Resources:

1. Scheduled zoom link
   https://zoom.us/j/400853409
   以后都会使用此链接。
2. Google Calendar
   https://calendar.google.com/event?action=TEMPLATE&tmeid=N2Zvam11djVhYXFrZTRrMDhydnFjMzNpbWRfMjAyMDAyMTNUMjEzMDAwWiBndWFueWluZ2p1bi5kYXRqdUBt&tmsrc=guanyingjun.datju%40gmail.com&scp=ALL
3. GitHub links: [data mining read group]
4. Course webpage: http://www.phontron.com/class/nn4nlp2020/schedule.html

---

02/20 Group Study - Lecture 4: RNN for text

Attendance: Lan Li, Lan Jiang, Yingjun Guan, Ziqi Jiang, Yiren Liu, Yixing Hu, Xiaoliang Jiang

Course Material:
- 辅助ppt: [RNN]
- 视频学习 https://youtu.be/aA14_J-8HXU
- 浏览课件 http://www.phontron.com/class/nn4nlp2020/assets/slides/nn4nlp-04-rnn.pdf
- 阅读材料 GoldBerg's NNM4NLP Chap14-16
https://www.morganclaypool.com/doi/abs/10.2200/S00762ED1V01Y201703HLT037
- 如果你对Recurrent Networks, Vanishing Gradient and LSTMs, Strengths and Weaknesses of Recurrence in Sentence Modeling, Pre-training for RNNs的基础知识不很了解，请移步这里的其他资源 http://www.phontron.com/class/nn4nlp2020/schedule/rnn.html

Study group adjustments (for future).

1. Group members will lead discussion in turn (presenting in teams is ok).
   a. Some future leadings: {Conditioned generation: Ziqi, Attention: Yiren}
2. Codes are suggested in future lectures to help better understand the material. According to the discussion, the group will spend 5-10 min on codes. Members should try running codes before class.
3. Reorganizing the slides like today is highly recommended, but not required.
4. We could start using github to share interesting codes, test, projects, etc.
5. Responsibility for presenters:
   a. [Required] Lead discussion; ask questions; run codes successfully and lead.
   b. [suggested] make slides; read and summarize readings; recommend other relative materials.
6. Responsibility for attenders:
   a. [Required] Watch youtube videos; check slides; try to run codes.
   b. [suggested] Reading.

Discussion Notes:

Why RNN….

Liu: sequential + long-distance

Lan Jiang: coreference : 同样词指代不同东西

Loss function ?

Regression Loss Functions:
Mean Squared Error Loss
Mean Squared Logarithmic Error Loss
Mean Absolute Error Loss

Binary Classification Loss Functions:
Binary Cross-Entropy
Hinge Loss
Squared Hinge Loss

Multi-Class Classification Loss Functions:
Multi-Class Cross-Entropy Loss
Sparse Multiclass Cross-Entropy Loss
Kullback Leibler Divergence Loss

Total loss→ separate loss

Backpropagation的目的：调整separate loss.

Lan:
Wish to discuss Attention in more detail in future. (Yes. In two more weeks, we will do.)

Ziqi jiang:
Shared Parameter & Accumulated derivatives
Updating process: total loss→ loss i (where i = 1,2,...,m) → derivative^(i) → use sum(derivative^(i)) update the weight
Advantage:

