RL for DL

Yoshua Bengio

December 11, 2015

NIPS’2015 Deep Reinforcement Learning Workshop
Deep Learning: Beyond Pattern Recognition, towards AI

- Many researchers believed that neural nets could at best be good at pattern recognition
- And they are really good at it!

- But many more ingredients needed towards AI. Recent progress:

  - ATTENTION & REASONING:
    - Machine translation, Memory networks & Neural Turing Machine
  - PLANNING & REINFORCEMENT LEARNING:
    - DeepMind (Atari game playing) & Berkeley (Robotic control)
How to train neural nets to take internal discrete decisions?

• Can we approximate the gradient through discrete (possibly stochastic decisions) so as to extend the reach of back-prop?

• Usage:
  • Conditional computation / dynamically routed architectures
  • Attention (internal and external)
  • Alignment, (hierarchical) segmentation, etc.
  • Long-term dependencies through many nonlinearities: almost not differentiable

• Simple solution: REINFORCE (+ conditional baseline)

\[
\frac{\partial}{\partial \theta} \sum_a p_\theta(a|x) L(a) = \sum_a p_\theta(a|x) L(a) \frac{\partial \log p_\theta(a|x)}{\partial \theta}
\]
Conditional Computation
Computation / Capacity Ratio

- N-grams, decision trees, etc.: capacity (and memory) can grow a lot while computation remains constant or grows as \( \log(\text{capacity}) \).

- Neural nets / deep learning: computation grows linearly with capacity (number of parameters). Each parameter is used for every example.

- To build much higher capacity models, we need to break that linear relationship.
Conditional Computation: only visit a small fraction of parameters / example

Bengio, Leonard & Courville
arXiv 1305.2982

• Deep nets vs decision trees
• Hard mixtures of experts (Collobert, Bengio & Bengio 2002)
• Conditional computation for deep nets: sparse distributed gaters selecting combinatorial subsets of a deep net
• Challenges:
  • Credit assignment for hard decisions
  • Gated architectures exploration
Credit Assignment for Discrete Actions

(Bengio, Leonard, Courville 2013): *Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation*

- Gating units take a hard decision
- Gradient through discrete function = 0
- Solution ideas in (Bengio et al 2013):
  - Heuristic back-prop (straight through estimator), also (Gregor et al ICML 2014).
  - Noisy rectifier:
  - Smooth times Stochastic bvp with b ~ Bin(vp)
  - REINFORCE with variance reduction i.e., RL, i.e. correlate with loss, no back-prop for gaters

Attention for MT, caption generation and Reasoning
Encoder-Decoder Framework

- Intermediate representation of meaning
  = ‘universal representation’
- Encoder: from word sequence to sentence representation
- Decoder: from representation to word sequence distribution
Attention Mechanism for Deep Learning

- Consider an input (or intermediate) sequence or image
- Consider an upper level representation, which can choose « where to look », by assigning a weight or probability to each input position, as produced by an MLP, applied at each position

![Diagram showing attention mechanism between lower and higher levels.]
Soft-Attention vs Stochastic Hard-Attention

- With soft-attention: input fed to higher level at location $i$ is a softmax-weighted sum of states at locations $j$ at lower level
  - Train by back-prop
  - Fast training

- With stochastic hard-attention: sample an input location according to the softmax output
  - Get a gradient on the decisions via REINFORCE - baseline
  - Noisy gradient, slower training but *works surprisingly well*
  - *Symmetry breaking*
End-to-End Machine Translation with Recurrent Nets and Attention Mechanism

- Reached the state-of-the-art in one year, from scratch

### (a) English → French (WMT-14)

<table>
<thead>
<tr>
<th></th>
<th>NMT(A)</th>
<th>Google</th>
<th>P-SMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMT</td>
<td>32.68</td>
<td>30.6*</td>
<td></td>
</tr>
<tr>
<td>+Cand</td>
<td>33.28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+UNK</td>
<td>33.99</td>
<td>32.7°</td>
<td>37.03*</td>
</tr>
<tr>
<td>+Ens</td>
<td>36.71</td>
<td>36.9°</td>
<td></td>
</tr>
</tbody>
</table>

