Visualizing, Measuring and Understanding Neural Networks

Zhiyuan Tang
tangzy@cslt.riit.tsinghua.edu.cn

August 1, 2016
Analyses of Deep Neural Networks

- visualizing
- quantitative vs. qualitative
- Task-specific knowledge
- interpretation of hidden feature
- pattern / event
- high-dimensional data
- discriminative vs. generative
- ...

...
Computer Vision (1)

- Interpreting high level features
- Stacked DAE (feedforward), MNIST
- activation maximization
- Optimization: non-convex, gradient descent
- Most random initializations yield same pattern
- sensitivity analysis: specific unit, general unit
- hierarchical representation

Computer Vision (2)

- contribution of different layers (CNN)
- map activities back to input pixel space
- Deconvolutional Network (prior)
- greater invariance at higher layers
Computer Vision (2)

- “dead” features
- translation, scale, and rotation, increasing invariance and class discrimination in layers
- Correspondence Analysis
- Sensitivity analysis (Occlusion Sensitivity)
- Compositionality

Computer Vision (3)

- computing the gradient (CNN)
- generating an image, which maximises the class score, google dream
- initialized with zero image, added training set mean
- saliency map: gradient based
- weakly supervised object segmentation
- gradient-based visualization and DeConv

NLP (1)

- Text prediction, character-level LM (LSTM)
- **time-scales**, the effect of the typo remains present in each layer
- Short time-scales in bottom layers, longer memory in top ones
- Long-term interactions: parenthesis
- Temporal hierarchy

- character-level LM, Linux Kernel Dataset (RNN, LSTM, GRU)
- Parentheses, the start of URLs, quotes, depth of expression (other syntactic patterns)
- Many cells not interpretable
- Saturation
- vs. n-gram

- Phrase classification (LSTM)
- negation, intensification
NLP (3)

- Phrase classification (LSTM)
- Important words: gradient-based
- Concessive clauses
- Phrase classification (LSTM)
- negative asymmetry

**CV & NLP**

- Text prediction & image-vector generation (2 GRUs)
- pays selective attention to lexical categories grammatical and semantic information
- Hidden units of VISUAL active for meaningful constructions


Attention/Encoder-decoder

Variations

- Depth,
- LSTM vs. GRU
- Architecture
- LSTM components

Platforms


Measurement

- General RNN: unidirectional, bidirectional, homogeneous
- **recurrent depth**: complexity in the long term
- **feedforward depth**: complexity in the short term
- **recurrent skip coefficients**: the speed of information flow
- “skipping” allowing unimpeded information flow

---

<table>
<thead>
<tr>
<th>Color</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Red</td>
<td>the longest path, with infinite time gap&lt;br&gt;growth rate (a limit over time) is <strong>recurrent depth</strong> $d_r$</td>
</tr>
<tr>
<td>Yellow</td>
<td>the longest input-output path, with least time gap&lt;br&gt;with $d_r$ to <strong>compute feedforward depth</strong></td>
</tr>
<tr>
<td>Blue</td>
<td>the shortest path, with infinite time gap&lt;br&gt;reciprocal of growth rate (a limit over time) is <strong>recurrent skip coefficients</strong></td>
</tr>
</tbody>
</table>
Measurement

- **feedforward depth**: might not help on long term dependency tasks
- **recurrent depth**: might yield performance improvements
- **recurrent skip coefficients**: largely improve performance on long term dependency tasks.

Rediscovery


How to do with ASR

DISCUSSION

- Acoustic knowledge: (vowel, consonant), (Initial, Final), ...
- Linguistic knowledge: syntactic info, ...
- Model attributes: long-term range, short term range, ...
- Discriminative model for generating
- Waveform/feature transform
- Discovery new knowledge
Thanks a lot.