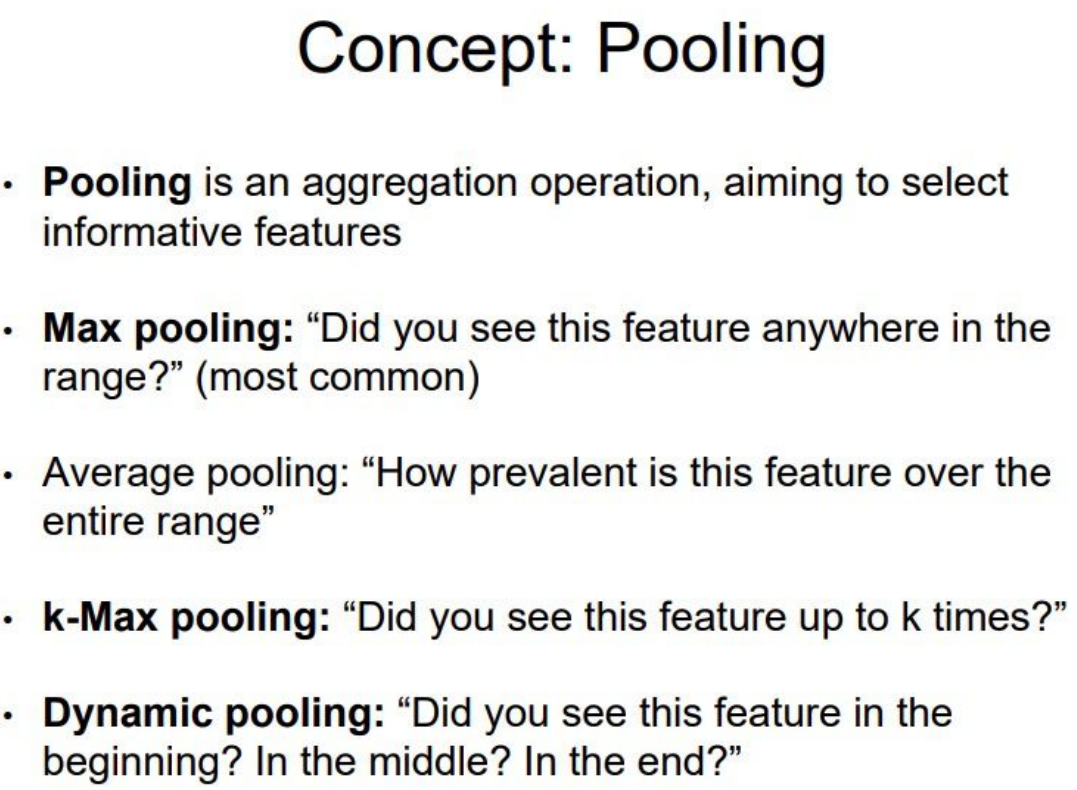
1. Updating parameters for the whole sentence
2. Computational complexity O(n) while updating in every step would require O(n^2) complexity.

02/18 Group Study - Lecture 3: CNN for Text

Attendance: Yiren Liu, Yingjun Guan, Liri Fang, Lan Li, Yixing Hu, Mengfei Lan, Gaozheng Liu.

Course Material
第三课： CNN for text
- 领讲：房丽日
- 视频学习
- 浏览课件
- 阅读材料 GoldBerg's NNM4NLP Chap13
- 如果你对n-gram, CNN, structured convolution, CNN visualization的基础知识不很了解，请移步这里的其他资源

Discussion Notes:
Fang:



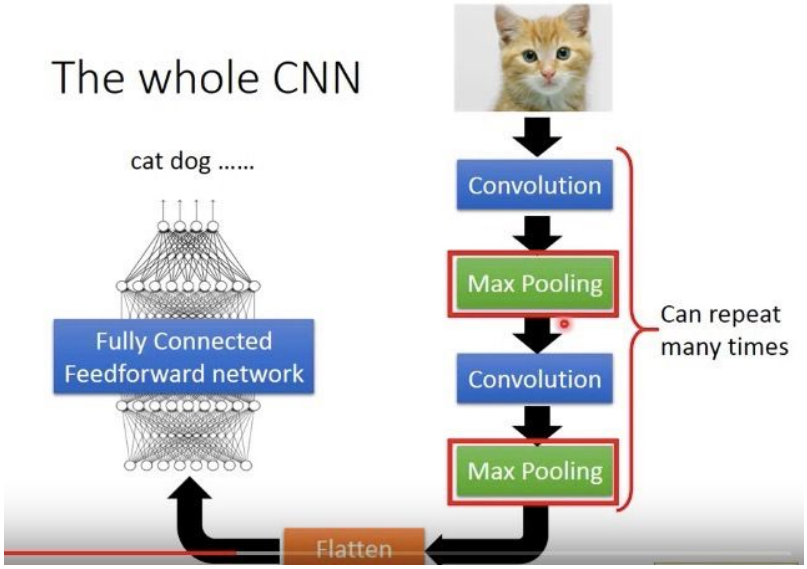
Workflow for NLP: (**Guan**)
Embedding (what's embedding?)→ Training → Testing
Related work:
BERT
General : Fine-tuning and Feature-based

