Highway Networks for Visual Question Answering

Aaditya Prakash
PhD advisor: James Storer
Brandeis University
Architecture
\[ y = H(x, W) + b \]
Highway Networks

\[ y = H(x, W_H) \cdot T(x, W_T) + x \cdot C(x, W_C) + b \]

\[ T(x) = \sigma(W_Tx + b_T) \]

\[ C(x) = 1 - T(x) \]
Highway Networks

- Allows training very deep networks
  - Srivastava et al trained 50+ layers [1]

- Overcomes vanishing/exploding gradient issues by learning gating mechanism, like LSTM

- Includes ‘Transform’ gate (T) and ‘Carry’ gate (C)
  - Simple Perceptron
    \[ y = H(x, W) + b \]
  - Highway Layer (MLP)
    \[ y = H(x, W_H) \cdot T(x, W_T) + x \cdot C(x, W_C) + b \]
    \[ T(x) = \sigma(W_T x + b_T) \]
    \[ C(x) = 1 - T(x) \]
Multimodal Learning

VQA

$x_1$ → Image

$x_2$ → Question
Multimodal Learning
VQA
Multimodal Learning
VQA
VISUAL QA

Highway Networks
After many layers later
And countless debugging
Note:
Figure does not mention the use of following techniques:

- Dropout and Batch-Normalization
- Image feature normalization
- Image augmentation before feature extraction
- Use of other word vectors like Word2Vec and ConceptNet
Results & Performance
### Results from VQA Challenge

#### Real Open-Ended Test Standard 2015* (%)

<table>
<thead>
<tr>
<th>Yes/No</th>
<th>Number</th>
<th>Other</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>82.11</td>
<td>37.73</td>
<td>51.91</td>
</tr>
<tr>
<td></td>
<td>62.88</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Real Multiple choice Test Standard 2015 (%)

<table>
<thead>
<tr>
<th>Yes/No</th>
<th>Number</th>
<th>Other</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>81.95</td>
<td>38.56</td>
<td>56.4</td>
</tr>
<tr>
<td></td>
<td>65.07</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Five model ensemble
  - Model 1 - VGGNet + 98% SF + Glove
  - Model 2 - VGGNet + 95% SF + Word2Vec
  - Model 3 - ResNet + 98% SF + Glove
  - Model 4 - ResNet + 98% SF + ConceptNet Numberbatch
  - Model 5 - ResNet + 95% SF + Word2Vec
- 10 Crop image inference ensembled into one answer
- SF - Statistical Filtering: restrict the answer to some percentage of answers within that question type
- Trained on train2014 + val2014 + finetuned on results from earlier model from test2015 [3]
- No SF for Real Multiple Choice (this might have been a bad idea)
# Comparison of Accuracy over depth

## VGGNet (4096 features)*

<table>
<thead>
<tr>
<th># Layers</th>
<th>Parameters (millions)</th>
<th>Accuracy (val)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>46.052</td>
<td>22.83</td>
</tr>
<tr>
<td>3</td>
<td>113.177</td>
<td>44.7</td>
</tr>
<tr>
<td>5</td>
<td>180.302</td>
<td>47.4</td>
</tr>
<tr>
<td>10</td>
<td>348.115</td>
<td>55.7</td>
</tr>
</tbody>
</table>

## ResNet (2048 features)*

<table>
<thead>
<tr>
<th># Layers</th>
<th>Parameters (millions)</th>
<th>Accuracy (val) %</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>14.638</td>
<td>22.1</td>
</tr>
<tr>
<td>3</td>
<td>31.423</td>
<td>45.85</td>
</tr>
<tr>
<td>5</td>
<td>48.208</td>
<td>49.21</td>
</tr>
<tr>
<td>10</td>
<td>90.172</td>
<td>57.1</td>
</tr>
</tbody>
</table>

* Trained on train2014 and tested on val2014
* Single model (no ensembling), No Statistical filtering
Comparison of accuracy & parameters over depth

* Trained on train2014 and tested on val2014
* Single model (no ensembling), No Statistical filtering
* Real Open-Ended only
Hyper Parameter Search

*Trained on train2014 and tested on val2014, ResNet
*Single model (no ensembling), No Statistical filtering (SF)
* Real OpenEnded only

Parameters

- Learning Rate
- Number of output (softmax)
- Initialization
  - Uniform
  - Xavier
  - Kaiming
  - heuristic
- Activation (tanh/relu/prelu)
- Num highway layers (1,2,3,4,6,10)
- Bias (Carry & Transfer)
- Decay factor
- Epoch at which to change optimizer
References


My thanks to -

- VQA Team for the challenge
- Aishwarya Agrawal for blazing fast replies to all my queries
- James Storer, my PhD advisor.
- NVIDIA for gifting us a Titan X.
- Following people from whose code I learned -
  - Yoon Kim @yoonkim (HarvardNLP)
  - Jin-Hwa Kim @jnhwkim (Element-Research)
  - Jainsen Lu @jiasenlu (VQA_LSTM_CNN)
  - François Chollet @fchollet (Keras)
  - Hyeonwoo Noh @HyeonwooNoh (DPPNet)
  - Bolei Zhou @metalbubble (VQAbaseline)
  - Matthew Honnibal @honnibal (Spacy)