LSTMs Overview

Subhashini Venugopalan
Neural Networks

- Input
- Hidden
- Hidden
- Output

Diagram:

- $X_t$
- $A$
- $B$
- $Z_t$
WHY RNNs/LSTMs?

Can we operate over sequences of inputs?

Limitations of vanilla Neural Networks

Outputs a fixed size vector.

Performs a fixed number of computations (#layers).

Accepts only fixed size input e.g 224x224 images.
Recurrent Neural Networks

They are networks with loops. [Elman ‘90]

Image Credit: Chris Olah
Un-Roll The Loop

Each time step has a layer with the same weights.
The repeating layer/module is a sigmoid or a tanh.
Learns to model \( h_t \mid x_1, \ldots, x_{t-1} \)

Recurrent Neural Network “unrolled in time”

Image Credit: Chris Olah
Simple RNNs

\[ h_t = g(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \]
\[ z_t = g(W_{hz}h_t + b_z) \]

Image Credit: Chris Olah
Problems with Simple RNNs

- Can’t seem to handle “long-term dependencies” in practice
- Gradients shrink through the many layers (Vanishing Gradients)

[Hochreiter ‘91]
[Bengio et. al. ‘94]

Image Credit: Chris Olah
Long Short Term Memory (LSTMs)

[Hochreiter and Schmidhuber ‘97]
LSTM UNIT

Memory Cell: Core of the LSTM Unit
Encodes all inputs observed

[Hochreiter and Schmidhuber ‘97]
[Graves ‘13]
LSTM UNIT

Memory Cell:
Core of the LSTM Unit
Encodes all inputs observed

Gates:
Input, Output and Forget
Sigmoid $[0,1]$

[Hochreiter and Schmidhuber ‘97]
[Graves ‘13]
LSTM UNIT

\[ i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1}) \]
\[ f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1}) \]

Update the Cell state
\[ c_t = f_t \odot c_{t-1} + i_t \odot \phi(W_{xc}x_t + W_{hc}h_{t-1}) \]
Learns long-term dependencies

\[ o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1}) \]
\[ h_t = o_t \odot \phi(c_t) \]

[Hochreiter and Schmidhuber ‘97]
[Graves ‘13]
Can Model Sequences

- Can handle longer-term dependencies
- Overcomes Vanishing Gradients problem
- GRUs - Gated Recurrent Units is a much simpler variant which also overcomes these issues. [Cho et. al. ‘14]
Putting Things Together

Encode a sequence of inputs to a vector.

\((h_t \mid x_1, \ldots, x_{t-1})\)

Decode from the vector to a sequence of outputs.

\(Pr(x_t \mid x_1, \ldots, x_{t-1})\)
SOLVE A WIDER RANGE OF PROBLEMS

one to many

many to one

many to many

many to many

Image Captioning
Vinyals et. al. ‘15, Donahue et. al. ‘15

Activity Recognition
Donahue et. al. ‘15

Sequence to Sequence
Machine Translation
Speech Recognition
Video Description
VQA, POS tagging, ...

Sutskever et. al. ‘14, Cho et. al. ‘14
Graves & Jaitly ‘14
V. et. al. ‘15, Li et. al. ‘15
3 of 4 papers to be discussed this class

Image Credit: Andrej Karpathy
Resources

- Chris’ Blog - LSTM unit explanation.
- Karpathy’s Blog - Applications.
- Tensorflow and Caffe - Code examples.
A monkey is pulling a dog’s tail and is chased by the dog.
Recurrent Neural Networks (RNNs) can map a vector to a sequence.

- English Sentence → RNN encoder → RNN decoder → French Sentence
  - [Sutskever et al. NIPS’14]

- Encode → RNN decoder → Sentence
  - [Donahue et al. CVPR’15]
  - [Vinyals et al. CVPR’15]

- Encode → RNN decoder → Sentence
  - [Venugopalan et. al. NAACL’15]

- RNN encoder → RNN decoder → Sentence
  - [Venugopalan et. al. ICCV’15] (this work)
Sequence to Sequence - Video to Text (S2VT)
S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, K. Saenko
1. Train on Imagenet

2. Take activations from layer before classification

Forward propagate
Output: “fc7” features
(activations before classification layer)
1. Train CNN on Activity classes

2. Use optical flow to extract flow images.

3. Take activations from layer before classification

Forward propagate
Output: “fc7” features
(activations before classification layer)

Frames: Flow

CNN (modified AlexNet)

UCF 101
101 Action Classes

fc7: 4096 dimension “feature vector”

[T. Brox et. al. ECCV '04]
Dataset: Youtube

- ~2000 clips
- Avg. length: 11s per clip
- ~40 sentence per clip
- ~81,000 sentences

