Language, Vision and Sequences

EECS 442 – David Fouhey
Winter 2022, University of Michigan

http://web.eecs.umich.edu/~fouhey/teaching/EECS442_W22/
Quick – what’s this?

Dog image credit: T. Gupta
Previously on EECS 442

Feature vector from image

<table>
<thead>
<tr>
<th>Cat weight vector</th>
<th>Dog weight vector</th>
<th>Hat weight vector</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>-0.5</td>
<td>0.1</td>
</tr>
<tr>
<td>1.5</td>
<td>1.3</td>
<td>2.1</td>
</tr>
<tr>
<td>0.0</td>
<td>0.3</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Weight matrix $W$ a collection of scoring functions, one per class

$W x_i$ where $j$th component is “score” for $j$th class.

Diagram by: Karpathy, Fei-Fei
Previously on EECS 442

Converting Scores to “Probability Distribution”

<table>
<thead>
<tr>
<th></th>
<th>Cat score</th>
<th>Dog score</th>
<th>Hat score</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.9</td>
<td>e^{-0.9}</td>
<td>e^{0.6}</td>
<td>e^{0.4}</td>
</tr>
<tr>
<td>0.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td>0.41</td>
<td>1.82</td>
<td>1.49</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.11</td>
<td>0.49</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td>P(cat)</td>
<td>P(dog)</td>
<td>P(hat)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>∑=3.72</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Generally P(class j): \[ \frac{\exp((Wx)_j)}{\sum_k \exp((Wx)_k)} \]
What’s a Big Issue?

Is it a dog? Is it a hat?
Diagram by: Karpathy, Fei-Fei
**Take 2**

Converting Scores to “Probability Distribution”

<table>
<thead>
<tr>
<th></th>
<th>Cat score</th>
<th>Dog score</th>
<th>Hat score</th>
<th>$\text{sgm}(x)$</th>
<th>$P(\text{cat})$</th>
<th>$P(\text{dog})$</th>
<th>$P(\text{hat})$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-1.9</td>
<td>1.2</td>
<td>0.9</td>
<td></td>
<td>0.13</td>
<td>0.77</td>
<td>0.71</td>
</tr>
</tbody>
</table>

77% dog  
71% hat  
13% cat?
Hmm…

- We’d like to say: “dog with a hat” or “husky wearing a hat” or something else.
- Naïve approach (given N words to choose from and up to C words). **How many?**
  \[ \sum_{i=1}^{C} N^i \] classes to choose from (~N^i)
- N=10k, C=5 -> 100 billion billion
- Can’t train 100 billion billion classifiers
Hmm...

- Pick N-word dictionary, call them class 1, ..., N
- New goal: emit sequence of $C N$-way classification outputs
- Dictionary could be:
  - All the words that appear in training set
  - All the ascii characters
  - Typically includes special “words”: START, END, UNK
VizWiz: Nearly Real-Time Answers to Visual Questions, Bigham et al., UIST 2010
VizWiz

Really interesting area. See Captioning Images Taken by People Who Are Blind, Gurari et al. ECCV 2020 and vizwiz.org
Option 1 – Sequence Modeling

Output at $i$ is linear transformation of hidden state

$$y_i = W_{yh} h_i$$

Hidden state at $i$ is linear function of previous hidden state and input at $i$, + nonlinearity

$$h_i = \sigma(W_{hx}x_i + W_{hh}h_{i-1})$$
Option 1 – Sequence Modeling

Can stack arbitrarily to create a function of multiple inputs with multiple outputs that’s in terms of parameters $W_{HX}, W_{HH}, W_{YH}$

$$y_i = W_{yh} h_i$$

$$h_i = \sigma(W_{hx} x_i + W_{hh} h_{i-1})$$
Option 1 – Sequence Modeling

Can define a loss with respect to each output and differentiate wrt to all the weights

Backpropagation through time

\[ y_i = W_{yh} h_i \]
\[ h_i = \sigma(W_{hx} x_i + W_{hh} h_{i-1}) \]
Captioning

\[ f_{im} \in \mathbb{R}^{4096} \]

START

Dog

in

a

hat

END

\[ h_0 \rightarrow h_1 \rightarrow h_2 \rightarrow h_3 \rightarrow h_4 \rightarrow h_5 \]

\[ x_1 \rightarrow x_2 \rightarrow x_3 \rightarrow x_4 \rightarrow x_5 \]
Captioning

Each step: look at input and hidden state (more on that in a second) and decide output. Can learn through CNN!
Results

A female tennis player in action on the court.

A group of young men playing a game of soccer.

A man riding a wave on top of a surfboard.

A baseball game in progress with the batter up to plate.

A brown bear standing on top of a lush green field.

