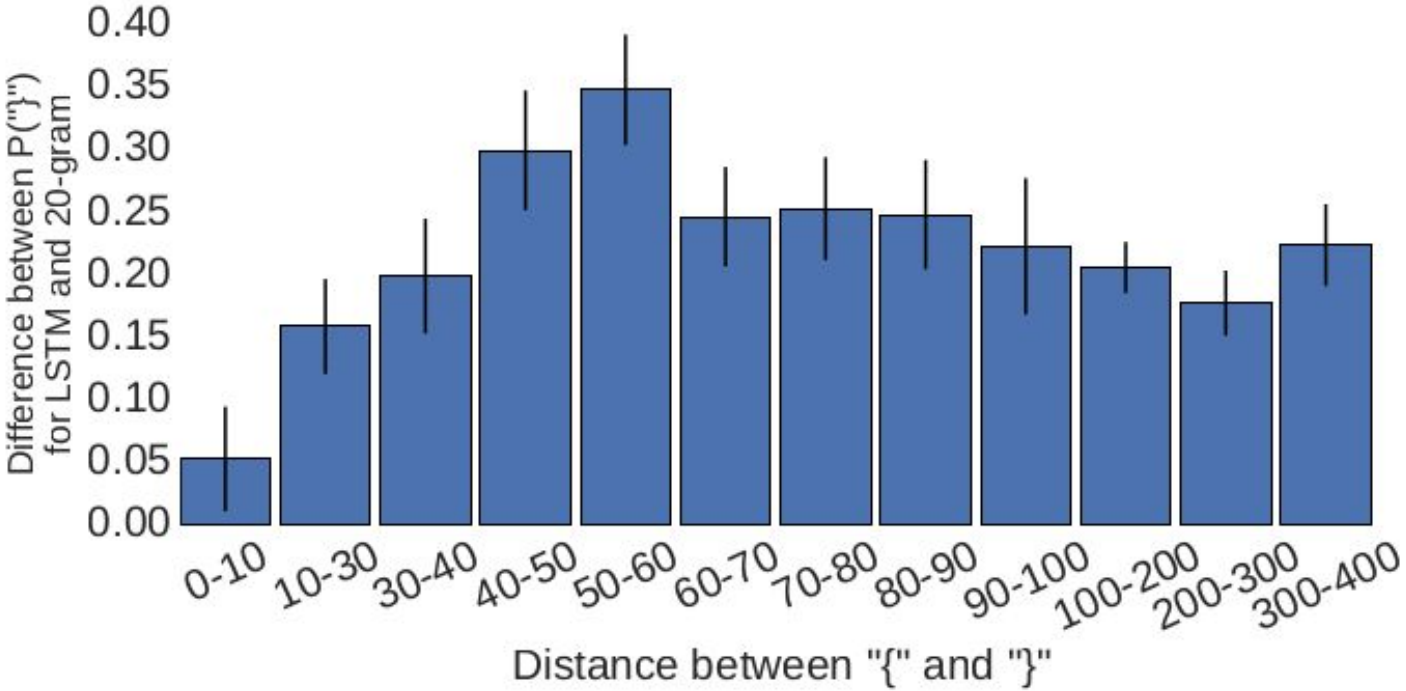


War and Peace



Slides from Andrej Karpathy

Closing brace (“}”) case study



Slides from Andrej Karpathy

Searching for interpretable cells

$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \text{sigm} \\ \text{sigm} \\ \text{sigm} \\ \tanh \end{pmatrix} W^l \begin{pmatrix} h_t^{l-1} \\ h_{t-1}^l \end{pmatrix}$$