- A man is walking on a rope.
- A man is walking across a rope.
- A man is balancing on a rope.
- A man is balancing on a rope at the beach.
- A man walks on a tightrope at the beach.
- A man is balancing on a volleyball net.
- A man is walking on a rope held by poles.
- A man balanced on a wire.
- The man is balancing on the wire.
- A man is walking on a rope.
- A man is standing in the sea shore.
**Results (Youtube)**

<table>
<thead>
<tr>
<th>Method</th>
<th>METEOR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean-Pool (VGG)</td>
<td>27.7</td>
</tr>
<tr>
<td>S2VT (randomized)</td>
<td>28.2</td>
</tr>
<tr>
<td>S2VT (RGB)</td>
<td>29.2</td>
</tr>
<tr>
<td>S2VT (RGB+Flow)</td>
<td>29.8</td>
</tr>
</tbody>
</table>

**METEOR**: MT metric. Considers alignment, para-phrases and similarity.
Correct descriptions.

S2VT: A man is doing stunts on his bike.

S2VT: A herd of zebras are walking in a field.

S2VT: A young woman is doing her hair.

S2VT: A man is shooting a gun at a target.

Relevant but incorrect descriptions.

S2VT: A small bus is running into a building.

S2VT: A man is cutting a piece of a pair of paper.

S2VT: A cat is trying to get a small board.

S2VT: A man is spreading butter on a tortilla.

Irrelevant descriptions.

S2VT: A man is pouring liquid in a pan.

S2VT: A polar bear is walking on a hill.

S2VT: A man is doing a pencil.

S2VT: A black clip to walking through a path.
## Evaluation: Movie Corpora

<table>
<thead>
<tr>
<th>MPII-MD</th>
<th>M-VAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>• MPII, Germany</td>
<td>• Univ. of Montreal</td>
</tr>
<tr>
<td>• DVS alignment: semi-automated and crowdsourced</td>
<td>• DVS alignment: automated speech extraction</td>
</tr>
<tr>
<td>• 94 movies</td>
<td>• 92 movies</td>
</tr>
<tr>
<td>• 68,000 clips</td>
<td>• 46,009 clips</td>
</tr>
<tr>
<td>• Avg. length: 3.9s per clip</td>
<td>• Avg. length: 6.2s per clip</td>
</tr>
<tr>
<td>• ~1 sentence per clip</td>
<td>• 1–2 sentences per clip</td>
</tr>
<tr>
<td>• 68,375 sentences</td>
<td>• 56,634 sentences</td>
</tr>
</tbody>
</table>
Looking troubled, someone descends the stairs. The Queen rushes into the courtyard. She then puts a head scarf on. . . . . . and gets into the driver’s side of a nearby Land Rover. The Land Rover pulls away. Three bodyguards quickly jump into a nearby car and follow her.

Processed:
Looking troubled, someone descends the stairs. Someone rushes into the courtyard. She then puts a head scarf on . . .
Results (MPII-MD Movie Corpus)

Best Prior Work
[Rohrbach et al. CVPR'15] 5.6

Mean-Pool 6.7

S2VT (RGB) 7.1
Results (M-VAD Movie Corpus)

- Best Prior Work
  - [Yao et al. ICCV'15]
  - Mean-Pool: 4.3

- Mean-Pool: 6.1

- S2VT (RGB): 6.7
S2VT: Someone sits on his bed, his head on his bed, his eyes open and he takes his hand.
GT: hikin up his pants, his father sits on the bed's edge and leans an arm over someone's legs.

M-VAD: https://youtu.be/pER0mjzSYaM
Discussion

- - -

- What are the advantages/drawbacks of this approach?
  - End-to-end, annotations
- Detaching recognition and generation.
- Why only METEOR (not BLEU or other metrics)?
- Domain adaptation, Re-use RNNs (youtube -> movies, activity recognition)
- Languages other than English.
- Features apart from Optical Flow, RGB; temporal representation.
<table>
<thead>
<tr>
<th>Edit-Distance</th>
<th>$k = 0$</th>
<th>$k \leq 1$</th>
<th>$k \leq 2$</th>
<th>$k \leq 3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSVD</td>
<td>42.9</td>
<td>81.2</td>
<td>93.6</td>
<td>96.6</td>
</tr>
<tr>
<td>MPII-MD</td>
<td>28.8</td>
<td>43.5</td>
<td>56.4</td>
<td>83.0</td>
</tr>
<tr>
<td>MVAD</td>
<td>15.6</td>
<td>28.7</td>
<td>37.8</td>
<td>45.0</td>
</tr>
</tbody>
</table>

Table 3. Percentage of generated sentences which match a sentence of the training set with an edit (Levenshtein) distance of less than 4. All values reported in percentage (%).
Code and more examples
http://vsubhashini.github.io/s2vt.html