Practical Tricks and Story Generation

CS 585, Fall 2019

Introduction to Natural Language Processing
http://people.cs.umass.edu/~miyyer/cs585/

Shufan Wang

College of Information and Computer Sciences

University of Massachusetts Amherst

Practical Tricks while working with NN

- Speed Tricks
- Memory Tricks
- Debugging Tricks

Speed tricks

Can you identify GPU vs CPU

CPU vs GPU

- CPU might be faster in many NLP applications/analysis tasks

- GPU will be better if:
  - Large hidden state size
  - Data processed by batches
  - Large matrix

CPU vs GPU

- CPU might be faster in many NLP applications/analysis tasks

- GPU will be better if:
  - Large hidden state size
  - Data processed by batches
  - Large matrix (attention, feedforward, softmax)

Reduce data transmission/augmentation

- Process tokens in bundles and move them to GPU once
- Group sequence of tokens by length in a batch
- Use pytorch/tensorflow dataloaders

Memory tricks

Memory is often an issue

- 1Byte = 8 bit
- 1KB = 1000 Byte
- 1MB = 1000 KB
- 1GB = 1000 MB
- 1TB = 1000 GB
- 10 KB = 10000 Byte

Memory components

- model parameter
- input/output
- batch_size * each sample

Memory components

- model parameter
- input/output
- batch_size * each sample

Missing something?

Memory components

- model parameter
- input/output
- batch_size * each sample
- Gradients !!!

GPT-2 medium

- 345 parameters, 1 float = 4 bytes
- 345 * 4, about 1.5 G
- Plus gradients, input/output, data
- Almost 5 G
- Most GPUs have 12 GB memory

- Easily get OOM even when batch size is small (2 or 3)

Handle the OOM issue

- Data parallel:

```
# declare model
model = myGreatModel()

# parallelize
model_para = nn.DataParallel(model, device_ids=DEVICES_IDS_LIST)

# move to device
model_para.to(device)
```