A person holding a cell phone in their hand.

Results

A close up of a person brushing his teeth.

A woman laying on a bed in a bedroom.

A black and white cat is sitting on a chair.

A large clock mounted to the side of a building.

A bunch of fruit that are sitting on a table.

A toothbrush holder sitting on top of a white sink.

Captioning – Looking at Each Step

Why might this be better than doing billions of classification problems?
What Goes On Inside?

• Great repo for playing with RNNs (Char-RNN)
  • https://github.com/karpathy/char-rnn
  • (Or search char-rnn numpy)
• Tokens are just the characters that appear in the training set

Result credit: A. Karpathy
/*
 * If this error is set, we will need anything right after that BSD.
 */

static void action_new_function(struct s_stat_info *wb)
{
    unsigned long flags;
    int lel_idx_bit = e->edd, *sys & ~((unsigned long) *FIRST_COMPAT);
    buf[0] = 0xFFFFFFFF & (bit << 4);
    min(inc, slist->bytes);
    printk(KERN_WARNING "Memory allocated %02x/%02x, "
        "original MLL instead\n"),
        min(min(multi_run - s->len, max) * num_data_in),
        frame_pos, sz + first_seg);
    div_u64_w(val, inb_p);
    spin_unlock(&disk->queue_lock);
    mutex_unlock(&s->sock->mutex);
    mutex_unlock(&func->mutex);
    return disassemble(info->pending_bh);
}

Result credit: A. Karpathy
Sample Trained on Names


Result credit: A. Karpathy
What Goes on Inside

Outputs of an RNN. Blue to red show timesteps where a given cell is active. What’s this?

Result credit: A. Karpathy
What Goes on Inside

Outputs of an RNN. Blue to red show timesteps where a given cell is active. **What’s this?**

```c
#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;
}
#endif
```

Result credit: A. Karpathy
What Goes on Inside

Outputs of an RNN. **Blue** to **red** show timesteps where a given cell is active. **What’s this?**

```c
/* Unpack a filter field's string representation from user-space */
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. */
```
What happens to really deep networks?
Remember $g^n$ for $g \neq 1$
Gradients explode / vanish
Nagging Detail #1 – Depth

• Typically use more complex methods that better manage gradient flowback (LSTM, GRU)

• General strategy: pass the hidden state to the next timestep as unchanged as possible, only adding updates as necessary
Nagging Detail #2

Lots of captions are in principle possible!

- A dog in a hat
- A dog wearing a hat
- Husky wearing a hat
- Husky holding a camera, sitting in grass
- A dog that’s in a hat, sitting on a lawn with a camera
Nagging Detail #2 – Sampling

Dog (P=0.3), A (P=0.2), Husky (P=0.15), ….

- Pick proportional to probability of each word
- Can adjust “temperature” parameter $\exp(\text{score}/t)$ to equalize probabilities
- $\exp(5) / \exp(1) \rightarrow 54.6$
- $\exp(5/5) / \exp(1/5) \rightarrow 2.2$
Effect of Temperature

• Train on essays about startups and investing

• Normal Temperature: “The surprised in investors weren’t going to raise money. I’m not the company with the time there are all interesting quickly, don’t have to get off the same programmers. There’s a super-angel round fundraising, why do you can do.”

• Low temperature: “is that they were all the same thing that was a startup is that they were all the same thing that was a startup is that they were all the same thing that was a startup is that they were all the same”
Nagging Detail #2 – Sampling

\[ h_0 \xrightarrow{} h_1 \xrightarrow{} h_2 \]

Dog (P=0.4), Husky (P=0.3), ….

\[ f_{im} \in \mathbb{R}^{4096} \]
Nagging Detail #2 – Sampling

\[ h_0 \xrightarrow{} h_1 \xrightarrow{} h_2 \]

Dog wearing (P=0.5), in (P=0.3), …. Each evaluation gives \( P(w_i|w_1,\ldots,w_{i-1}) \)

Can expand a finite tree of possibilities (beam search) and pick most likely sequence

\[ f_{im} \in \mathbb{R}^{4096} \]
Nagging Detail #3 – Evaluation

Computer: “A husky in a hat”
Human: “A dog in a hat”

How do you decide?

