10707
Deep Learning
Russ Salakhutdinov

Machine Learning Department
rsalakhu@cs.cmu.edu
http://www.cs.cmu.edu/~rsalakhu/10707/

Sequence to Sequence II
Slides borrowed from ICML Tutorial

Seq2Seq ICML Tutorial

Oriol Vinyals and Navdeep Jaitly
@OriolVinyalsML | @NavdeepLearning
Site: https://sites.google.com/view/seq2seq-icml17
Sydney, Australia, 2017
Applications
Sentence to Constituency Parse Tree

1. Read a sentence
2. Flatten the tree into a sequence (adding (,))
3. "Translate" from sentence to parse tree

John has a dog. →

(NP NNP VBZ NP DT NN)VP . )S

Speech Recognition

$p(y_{i+1}|y_{1..i}, x)$

Attention Example

Prediction derived from “attending” to segment of input

Attention vector - where the model thinks the relevant information is to be found

Attention Example

Attention Example

Attention Example

Attention Example

Attention Example

Attention Example

A man riding a horse in a field.

Xu et al, ICML 2015
Caption Generation with Visual Attention

A woman holding a clock in her hand.

A large white bird standing in a forest.

Xu et al, ICML 2015
Listen Attend and Spell (LAS)

- Reducing time resolution with a pyramidal encoder

## LAS Results

<table>
<thead>
<tr>
<th>Beam</th>
<th>Text</th>
<th>LogProb</th>
<th>WE R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Truth</td>
<td>call aaa roadside assistance</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>call aaa roadside assistance</td>
<td>-0.5740</td>
<td>0.00</td>
</tr>
<tr>
<td>2</td>
<td>call triple a roadside assistance</td>
<td>-1.5399</td>
<td>50.0</td>
</tr>
<tr>
<td>3</td>
<td>call trip way roadside assistance</td>
<td>-3.5012</td>
<td>50.0</td>
</tr>
<tr>
<td>4</td>
<td>call xxx roadside assistance</td>
<td>-4.4375</td>
<td>25.0</td>
</tr>
</tbody>
</table>
Lip Reading

<table>
<thead>
<tr>
<th>Channel</th>
<th>Series name</th>
<th># hours</th>
<th># sent.</th>
</tr>
</thead>
<tbody>
<tr>
<td>BBC 1 HD</td>
<td>News†</td>
<td>1,584</td>
<td>50,493</td>
</tr>
<tr>
<td>BBC 1 HD</td>
<td>Breakfast</td>
<td>1,997</td>
<td>29,862</td>
</tr>
<tr>
<td>BBC 1 HD</td>
<td>Newsnight</td>
<td>590</td>
<td>17,004</td>
</tr>
<tr>
<td>BBC 2 HD</td>
<td>World News</td>
<td>194</td>
<td>3,504</td>
</tr>
<tr>
<td>BBC 2 HD</td>
<td>Question Time</td>
<td>323</td>
<td>11,695</td>
</tr>
<tr>
<td>BBC 4 HD</td>
<td>World Today</td>
<td>272</td>
<td>5,558</td>
</tr>
<tr>
<td>All</td>
<td></td>
<td>4,960</td>
<td>118,116</td>
</tr>
</tbody>
</table>

http://www.robots.ox.ac.uk/~vgg/data/lip_reading/

Lip Reading

Separate embedding and attention for audio and visual streams

Google Neural Machine Translation System

Google Neural Machine Translation System

Closes gap between old system and human-quality translation by 58% to 87%
Loss Functions
Loss Functions

- Cross Entropy
- Scheduled Sampling [1]
- Expected Loss [2]
- Augmented Loss [3]
- Sequence to Sequence as a beam search optimization [4]
- Learning decoders with different loss function [5]

Cross Entropy (Negative Log Likelihood) Loss

- Log Likelihood, by chain rule is sum of next step log likelihoods

\[ \log p(y|x) = \sum_{i=1}^{N} \log p(y_i|y_{<i}, x) \]

- Supervised classification for each time step
  - depends on input, past outputs, which are known during training
Training and Inference Mismatch

Training

\[ P(y_t|h_t) \text{ with } h_t = f(h_{t-1}, y_{t-1}; \theta) \]

Training and Inference Mismatch

Inference

\[ P(y_t|h_t) \text{ with } h_t = f(h_{t-1}, y_{t-1}; \theta) \]

Scheduled Sampling

\[ P(y_t | h_t) \text{ with } h_t = f(h_{t-1}, \hat{y}_{t-1}; \theta) \]