Handle the OOM issue

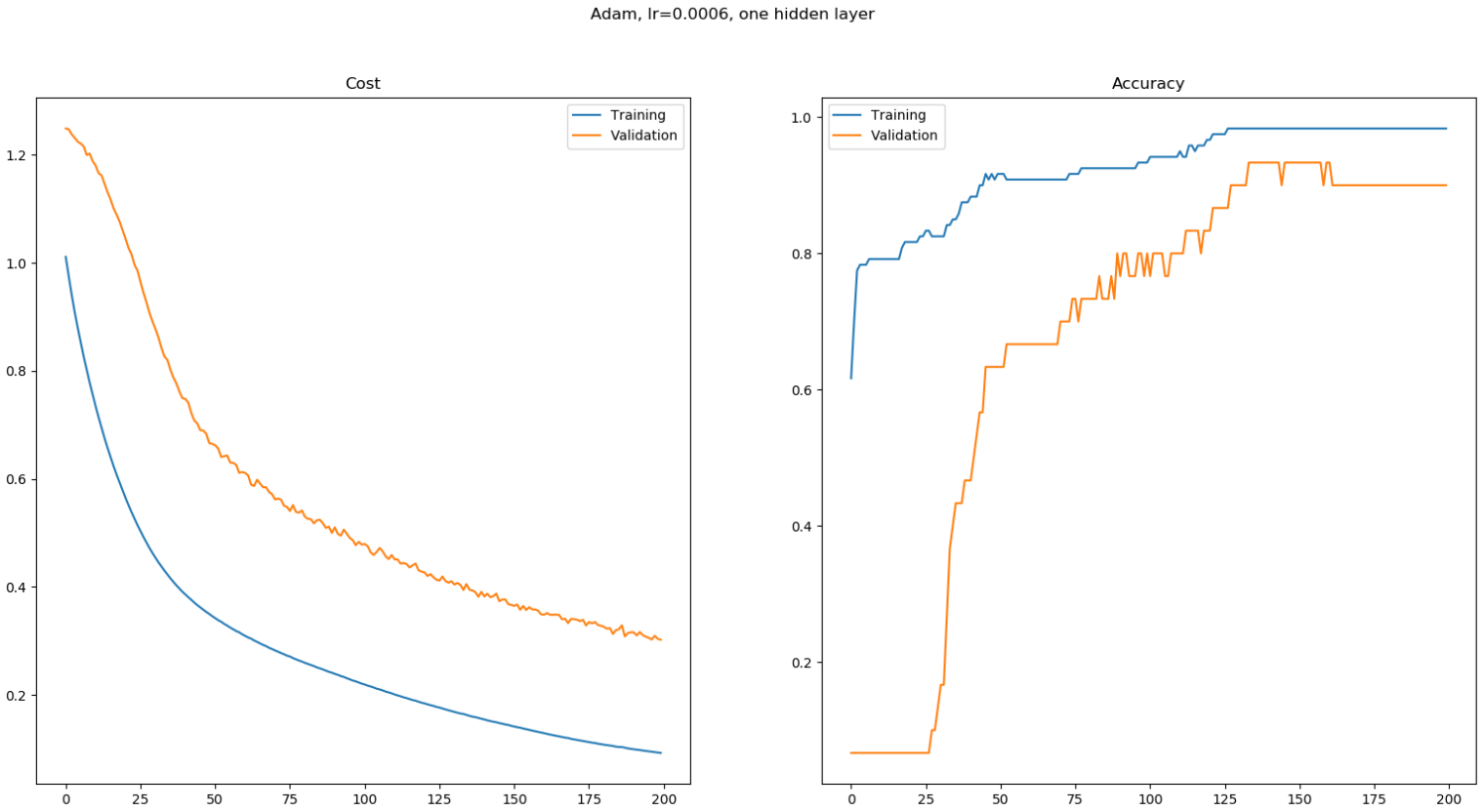
- Model parallelism

https://pytorch.org/tutorials/intermediate/model_parallel_tutorial.html

- 16-bit floating point precision (Apex)
  - But, only works on 2080ti

Debugging Tricks

Plot the loss curve



Sanity check

- Overfit a small dataset

- Select a small subset of sentences for classification
- Train it with 100 epochs
- see if accuracy approaches 100%

- &^$*(#$$!^&#*! generation

Weak model

- Hidden states need to be sufficiently large
- Generation tasks typically require more parameters than analysis/classification tasks (both in terms of hidden units and number of layers)
- Character-level models usually require more layers

Optimization

- SGD = stochastic gradient descent

- There are other gradient-based optimization algorithms:

  Nesterov, RMSPROP, Adam ……

- Each one has their own default parameters

  eg, Adam usually works with lr = 1e-4

Overfitting

- Overfitting is more and more of common in NLP
- Finetuning large LMs (GPT2/Bert) can easily overfit your dataset
- Early stopping, learning rate decay, dropout

Story Generation

Story Generation

- ROCStories (Mostafazadeh et al 2017)
- Hierarchical Neural Story Generation (Fan et al 2018)
- Plan-and-Write (Peng et al 2019)

ROCStories (Mostafazadeh et al 2017)

ROCStories

- Stories of five sentences

ROCStories

Title	Five-sentence Story
The Test	Jennifer has a big exam tomorrow. She got so stressed, she pulled an all-nighter. She went into class the next day, weary as can be. Her teacher stated that the test is postponed for next week. Jennifer felt bittersweet about it.
The Hurricane	Morgan and her family lived in Florida. They heard a hurricane was coming. They decided to evacuate to a relative's house. They arrived and learned from the news that it was a terrible storm. They felt lucky they had evacuated when they did.
Spaghetti Sauce	Tina made spaghetti for her boyfriend. It took a lot of work, but she was very proud. Her boyfriend ate the whole plate and said it was good. Tina tried it herself, and realized it was disgusting. She was touched that he pretended it was good to spare her feelings.

ROCStories

- Story Cloze Test

Story Cloze Test

Context	Right Ending	Wrong Ending
Karen was assigned a roommate her first year of college. Her roommate asked her to go to a nearby city for a concert. Karen agreed happily. The show was absolutely exhilarating.	Karen became good friends with her roommate.	Karen hated her roommate.
Jim got his first credit card in college. He didn't have a job so he bought everything on his card. After he graduated he amounted a $10,000 debt. Jim realized that he was foolish to spend so much money.	Jim decided to devise a plan for repayment.	Jim decided to open another credit card.
Gina misplaced her phone at her grandparents. It wasn't anywhere in the living room. She realized she was in the car before. She grabbed her dad's keys and ran outside.	She found her phone in the car.	She didn't want her phone anymore.