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

$$h_t^l = o \odot \tanh(c_t^l)$$

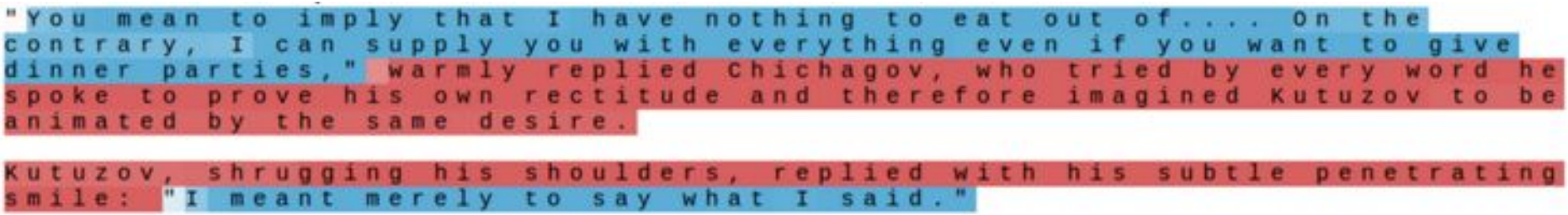
Slides from Andrej Karpathy

Searching for interpretable cells

```
/* Unpack a filter field's string representation from user-space
 * buffer. */
char *audit_unpack_string(void **bufp, size_t *remain, size_t len)
{
  char *str;
  if (!*bufp || (len == 0) || (len > *remain))
    return ERR_PTR(-EINVAL);
  /* Of the currently implemented string fields, PATH_MAX
   * defines the longest valid length.
   */
```

Slides from Andrej Karpathy

Searching for interpretable cells



quote detection cell

Slides from Andrej Karpathy

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

Slides from Andrej Karpathy

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.

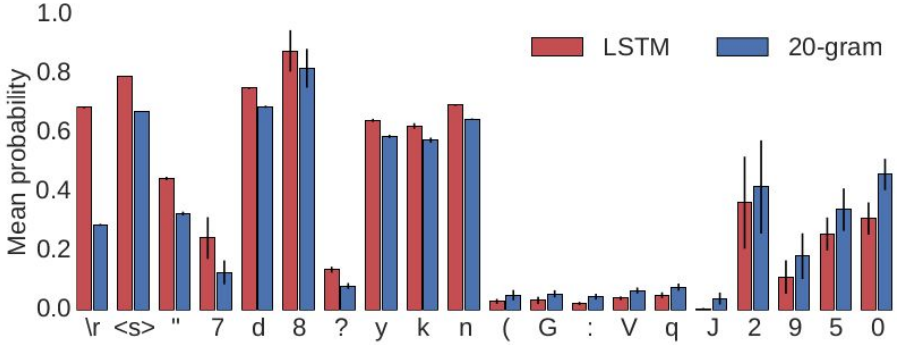
line length tracking cell

Slides from Andrej Karpathy

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.



Slides from Andrej Karpathy

Searching for interpretable cells

```
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

Slides from Andrej Karpathy

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

Slides from Andrej Karpathy

Searching for interpretable cells

```
#ifdef CONFIG_AUDITSYSCALL
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

Slides from Andrej Karpathy

Searching for interpretable cells

```
char *audit_unpack_string(void **bufp, size_t *remain, si
{
  char *str;
  if (!*bufp || (len == 0) || (len > *remain))
    return ERR_PTR(-EINVAL);
  /* Of the currently implemented string fields, PATH_MAX
   * defines the longest valid length.
   */
  if (len > PATH_MAX)
    return ERR_PTR(-ENAMETOOLONG);
  str = kmalloc(len + 1, GFP_KERNEL);
  if (unlikely(!str))
    return ERR_PTR(-ENOMEM);
  memcpy(str, *bufp, len);
  str[len] = 0;
  *bufp += len;
  *remain -= len;
  return str;
}
```

something interesting cell
(not quite sure what)

Slides from Andrej Karpathy

Learning Dynamics

training iterations

```
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
```

```
"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."
```

```
we counter. He stutn co des. His stanted out one ofler that concossions and was
to gearang reay Jotrets and with fre colt otf paitt thin wall. Which das stimn
```

```
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.
```

```
"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 aking his soul came to the packs and drove up his father-in-law women.
```

Slides from Andrej Karpathy

Learning Dynamics

training iterations

```
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
```