### (b) English → German (WMT-15)

<table>
<thead>
<tr>
<th>Model</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>24.8</td>
<td>Neural MT</td>
</tr>
<tr>
<td>24.0</td>
<td>U.Edinburgh, Syntactic SMT</td>
</tr>
<tr>
<td>23.6</td>
<td>LIMSI/KIT</td>
</tr>
<tr>
<td>22.8</td>
<td>U.Edinburgh, Phrase SMT</td>
</tr>
<tr>
<td>22.7</td>
<td>KIT, Phrase SMT</td>
</tr>
</tbody>
</table>

### (c) English → Czech (WMT-15)

<table>
<thead>
<tr>
<th>Model</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>18.3</td>
<td>Neural MT</td>
</tr>
<tr>
<td>18.2</td>
<td>JHU, SMT+LM+OSM+Sparse</td>
</tr>
<tr>
<td>17.6</td>
<td>CU, Phrase SMT</td>
</tr>
<tr>
<td>17.4</td>
<td>U.Edinburgh, Phrase SMT</td>
</tr>
<tr>
<td>16.1</td>
<td>U.Edinburgh, Syntactic SMT</td>
</tr>
</tbody>
</table>
TED talk MT, English-German

**BLEU (CASED)**

<table>
<thead>
<tr>
<th>Country</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stanford</td>
<td>30.85</td>
</tr>
<tr>
<td>Karlsruhe</td>
<td>26.18</td>
</tr>
<tr>
<td>Edinburgh</td>
<td>26.02</td>
</tr>
<tr>
<td>Heidelberg</td>
<td>24.96</td>
</tr>
<tr>
<td>PJAIT</td>
<td>22.51</td>
</tr>
<tr>
<td>Baseline</td>
<td>20.08</td>
</tr>
</tbody>
</table>

**HTER (HE SET)**

<table>
<thead>
<tr>
<th>Country</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stanford</td>
<td>16.16</td>
</tr>
<tr>
<td>Edinburgh</td>
<td>21.84</td>
</tr>
<tr>
<td>Karlsruhe</td>
<td>22.67</td>
</tr>
<tr>
<td>Heidelberg</td>
<td>23.42</td>
</tr>
<tr>
<td>PJAIT</td>
<td>28.18</td>
</tr>
</tbody>
</table>

-26%
Image-to-Text: Caption Generation with Attention

\[ f = (\text{a, man, is, jumping, into, a, lake, .).} \]

**Word Sample**

**Recurrent State**

**Attention Mechanism**

\[ \sum a_j = 1 \]

\( \text{Xu et al., 2015, Yao et al., 2015} \)
Paying Attention to Selected Parts of the Image While Uttering Words
Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

Results from (Xu et al, ICML 2015)

Table 1. BLEU-1,2,3,4/METEOR metrics compared to other methods, † indicates a different split, (—) indicates an unknown metric, ○ indicates the authors kindly provided missing metrics by personal communication, Σ indicates an ensemble, a indicates using AlexNet

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th>BLEU</th>
<th>METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>B-1</td>
<td>B-2</td>
</tr>
<tr>
<td>Flickr8k</td>
<td>Google NIC (Vinyals et al., 2014)†Σ</td>
<td>63</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td>Log Bilinear (Kiros et al., 2014a)○</td>
<td>65.6</td>
<td>42.4</td>
</tr>
<tr>
<td></td>
<td>Soft-Attention</td>
<td>67</td>
<td>44.8</td>
</tr>
<tr>
<td></td>
<td>Hard-Attention</td>
<td>67</td>
<td>45.7</td>
</tr>
<tr>
<td>Flickr30k</td>
<td>Google NIC† Σ</td>
<td>66.3</td>
<td>42.3</td>
</tr>
<tr>
<td></td>
<td>Log Bilinear</td>
<td>60.0</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>Soft-Attention</td>
<td>66.7</td>
<td>43.4</td>
</tr>
<tr>
<td></td>
<td>Hard-Attention</td>
<td><strong>66.9</strong></td>
<td><strong>43.9</strong></td>
</tr>
<tr>
<td>COCO</td>
<td>CMU/MS Research (Chen &amp; Zitnick, 2014)α</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>MS Research (Fang et al., 2014)†α</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>BRNN (Karpathy &amp; Li, 2014)○</td>
<td>64.2</td>
<td>45.1</td>
</tr>
<tr>
<td></td>
<td>Google NIC† Σ</td>
<td>66.6</td>
<td>46.1</td>
</tr>
<tr>
<td></td>
<td>Log Bilinear○</td>
<td>70.8</td>
<td>48.9</td>
</tr>
<tr>
<td></td>
<td>Soft-Attention</td>
<td>70.7</td>
<td>49.2</td>
</tr>
<tr>
<td></td>
<td>Hard-Attention</td>
<td><strong>71.8</strong></td>
<td><strong>50.4</strong></td>
</tr>
</tbody>
</table>
The Good