Deep Structured Semantic Model

- Construct several features based on semantic contexts (distribution of tri-gram letters)

- A pretty basic baseline (59.5 classification accuracy on a binary classification tasks)

A hard binary classification tasks

Rank	CodaLab Id	Model	ROCStories	Pre-trained Embeddings	Other Resources	Accuracy
1	**msap**	Logistic regression	Spring 2016, Winter 2017	–	NLTK Tokenizer, Spacy POS tagger	**0.752**
2	**cogcomp**	Logistic regression	Spring 2016, Winter 2017	Word2Vec	UIUC NLP pipeline, FrameNet, two sentiment lexicons	0.744
3	**tbmihaylov**	LSTM	–	Word2Vec	–	0.728
4	**ukp**	BiLSTM	Spring 2016, Winter 2017	GloVe	Stanford CoreNLP, DKPro TC	0.717
5	**acoli**	SVM	–	GloVe, Word2Vec	–	0.700
6	**roemmele**	RNN	Spring 2016, Winter 2017	Skip-Thought	–	0.672
7	**mflor**	Rule-based	–	–	VADER sentiment lexicon, Gigaword corpus PMI scores	0.621
8	**Pranav_Goel**	Logistic regression	Spring 2016, Winter 2017	Word2Vec	VADER sentiment lexicon, SICK data set	0.604
9	**ROCNLP (baseline)**	DSSM	Spring 2016, Winter 2017	–	–	0.595

Pretrained bert

Methods	Accuracy (%)
$\text{BERT}_{\text{BASE}}$ (multilingual, uncased)	75.9
$\text{BERT}_{\text{BASE}}$ (multilingual, cased)	80.2
$\text{BERT}_{\text{BASE}}$ (monolingual, cased)	87.4
$\text{BERT}_{\text{BASE}}$ (monolingual, uncased)	88.1
$\text{BERT}_{\text{LARGE}}$ (monolingual, uncased)	89.2
$\text{BERT}_{\text{LARGE}}$ (monolingual, cased)	**90.0**

Challenges with story generation

- Despite the fact that pretrained models can improve the analysis/classification tasks on story common sense significantly, generation remains hard

- **Title:** bicycle path accident
- **Generated Story:** sam bought a new bicycle. his bicycle was in an accident. his bicycle was in an accident. his bicycle was in an accident. his bicycle was totaled.
- **Title:** darth vader on earth
- **Generated Story:** it was a very windy day. i 've never been to it before. i do not know what to do. i do not know what to do. i think it is a good idea.

- Slide credit: Nanyun Peng, NAACl 2019

How about GPT2?

- **bicycle path accident**, but it's still a pretty bad event. But, you know, they're also doing other events. ... I'm pretty excited, actually. I'm really excited. And I hope, you know, there's some other people out there feeling it. …
- **darth vader on earth** !!!
  HOLY HOP! HE JUST WOKE UP ON HIM!!! He did nothing wrong, he's a guy just like us. But it's funny how when someone says, "I'm so grateful he's alive. I'm so happy to see you're still alive."
  …

From analyzing to generating

- Requires a deeper understanding and more comprehensive modelling on stories:
  - Themes
  - Characters
  - Genres
  - Point of views
  - Events
  - Ending spirits

Hierarchical Neural Story Generation (Fan et al 2018)