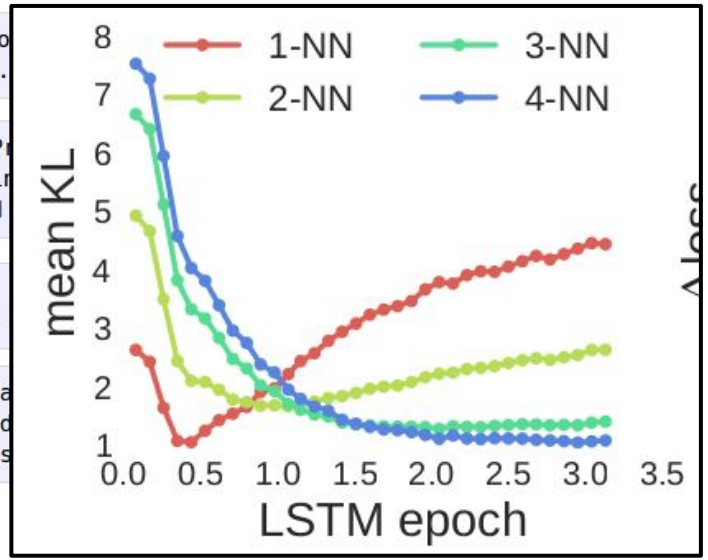
```
"Tmont thithey" fomesscerliund
Keushey. Thom here
sheulke, anmerenith ol sivh I lalterthend Bleipile shuwy fil on aseterlome
coaniogennc Phe lism thond hon at. MeiDimorotion in ther thize."
```

```
we counter. He stutn co des. His stanted out one ofler that co
to gearang reay Jotrets and with fre colt otf paitt thin wall.
```

```
Aftair fall unsuch that the hall for Pr
her hearly, and behs to so arwage fivir
how, and Gogition is so overelical 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, fed
Pierre aking his soul came to the packs
```



Slides from Andrej Karpathy

Learning Dynamics

training iterations

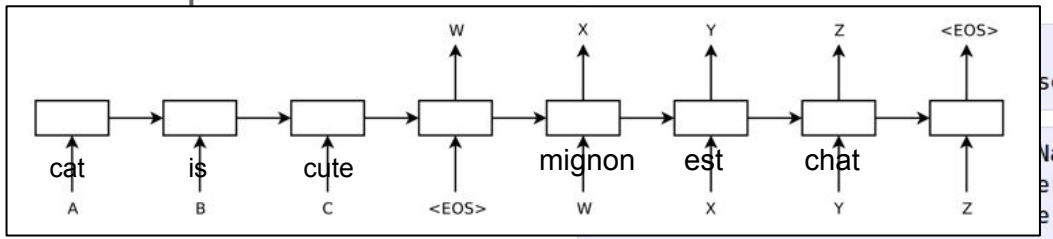
```
tyntd-iafhatawiaoihrdemot  lytdws  e ,tfti, astai f ogoh eoase rrranbyne 'nhthnee e
plia tklrgd t o idoe ns,smtt   h ne etie h,hregtrs nigtike,aoaenns lng
```