1) Ask humans. **Why might this be an issue?**
2) In practice: use something like precision (how many generated words appear in ground-truth sentences) or recall. Details very important to prevent gaming (e.g., “A a a a a”)
More General Sequence Models

Can have multiple inputs, single output

Positive Review

\[ h_0 \rightarrow h_1 \rightarrow h_2 \rightarrow h_3 \rightarrow h_4 \rightarrow h_5 \]

\[ x_1 \rightarrow x_2 \rightarrow x_3 \rightarrow x_4 \rightarrow x_5 \]

I loved my meal here
More General Sequence Models

Could be a feature vector!

h_0 \rightarrow h_1 \rightarrow h_2 \rightarrow h_3 \rightarrow h_4 \rightarrow h_5

x_1 \rightarrow x_2 \rightarrow x_3 \rightarrow x_4 \rightarrow x_5

What \rightarrow is \rightarrow the \rightarrow dog \rightarrow wearing
More General Models
Visual Question-Answering

What color are her eyes? What is the mustache made of?

How many slices of pizza are there? Is this a vegetarian pizza?

Is this person expecting company? What is just under the tree?

Does it appear to be rainy? Does this person have 20/20 vision?

Top-Performing Methods

Top methods now look at objects in the image as opposed to one big image vector.

Two men playing frisbee in a dark field.
Top-Performing Methods

Let’s Revisit A Number

• How many 20-word sentences with a vocabulary of 10k words are there really?
• Is it really \((10k)^20\)? **Why not?**
• Let’s look at some giraffes (I swear this is relevant)

Giraffe example credit: L. Zitnick
What do Giraffes Do All Day?

A giraffe grazing in its enclosure

A giraffe sitting and resting

A giraffe wandering around

A giraffe grazing in its enclosure

With apologies to both giraffes and people who study giraffes, I’m sure they’re fascinating
Alternate Idea – Retrieval

$\mathbb{R}^{4096}$

New “test” image

Training images + captions

“Giraffe that’s outstanding in its field”

“Giraffe sitting & relaxing”
Alternate Idea – Retrieval

\[ \mathbb{R}^{4096} \]

Training images + captions

“Giraffe that’s outstanding in its field”

“Giraffe sitting & relaxing”

“Giraffe sitting & relaxing”
Retrieval Results

A man riding a wave on a surfboard.

A man riding a wave on a surfboard in the ocean.

A person flying a kite in the sky.

A person flying a kite in the sky.

A cat sitting in a bathroom sink.

A black and white cat sitting in a bathroom sink.

Exploring Nearest Neighbor Approaches for Image Captioning. Devlin et al. 2015
Retrieval Results

A wooden bench in front of a building.

A window display on the front of a building.

A building with a clock on the top.

A clock tower on the top of a building.

The side of a passenger train at a train station.

A bus that is on the side of a road.

Exploring Nearest Neighbor Approaches for Image Captioning. Devlin et al. 2015
Retrieval Results

• In practice: humans don’t like retrieved captions as much
• Can’t generate anything new!
• Works well, but not as well as good systems
Latest in This Space

• Very Quick summary
• Great resources include
  • EECS 487 (Lu Wang)
  • EECS 498 (Justin Johnson)
  • Transformers from scratch
    http://peterbloem.nl/blog/transformers
Suppose you’re in charge of the hidden state. What’s the problem?

Have to keep track of everything! Very tedious if the sequence is long.

Diagram is remake of one from J. Johnson
Self-Attention

Output: $\sum_i a_i h_i$

Attention: softmax over $e_i$

Alignment Score: $e_i = h_i^T s_0$

When do we eat

Diagram is remake of one from J. Johnson
When do we eat?

Output given to decoder that handles making predictions.

Diagram is remake of one from J. Johnson
Transformers

Old Output: $\sum_i a_i h_i$

New Output: $\sum_i a_i (Vh_i)$

Attention: softmax over $e_i$

New Alignment: $e_i = h_i^T s_0$

Old Alignment: $e_i = h_i^T K s_0$

Vector doesn’t do it all

New: scale dot product between F-D vectors by $\sqrt{F}$

Diagram is remake of one from J. Johnson
Transformers

Does the alignment score know about position?

Alignment Score: $e_i = h_i^T s_0$

Solution: add vectors that are function of location

Diagram is remake of one from J. Johnson
Positional Encodings

Set of sinusoids of different frequencies. Really effective trick for encoding locations for networks

\[ f(p, 2i) = \sin \left( \frac{p}{10000^{2i/d}} \right) \]
\[ f(p, 2i + 1) = \cos \left( \frac{p}{10000^{2i/d}} \right) \]

Diagram: Amirhoseein Kazemnejad, https://kazemnejad.com/blog/transformer_architecture_positional_encoding/
Large-scale Language Models

• Can train these models at huge scale
• GPT-3:
  • 410 billion tokens of training data
  • 17 billion parameters
Positional Encodings Elsewhere

Learn network mapping from \((x,y) \rightarrow (r,g,b)\)

Learned on \((x,y)\) w/positional encodings

How Might We Use a Transformer?