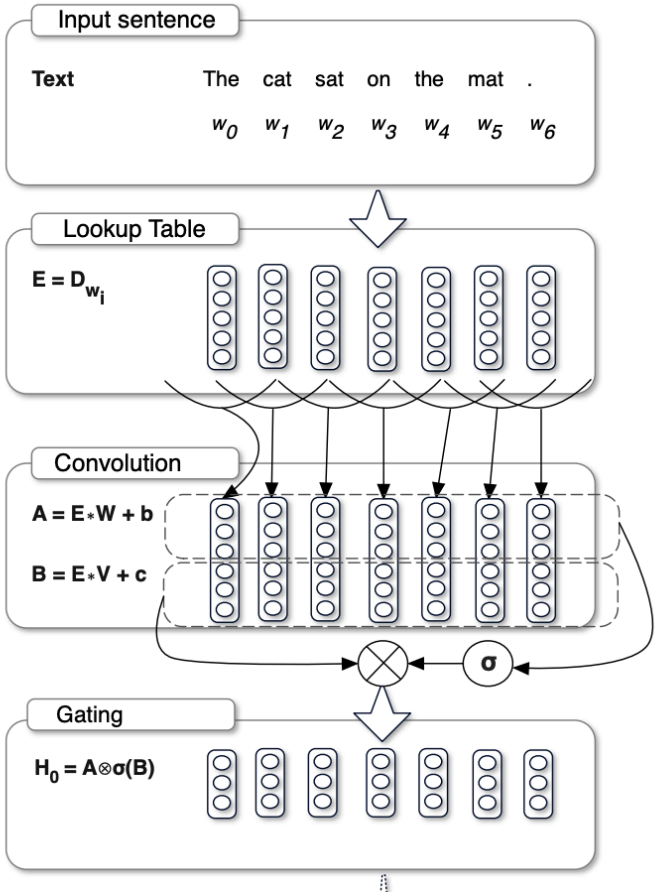
Hierarchical Neural Story Generation

Motivation:

- Most sequence to sequence models tend to ignore prompts
- Story needs a grounding

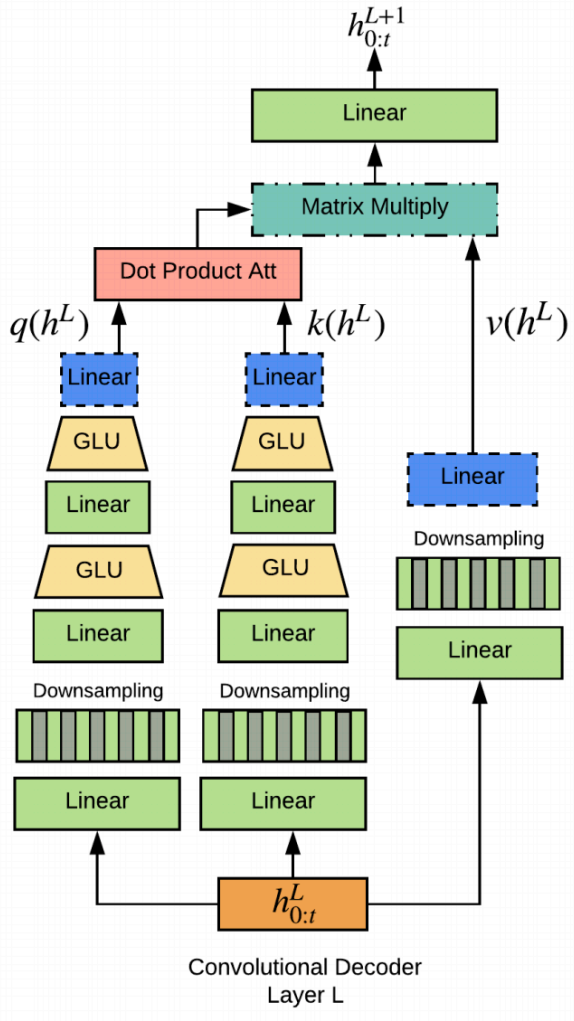
Gated Linear Unit

- A is trained to take in or shut down B, essentially acting as a gate
- Help with gradient vanishing



Convolutional seq2seq model

- Parallelable
- Fewer parameters (shared weights and location pooling)
- Allow longer contexts to be captured



Multi-Scale Attention

- have each head operating at a different time scale

- The downsampling operation limits the number of tokens in the attention maps, making them sharper

Examples

Example Prompt 2: The scientists have discovered something terrible .

The scientist stood there, a little dazed as he stared.
“What is it?” He asked.
“This...this...Thing...This is a virus. A chemical that can destroy entire planet and it is a very small, complex, chemical that could destroy any planet.” The scientist replied. His lab assistant looked down at the tablet.
“I’ve just discovered it. I can’t believe it. It looks like it’s made of some sort of chemical that’s very dangerous.”
“ Well, there’s a virus on the ground. It’s very effective...” “ I can’t believe what it is.” he said, pointing to the scientist .
“ We don’t know what this thing is. We haven’t seen anything like it . We can’t even see anything like this. ” Dr. Jones stared at the scientist for a moment.
“What do you mean what does it do ?”
“It...It ’s a monster.”

Plan-and-Write (Peng et al 2019)

Controlled story line

- Extracting Story Line
  - RAKE: Rapid Automatic Keyword Extraction

Title: christmas shopping

Story: frankie had christmas shopping to do.

she went to the store.

inside, she walked around looking for gifts.

soon her cart was full.

she paid and took her things home.