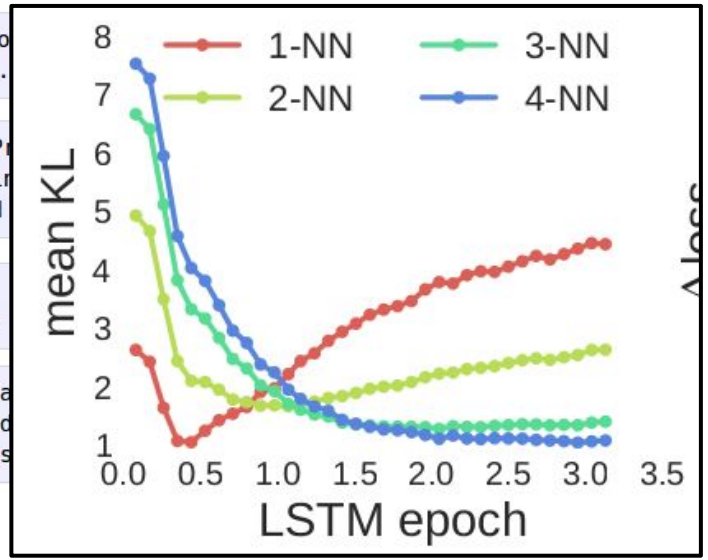
```
"Tmont thithey" fomesscerliund
Keushey. Thom here
sheulke, anmerenith ol sivh I lalterthend Bleipile shuwy fil on aseterlome
coaniogennc Phe lism thond hon at. MeiDimorotion in ther thize."
```

```
we counter. He stutn co des. His stanted out one ofler that co
to gearang reay Jotrets and with fre colt otf paitt thin wall.
```

```
Aftair fall unsuch that the hall for Pr
her hearly, and behs to so arwage fivir
how, and Gogition is so overelical and
```

cat is cute → chat est mignon





Sequence to Sequence Learning with Neural Networks, Sutskever et al.

Slides from Andrej Karpathy

Error Analysis

Slides from Andrej Karpathy

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 ngram model, for n = 1 .. 9

1-gram to 5-gram

```
"My wife," continued Prince Andrew, "is an excellent woman, one of those
rare women with whom a man's honor is safe; but, O God, what would I not
give now to be unmarried! You are the first and only one to whom I
mention this, because I like you."
```

18% of errors

Slides from Andrej Karpathy

dynamic memory oracle

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

Slides from Andrej Karpathy

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

Slides from Andrej Karpathy

rare words oracle

Remove errors for words that occur very infrequently in the training data (n = 0...5).

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

Slides from Andrej Karpathy

difficult next letter oracles

After space, quote, new line

After space or quote

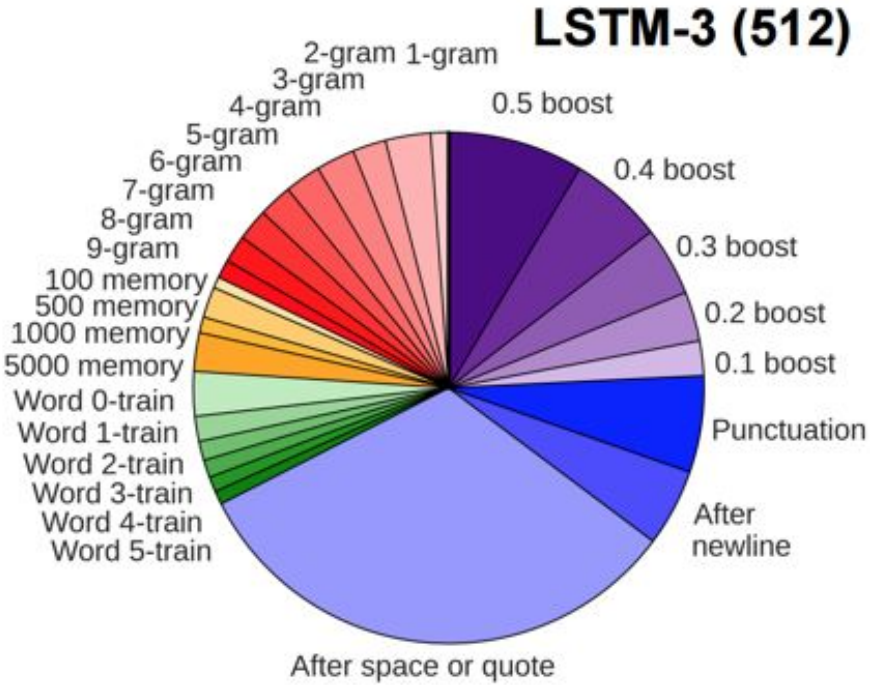
"No, impossible!" said Prince Andrew, laughing and pressing Pierre's hand to show that there was no need to ask the question. He wished to