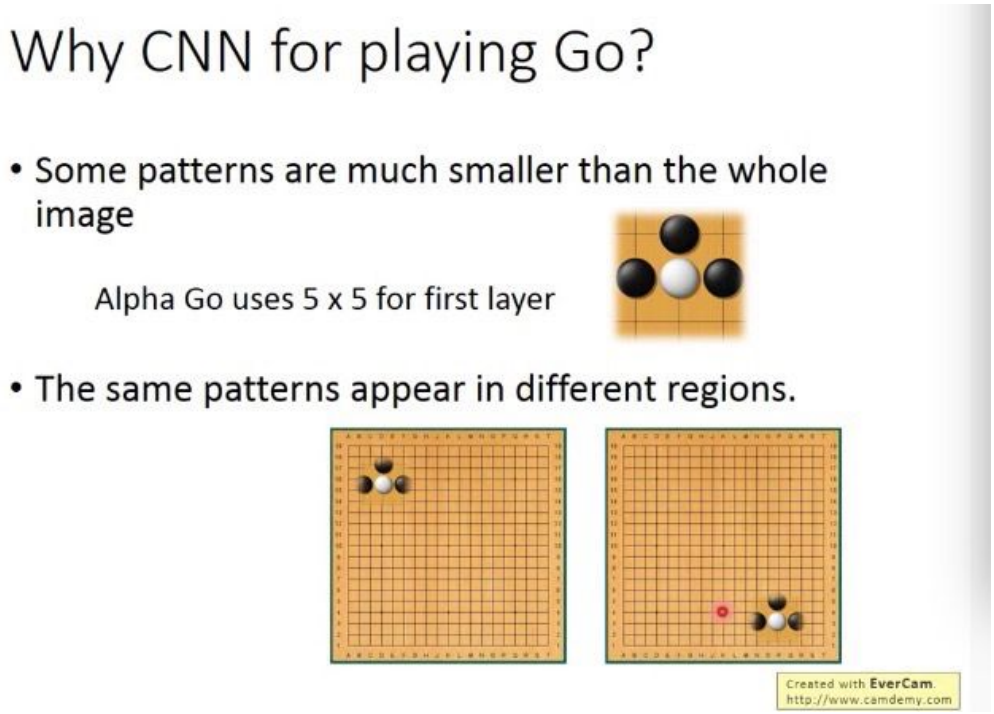
Yiren Liu: suggested : first fine-tuning with BERT, then training (https://github.com/google-research/bert)

> Using BERT has two stages: *Pre-training* and *fine-tuning*.
>
> **Pre-training** is fairly expensive (four days on 4 to 16 Cloud TPUs), but is a one-time procedure for each language (current models are English-only, but multilingual models will be released in the near future). We are releasing a number of pre-trained models from the paper which were pre-trained at Google. Most NLP researchers will never need to pre-train their own model from scratch.
>
> **Fine-tuning** is inexpensive. All of the results in the paper can be replicated in at most 1 hour on a single Cloud TPU, or a few hours on a GPU, starting from the exact same pre-trained model. SQuAD, for example, can be trained in around 30 minutes on a single Cloud TPU to achieve a Dev F1 score of 91.0%, which is the single system state-of-the-art.

Why padding? For boundary :
If the final number for the filter is 0, the last will not be sampled...

Lan: https://www.youtube.com/watch?v=FrKWiRv254g





Mengfei:
In NLP, can vector be parsed and CNNed?

