CMPUT 367

Roderick Lan

Contents

Lecture 17 - Mar 19

Prior knowledge abt gradient
(finish - first 9 min)

0.1 Prior Knowledge

CNN - spatial neighborhood; sliding window
RNN - ordered info; sequential proc
LLM has momentum, variance, gradient info

Don't want to compress weight, want to compress gradient information without losing information

> pytorch and tf give gradient of all weights (whole memory space)
>
> can also do it layer by layer

grad = loss.back()
parm = opt.adam(param, grad)

1 Structured Network

Has parse tree structure info.

1.1 Recursive Propagation

RNN that incorporates parse structure

2 Tree Based Convolution

Tree structure instead of sequence

3 Seq2Seq and Attention

3.0.1 Sidenote:

In compilers, CFG is NP-complete.
Avoid backtracking, look forward languages preferred

3.1 Seq2Seq

Given sequence of input signals (usually time series info. like text)
Encoder / Decoder structure
(finish; 41 min in lec)
Important: Need to feed data back

3.1.1 Training

Decoder input layer:

Attempt 1 - feed in predicted

Attempt 2 - feed in ground truth

Cannot feed ground ground truth during inference

3.1.2 Inference

Decoder input layer - feed in predicted words
Terminate w/ EOS token (EOS predicted $\rightarrow$ sentence terminated)

3.1.3 Caveat

Batch implementation

Padding EOS or 0 vector $\rightarrow$ incorrect

Masking $\rightarrow$ correct; 1 = actual step, 0 = virtual step

3.1.4 Inference Criteria

Single output classification

MAP inference $\iff$ minimal empirical loss

(get most likely category given inference)

$$y = \arg\max p(y|x)$$

Sentence Generation

generate best sentence

$$\mathbf{y} = \arg\max_{\mathbf{y}} p(\mathbf{y}|\mathbf{x})$$

arg max over entire sentence space

greedy can sometimes work

exhaustive search too long ($O(b^l)$)

Use something in between greedy and exhaustive $\rightarrow$ beam search

try all words or a few best

dont maintain all; maintian only several good partial seq.

3.2 Beam Search

3.3 Isuses with Autoreg

Error accumulation

fix by feed in self gen words and anneal ground truth

Label bias

Locally normalized models prefer highly probable (but possibly unimportant) words in a more conentrated pred.