Let’s plug an image into a transformer

Break into patches, treat like words
Vision Transformer

Key idea: put in sequence of image tokens
Why is [person4] pointing at [person1]?
VCR

Why is [person4] pointing at [person1]? 

- a) He is telling [person3] that [person1] ordered the pancakes.
- b) He just told a joke.
- c) He is feeling accusatory towards [person1].
- d) He is giving [person1] directions.

RedCaps

RedCaps: web-curated image-text data created by the people, for the people. Desai et al. NeurIPS 2021.
Some Concluding Thoughts

• Getting this right is really hard!
• Deep learning is trying to do solve any problem you pose with as little effort as possible.
• A lot of this has to do with data and people
Novel Captions

Deep Compositional Captioning: Describing Novel Object Categories without Paired Training Data.
L. Hendricks et al. CVPR 2016
Simple Baseline for VQA

- Construct a vocabulary of 5000 most frequent answers
- Extract all the information from the image, $I$
  - Construct an image representation using a CNN
- Represent the question, $Q$ with BoW
- Compute distribution of answers, $P(A|Q,I)$


Slide credit: T. Gupta
Qualitative Results

**Question:** what are they doing

**Predictions:**
- playing baseball (score: 10.67 = 2.01 [image] + 8.66 [word])
- baseball (score: 9.65 = 4.84 [image] + 4.82 [word])
- grazing (score: 9.34 = 0.53 [image] + 8.81 [word])

Based on image only: umpire (4.85), baseball (4.84), batter (4.46)

Based on word only: playing wii (10.62), eating (9.97), playing frisbee (9.24)

**Question:** how many people inside

**Predictions:**
- 3 (score: 13.39 = 2.75 [image] + 10.65 [word])
- 2 (score: 12.76 = 2.49 [image] + 10.27 [word])
- 5 (score: 12.72 = 1.83 [image] + 10.89 [word])

Based on image only: umpire (4.85), baseball (4.84), batter (4.46)

Based on word only: 8 (11.24), 7 (10.95), 5 (10.89)

Slide credit: T. Gupta
Qualitative Results

**Question:** which brand is the laptop

**Predictions:**
- apple (score: 10.87 = 1.10 [image] + 9.77 [word])
- dell (score: 9.83 = 0.71 [image] + 9.12 [word])
- toshiba (score: 9.76 = 1.18 [image] + 8.58 [word])

Based on image only: books (3.15), yes (3.14), no (2.95)
Based on word only: apple (9.77), hp (9.18), dell (9.12)

- Language prior prunes the answer space significantly


Slide credit: T. Gupta
## Quantitative Evaluation

<table>
<thead>
<tr>
<th>Method</th>
<th>Open-Ended</th>
<th></th>
<th>Multiple-Choice</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Overall</td>
<td>yes/no</td>
<td>number</td>
<td>others</td>
</tr>
<tr>
<td>IMG [2]</td>
<td>28.13</td>
<td>64.01</td>
<td>00.42</td>
<td>03.77</td>
</tr>
<tr>
<td>BOW [2]</td>
<td>48.09</td>
<td>75.66</td>
<td>36.70</td>
<td>27.14</td>
</tr>
<tr>
<td>BOWIMG [2]</td>
<td>52.64</td>
<td>75.55</td>
<td>33.67</td>
<td>37.37</td>
</tr>
<tr>
<td>LSTMIMG [2]</td>
<td>53.74</td>
<td>78.94</td>
<td>35.24</td>
<td>36.42</td>
</tr>
<tr>
<td>CompMem [6]</td>
<td>52.62</td>
<td>78.33</td>
<td>35.93</td>
<td>34.46</td>
</tr>
<tr>
<td>NMN+LSTM [1]</td>
<td>54.80</td>
<td>77.70</td>
<td>37.20</td>
<td>39.30</td>
</tr>
<tr>
<td>WR Sel. [13]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ACK [16]</td>
<td>55.72</td>
<td>79.23</td>
<td>36.13</td>
<td>40.08</td>
</tr>
<tr>
<td>DPPnet [11]</td>
<td>57.22</td>
<td>80.71</td>
<td>37.24</td>
<td>41.69</td>
</tr>
<tr>
<td>iBOWIMG</td>
<td>55.72</td>
<td>76.55</td>
<td>35.03</td>
<td>42.62</td>
</tr>
</tbody>
</table>

Evaluated on the VQA dataset – although now results are quite a bit higher
Does the model learn to localize?

Class Activation Mapping applied to VQA Baseline

Slide credit: T. Gupta
Recent Developments

Can balance data to make things difficult

Who is wearing glasses?
- man
- woman

Where is the child sitting?
- fridge
- arms

Is the umbrella upside down?
- yes
- no

How many children are in the bed?
- 2
- 1

Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. Goyal et al. 2017