A woman is throwing a frisbee in a park.
A dog is standing on a hardwood floor.
A stop sign is on a road with a mountain in the background.
A little girl sitting on a bed with a teddy bear.
A group of people sitting on a boat in the water.
A giraffe standing in a forest with trees in the background.
And the Bad

A large white **bird** standing in a forest.

A woman holding a **clock** in her hand.

A man wearing a hat and a hat on a **skateboard**.

A person is standing on a beach with a **surfboard**.

A woman is sitting at a table with a large **pizza**.

A man is talking on his cell phone while another man watches.
Attention Mechanisms for Memory Access

- Neural Turing Machines (Graves et al 2014)
- and Memory Networks (Weston et al 2014)
- Use a form of attention mechanism to control the read and write access into a memory
- The attention mechanism outputs a softmax over memory locations
- For efficiency, the softmax should be sparse (mostly 0’s), e.g. maybe using a hash-table formulation.
- Both soft and (stochastic) hard attention are used
Training a Critic
Alternatives to REINFORCE?

• Train a critic (estimate future reward) and backprop through it
  - “Generative Adversarial Networks”, Goodfellow et al, NIPS 2014

• NN estimates loss that would be obtained for any discrete choice of sequence of actions
• At test-time, use beam-search (planning) to find approximately optimal seq. of actions

• Train a stochastic credit-assignment machine, e.g., by
  Reweighted Wake-Sleep (Bornshein & Bengio 2014), or
  Variational Auto-Encoder (Kingma et al 2014) which amounts to
  learning to predict the credit to attribute, given the future
  reward (observed outputs), in the form of a posterior probability
  distribution for the discrete latent variable
(Conditional) Reweighted Wake-Sleep

(Bornschein & Bengio ICLR 2015)

- If $H$ is continuous, can use a conditional VAE framework
- Otherwise a conditional reweighted wake-sleep
- See also Ba et al NIPS’2015 for an application of similar idea to a recurrent attention model

$$P(H|X) = \text{stochastic policy}$$

$$Q(H|X,Y) = \text{how } H \text{ should be in order to maximize reward}$$

$$\log P(Y|X,H) = \text{reward}$$

visible input $X$

visible target $Y$

discrete stochastic variable $H$
Training a Critic for Generative Models
Maximizing the Probability of Passing a Turing Test for Generative Models

• Generate sample $y$ (possibly given $x$) OR get $(y,x)$ from data generating distribution

• Ask human if answer $y | x$ comes from data generating distribution (good) or from the computer (bad)

• If human cannot statistically distinguish the two distributions, then the computer passes that Turing test

• Can we train a critic that predicts the human answer?
GAN: Generative Adversarial Networks

Goodfellow et al NIPS 2014
LAPGAN: Visual Turing Test

(Denton + Chintala, et al 2015)

• 40% of samples mistaken by humans for real photos

• Sharper images than max. lik. proxys (which min. KL(data|model)):
• GAN objective = compromise between KL(data|model) and KL(model|data)
Convolutional GANs

(Radford et al, arXiv 1511.06343)

Strided convolutions, batch normalization, only convolutional layers, ReLU and leaky ReLU
Biologically Plausible and Memory-Efficient Online Training of RNNs?

• The brain is a big RNN
• Backprop in a feedforward net may have plausible biological implementations (using the feedback weights to propagate credit information, targets or gradients)
• But what about backprop through time?
  - Requires storing the state of the network (i.e. the activations of all neurons) for an indefinitely long duration
  - Need to wait for the “end of the episode” (your life) to start learning
  - Some form of online learning is necessary
• If we had a strong critic trained to predict discounted future rewards (and we have it in our brain), then we would just need to backprop through it to update our policy
• Critic is missing information (intermediate actions and observations)
MILA: Montreal Institute for Learning Algorithms