After newline

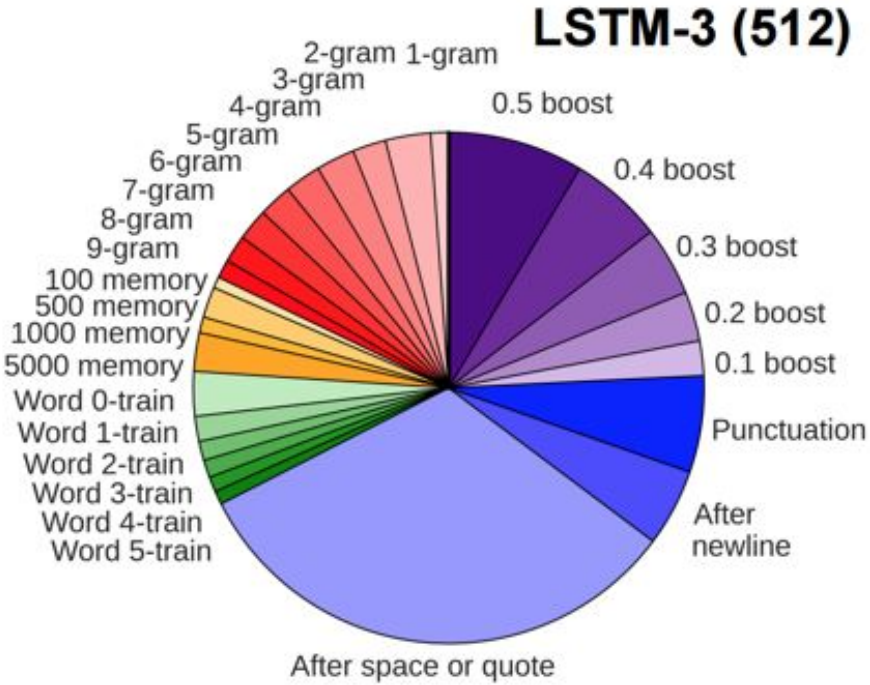
Anna Pavlovna smiled and promised to take Pierre in hand. She knew his father to be a connection of Prince Vasili's. The elderly lady who had been sitting with the old aunt rose hurriedly and overtook Prince Vasili