## Scheduled Sampling

<table>
<thead>
<tr>
<th>Machine Translation Model</th>
<th>Bleu-4</th>
<th>Meteor</th>
<th>Cider</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>28.8</td>
<td>24.2</td>
<td>89.5</td>
</tr>
<tr>
<td>Baseline with dropout</td>
<td>28.1</td>
<td>23.9</td>
<td>87.0</td>
</tr>
<tr>
<td>Scheduled sampling</td>
<td>30.6</td>
<td>24.3</td>
<td>92.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parsing Model</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline LSTM with dropout</td>
<td>87.00</td>
</tr>
<tr>
<td>Scheduled sampling with dropout</td>
<td>88.68</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Speech Recognition Model</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>LAS + LM Rescoring</td>
<td>12.6</td>
</tr>
<tr>
<td>LAS + <strong>Sampling</strong> + LM Rescoring</td>
<td>10.3</td>
</tr>
</tbody>
</table>

Rewards (-loss) used in Structured Prediction

<table>
<thead>
<tr>
<th>TASK</th>
<th>REWARD</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification</td>
<td>0/1 rewards</td>
<td>$r(y, y^<em>) = 1[y = y^</em>]$</td>
</tr>
<tr>
<td>Segmentation</td>
<td>Intersection over Union</td>
<td>$r(y, y^<em>) = \frac{\cap(y, y^</em>)}{\cup(y, y^*)}$</td>
</tr>
<tr>
<td>Speech Recognition</td>
<td>Edit Distance</td>
<td>$r(y, y^*) = (#d + #i + #s)$</td>
</tr>
<tr>
<td>Machine Translation</td>
<td>BLEU</td>
<td></td>
</tr>
</tbody>
</table>
Expected reward (-loss)

Given a dataset of input output pairs, \( \mathcal{D} \equiv \{ (x^{(i)}, y^{(i)*}) \}_{i=1}^{N} \)

learn a conditional distribution \( p_\theta(y \mid x) \) that minimizes expected loss:

\[
\mathcal{L}_{\text{RL}}(\theta) = \sum_{(x, y^{*}) \in \mathcal{D}} - \sum_{y \in \mathcal{Y}} p_\theta(y \mid x) \ r(y, y^{*})
\]

Sample from the model distribution

Difficult / Impossible to train from scratch!!
Mixed Incremental Cross-Entropy Reinforce (MIXER)

● Gradually interpolate from Cross-Entropy to Expected Loss

Data: a set of sequences with their corresponding context.
Result: RNN optimized for generation.
Initialize RNN at random and set $N_{XENT}$, $N_{XE+R}$ and $\Delta$;
for $s = T, 1, -\Delta$ do
  if $s == T$ then
    train RNN for $N_{XENT}$ epochs using XENT only;
  else
    train RNN for $N_{XE+R}$ epochs. Use XENT loss in the first $s$ steps, and REINFORCE (sampling from the model) in the remaining $T - s$ steps;
end

More expected loss optimization as training proceeds

Mixed Incremental Cross-Entropy Reinforce (MIXER)

<table>
<thead>
<tr>
<th>TASK</th>
<th>XENT</th>
<th>DAD</th>
<th>E2E</th>
<th>MIXER</th>
</tr>
</thead>
<tbody>
<tr>
<td>summarization</td>
<td>13.01</td>
<td>12.18</td>
<td>12.78</td>
<td><strong>16.22</strong></td>
</tr>
<tr>
<td>translation</td>
<td>17.74</td>
<td>20.12</td>
<td>17.77</td>
<td><strong>20.73</strong></td>
</tr>
<tr>
<td>image captioning</td>
<td>27.8</td>
<td>28.16</td>
<td>26.42</td>
<td><strong>29.16</strong></td>
</tr>
</tbody>
</table>

Reward Augmented Maximum Likelihood (RML)

Finding the *right output sequence*, for tasks like speech recognition or machine translation is like finding a *needle in a haystack*. It is very risky to shoot *only* for the *true target*. What if we expand the targets to make learning easier? *E.g.* by inserting, deleting random words...