02/13 Group Study - Lecture 2: Language Modeling

Attendance: Yingjun Guan, Xiaoliang Jiang, Ziqi Jiang, Lan Li, Liri Fang, Yiren Liu
Notification:

5. Scheduled zoom link
   https://zoom.us/j/400853409 以后都会使用此链接。
6. Google Calendar
   https://calendar.google.com/event?action=TEMPLATE&tmeid=N2Zvam11djVhYXFrZTRrMDhydnFjMzNpbWRfMjAyMDAyMTNUMjEzMDAwWiBndWFueWluZ2p1bi5kYXRqdUBt&tmsrc=guanyingjun.datju%40gmail.com&scp=ALL
7. Other **Useful links: [data mining read group]**

Course material:
第二课： language modeling.
- 视频学习
- 浏览课件
- 阅读资料 GoldBerg's NNM4NLP chap8-9
- 如果你对language modeling, feed-forward NN, optimization, language model evaluation的基础知识不熟悉的话，请移步这里的其他资源：

Discussion Notes:

Lan:

- potential related work：[Attention is all you need]

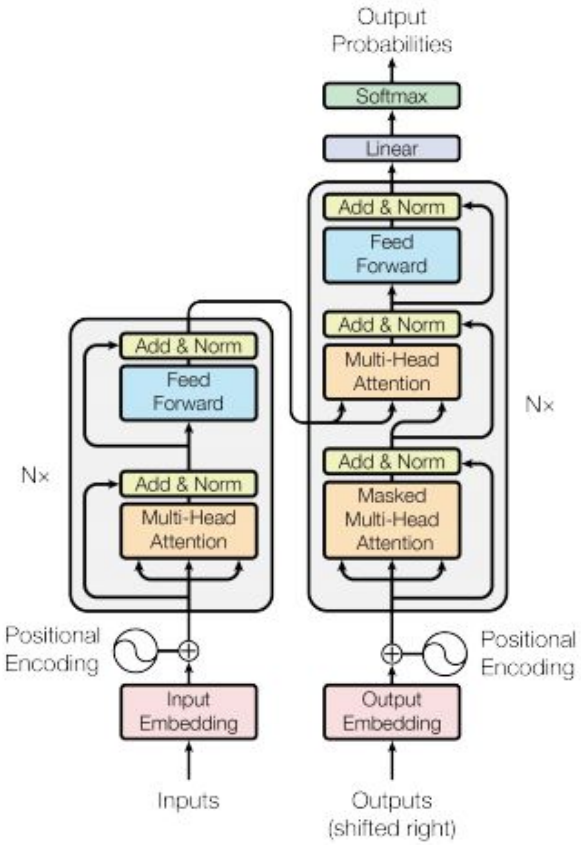


Figure 1: The Transformer - model architecture.

- Add on positional parameter
- Improve computation efficiency : parallelisation
- Page 6: self-attention could help with long-range dependencies
- The third is the path length between long-range dependencies in the network. Learning long-range dependencies is a key challenge in many sequence transduction tasks. One key factor affecting the ability to learn such dependencies is the length of the paths forward and backward signals have to traverse in the network. The shorter these paths between any combination of positions in the input and output sequences, the easier it is to learn long-range dependencies [12]. Hence we also compare the maximum path length between any two input and output positions in networks composed of the different layer types.

Xiaoliang:

1. **Count-based language modeling**
2. Bow, sequence/sequential data: the word after A and B vs the third word after 1st 2nd and before 4th 5th… ->speed up
3. **Generative model &&**
4. **Smoothing https://en.wikipedia.org/wiki/Kneser%E2%80%93Ney_smoothing**
5. **Back-off smoothing https://en.wikipedia.org/wiki/Katz%27s_back-off_model**
6. Log加和概率乘积
7. Loglikelihood: 整句， Per-word Log Likelihood: 除词个数， 词，

Ziqi
P26: What are Input Words, Output Words, context?

Liri:
Q: 同时生成第3词和第5词 paper:

- Non-Monotonic Sequential Text Generation by KH Cho, etc.