Storyline (unsupervised extraction): frankie store gifts cart paid

Title: farm

Story: bogart lived on a farm.

he loved bacon.

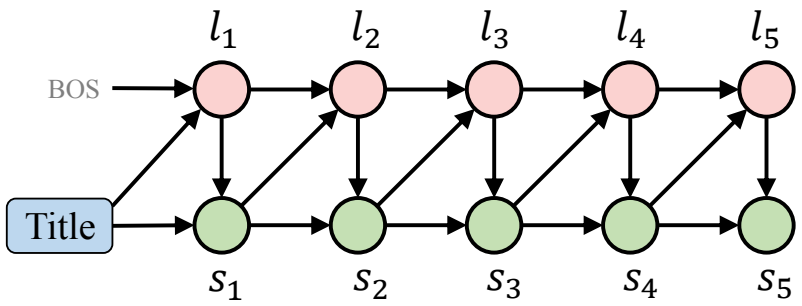
he decided to buy a pig. shortly after, he grew fond of the pig.

bogart stopped eating bacon.

Storyline (unsupervised extraction): farm bacon decided pig bogart

Dynamic vs Static Schema

Dynamic Schema



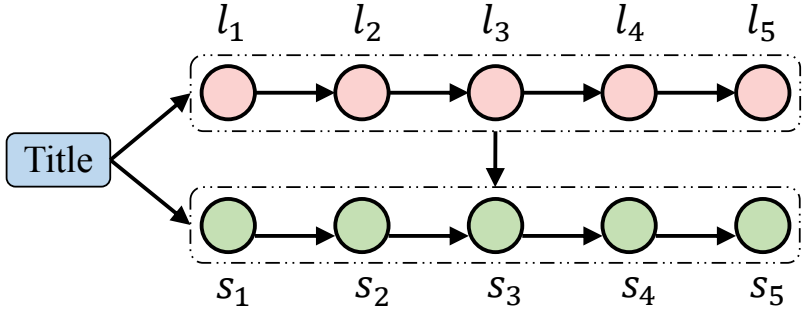
We define context as: $\boldsymbol{ctx} = [\boldsymbol{t}; \boldsymbol{s}_{1:i-1}]$

At the plan step, we model: $P(l_i|\boldsymbol{ctx}, l_{1:i-1})$

At the write step, we model: $P(s_i|\boldsymbol{ctx}, l_{1:i})$

The probabilities are computed by some specifically designed fusion-RNN cells.

Static Schema



At the plan step, we model: $P(l_i|\boldsymbol{t}, l_{i-1})$

At the write step, we model: $P(s_i|\boldsymbol{ctx}, l_{1:5})$

Examples from Plan and Write

Title: The Virus		
Baselines	Inc-S2S	His computer was fixed and he fixed it. John got a new computer on his computer. John was able to fix it himself. John was able to fix his computer and was able to fix his computer. John was able to fix his computer and had a virus and was able to fix his computer.
	Cond-LM	Tim was working on a project. He was working on a project. Tim was working on a project. The project was really good. Tim was able to finish the project.
Dynamic	Storyline	computer → use → anywhere → house → found
	Story	I was working on my computer today. I was trying to use the computer. I couldn't find it anywhere. I looked all over the house for it. Finally, i found it.
Static	Storyline	work → fix → called → found → day
	Story	I had a virus on my computer. I tried to fix it but it wouldn't work. I called the repair company. They came and found the virus. The next day, my computer was fixed.

Some examples generated by us

- The story prompt:

A soldier on the front dies in the middle of writing a letter home . It is finished and sent by the man who killed him .

- Generated by finetuned GPT2 model:

`` Sir, I am sorry for your loss. ''

`` I understand . ''

In the years since his loss, I have lost count in my mind as to these men's sacrifice. They were the most brave men I would ever kill. Not just one, but two, and as I said in the beginning of this post, each of the men I have come to know and love died in the same way. The same could not be said for your fallen comrade. You were there in your own flesh and blood, not just because you didn't die for them in a hundred thousand ways, but because you chose to. But you know not one of the others who chose to do so.

I have heard the stories of the brave men whose stories I have never heard before. Some tell of the ones who have died or survived

In-class Exercise

- Please read/finish the colab notebook and submit a pdf by 11:59pm