Reward Augmented maximum likelihood (RML)

\[
\mathcal{L}_{\text{RML}}(\theta; \tau) = \sum_{(x, y^*) \in \mathcal{D}} \left\{ - \sum_{y \in \mathcal{Y}} q(y \mid y^*; \tau) \log p_\theta(y \mid x) \right\}
\]

Optimal \( p_\theta(y \mid x) \):

\[
q(y \mid y^*; \tau) = \frac{1}{Z(y^*, \tau)} \exp \left\{ \tau(y, y^*) / \tau \right\}
\]

\[
\mathcal{L}_{\text{RML}}(\theta; \tau) = \sum_{(x, y^*) \in \mathcal{D}} D_{\text{KL}}(q(y \mid y^*; \tau) \parallel p_\theta(y \mid x)) + \text{constant}
\]

RML - Impact of temperature $\tau$

Temperature impacts spread of distribution we sample from

Cross Entropy Targets $\quad$ More spread

$q(y \mid y^*; \tau = 0)$ $\quad$ $q(y \mid y^*; \tau = .1)$ $\quad$ $q(y \mid y^*; \tau = .2)$

\[
\mathcal{L}_\text{RML}(\theta; \tau) = \sum_{(x,y^*) \in \mathcal{D}} D_{KL}(q(y \mid y^*; \tau) \parallel p_\theta(y \mid x)) + \text{constant}
\]

Margin Loss

- Perform beam search until correct hypothesis falls out of the beam
- Restart beam whenever there is a violation
- Extract correct hypothesis and competing hypotheses

Margin Loss

- Add a margin score for all time steps where the correct hypothesis is not better than the Kth best hypothesis by a certain margin.

\[
\mathcal{L}(f) = \sum_{t=1}^{T} \Delta(\hat{y}_{1:t}^{(K)}) \left[ 1 - f(y_t, h_{t-1}) + f(\hat{y}_t^{(K)}, \hat{h}_t^{(K)}) \right]
\]

- Loss for error; 0 when margin constraint is satisfied.
- Score function for prediction of current output.
- Score function for prediction of Kth best output.

Wiseman, S., Rush, A. “Sequence-to-sequence learning as beam-search optimization.” *EMLP (2016).*
## Margin Loss

<table>
<thead>
<tr>
<th></th>
<th>Machine Translation (BLEU)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$K_{te} = 1$</td>
</tr>
<tr>
<td>seq2seq</td>
<td>22.53</td>
</tr>
<tr>
<td>BSO, SB-Δ</td>
<td>23.83</td>
</tr>
<tr>
<td>XENT</td>
<td>17.74</td>
</tr>
<tr>
<td>DAD</td>
<td>20.12</td>
</tr>
<tr>
<td>MIXER</td>
<td>20.73</td>
</tr>
</tbody>
</table>

Wiseman, S., Rush, A. “Sequence-to-sequence learning as beam-search optimization.” *EMLP (2016).*
Autoregressive
Generative Models
Pixel RNN Model

$$p(x) = \prod_{i=1}^{n^2} p(x_i | x_1, \ldots, x_{i-1})$$

- Fully visible
- Similar to language models with RNNs
- Model pixels with Softmax

Softmax Sampling

Softmax Sampling

Softmax Sampling

Softmax Sampling

Softmax Sampling

Softmax Sampling

Softmax Sampling

Softmax Sampling

Softmax Sampling

Softmax Sampling

Softmax Sampling

Softmax Sampling

Softmax Sampling

Pixel RNN

Sequence of Words == Sequence of Pixels

Pixel RNN

Pixel RNN

## Video Pixel Network (VPN)

<table>
<thead>
<tr>
<th>Model</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Shi et al., 2015)</td>
<td>367.2</td>
</tr>
<tr>
<td>(Srivastava et al., 2015a)</td>
<td>341.2</td>
</tr>
<tr>
<td>(Brabandere et al., 2016)</td>
<td>285.2</td>
</tr>
<tr>
<td>(Patraucean et al., 2015)</td>
<td>179.8</td>
</tr>
<tr>
<td>Baseline model</td>
<td>110.1</td>
</tr>
<tr>
<td><strong>VPN</strong></td>
<td><strong>87.6</strong></td>
</tr>
<tr>
<td><strong>Lower Bound</strong></td>
<td><strong>86.3</strong></td>
</tr>
</tbody>
</table>

New Architectures


Conv seq2seq, Gehring, et al, 2017

Att is all you need, Vaswani, et al, 2017
Self-Attention

Convolution

Self-Attention
Self-Attention

Convolution

Self-Attention
MultiModel

“Last week, Kigali raised the possibility of military retaliation after shells...”

“Can you give our readers some details on this?”

The above represents a triumph of either apathy or civility

“S NP DT JJS /NP VP VBZ NP NP DT NN /NP PP IN NP NP NN /NP CC NP NN /NP /NP /PP /NP /VP . /S”