N: video on Momentum: https://www.youtube.com/watch?v=k8fTYJPd3_I

02/12 CBOW discussion

CBOW: page 117-118 [chap8-9]
CBOW is very similar to the traditional bag-of-words representation in which we discard order information, and works by either summing or averaging the embedding vectors of the corresponding features:

$$\text{CBOW}(f_1, \ldots, f_k) = \frac{1}{k} \sum_{i=1}^{k} v(f_i). \tag{8.1}$$

A simple variation on the CBOW representation is weighted CBOW, in which different vectors receive different weights:

$$\text{WCBOW}(f_1, \ldots, f_k) = \frac{1}{\sum_{i=1}^{k} a_i} \sum_{i=1}^{k} a_i v(f_i). \tag{8.2}$$

Here, each feature f_i has an associated weight a_i, indicating the relative importance of the feature. For example, in a document classification task, a feature f_i may correspond to a word in the document, and the associated weight a_i could be the word's TF-IDF score.

02/11 Group study - Lecture 1: General NLP

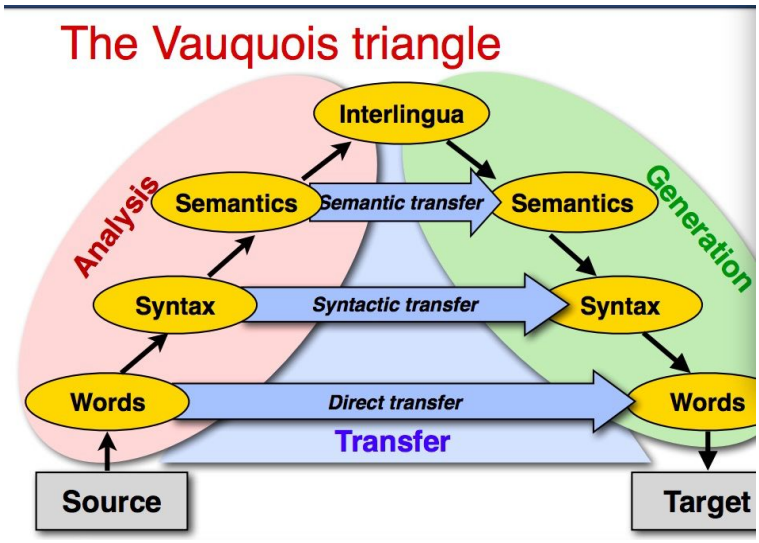
Attendence: Yingjun Guan, Liri Fang, Mengfei Lan, Yixing Hu, Yiren Liu, Xiaoliang Jiang, Lan Li

1. Self introduction.
2. http://phontron.com/class/nn4nlp2020/description.html
3. L1 revision. (http://phontron.com/class/nn4nlp2020/assets/slides/nn4nlp-01-intro.pdf)
4. L2 revision.
5. Phenomena to Handle
    a. **Morphology** eg. go/goed
    b. **Syntax** 句法？
        i. 主谓宾顺序？
        ii. 主谓宾缺失？
    c. **Semantics/World Knowledge**: eg. The store went to Jane.
    d. **Discourse**:
        i. 1.远距离？答非所问？Q：吃了啥 A:管老师好
        ii. 2. 代词？
    e. Pragmatics

    Pragmatic Analysis is part of the process of extracting information from text. Specifically, it's the portion that focuses on taking structures set of text and figuring out what the actual meaning was. It actually comes from the field of linguistics (as a lot of **NLP** does), where the context is considered from the text. → google search

    [dictionary]: the branch of linguistics dealing with language in use and the contexts in which it is used, including such matters as deixis, the taking of turns in conversation, text organization, presupposition, and implicature.

    f. Multilinguality



Lanjie 提到的三角形

Liri Fang: dialect → ?

Xiaoliang: 我就来记几个专业名词

1. **Sentence Classification**
2. **Computation Graphs**
3. **Forward propagation; backward propagation.**
4. **Keras**
5. **AutoGrad**
6.

BOW vs CBOW vs deepCBOW

6. Study plan in future:

Time: Tuesday, Thursday 3:30-4:40pm

https://github.com/uiuc-dm-group/DMRG-20SP