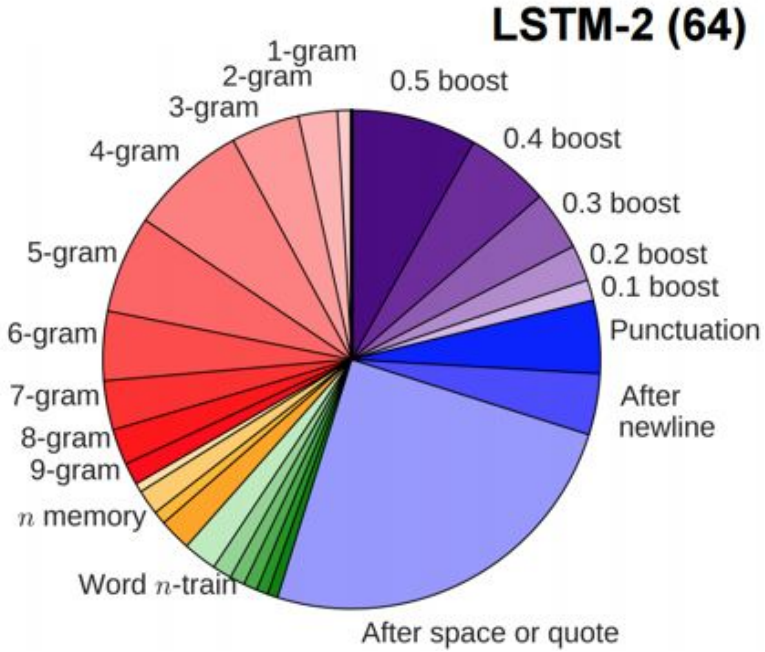
37% of errors

Slides from Andrej Karpathy

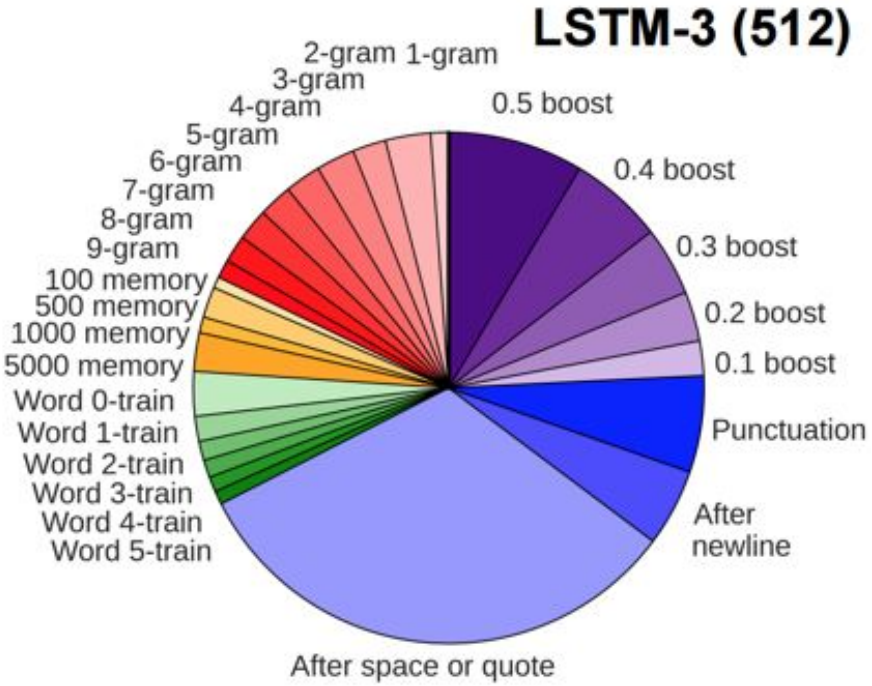


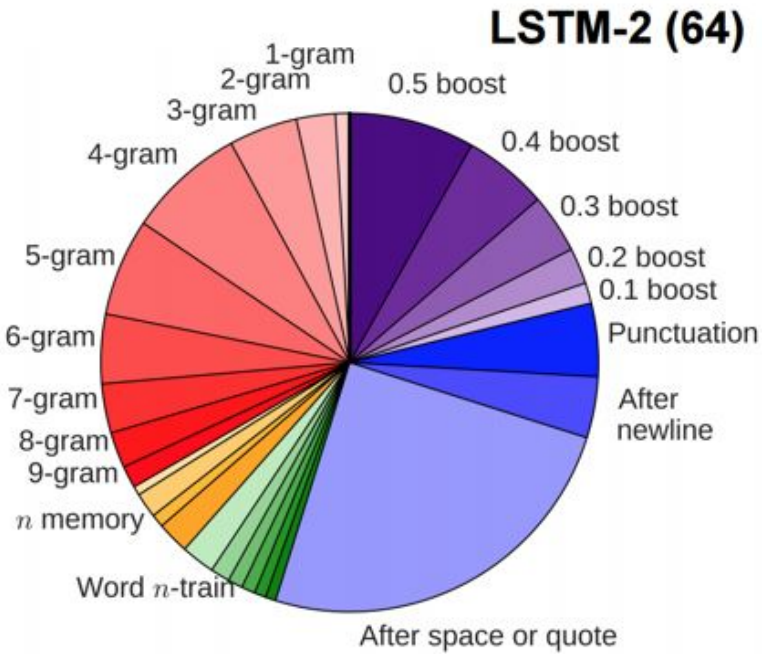
Slides from Andrej Karpathy





Slides from Andrej Karpathy





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!

Slides from Andrej Karpathy

Conclusions

- LSTMs are powerful models and **do** learn interesting, <u>interpretable</u>, <u>long-term</u> 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

Slides from Andrej Karpathy

Thank you!