自然语言处理基础
多标签分类

杨鹏程
北京大学
yang_pc@pku.edu.cn

Outline

1. Background knowledge
   - Highway network.
   - Sequence-to-sequence model.
   - **Reinforcement learning:** policy gradient, SeqGAN, self-critical training.
2. Introduction to multi-label classification.
3. Sequence generation model for multi-label classification.
4. Experience.

Highway Network (Srivastava et al., 2015)

Motivation:

1. The depth of neural networks is crucial.
2. Training becomes more difficult as depth increases.
    - The problem of vanishing gradients.

Highway Network (Srivastava et al., 2015)

Motivation:

1. The depth of neural networks is crucial.
2. Training becomes more difficult as depth increases.
   - The problem of vanishing gradients.

Proposal:

1. Allow unimpeded information flow across many layers on *information highways*.
2. Use the adaptive gate.

Highway Network (Srivastava et al., 2015)

Motivation:

1. The depth of neural networks is crucial.
2. Training becomes more difficult as depth increases.
   - The problem of vanishing gradients.

Proposal:

1. Allow unimpeded information flow across many layers on *information highways*.
2. Use the adaptive gate.

Conclusion:

1. Can be trained directly using SGD, in contrast to plain networks which become hard to optimize. as depth increases.

Highway Network (Srivastava et al., 2015)

Plain network: $y = H(x, \boldsymbol{W_H})$

Highway Network (Srivastava et al., 2015)

Plain network: $y = H(x, \boldsymbol{W_H})$

Highway network:

1. General:

$$y = H(x, \boldsymbol{W_H}) \odot T(x, \boldsymbol{W_T}) + x \odot C(x, \boldsymbol{W_C}) \tag{1}$$

2. Couple:

$$y = H(x, \boldsymbol{W_H}) \odot T(x, \boldsymbol{W_T}) + x \odot (1 - T(x, \boldsymbol{W_T})) \tag{2}$$

3. Two gates:
   - **Transform gate** H**:** determine the information that x need to be transformed.
   - **carry gate** C**:** determine the information that x need to be carried.

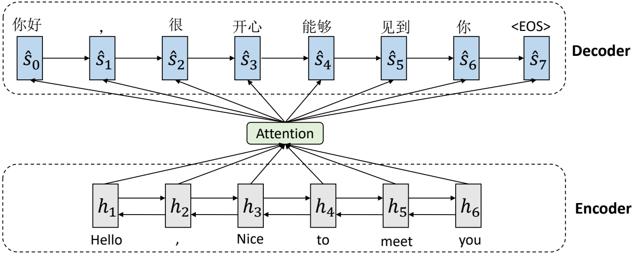
Highway Network (Srivastava et al., 2015)

1. **Highway network**: $y = H(x, \boldsymbol{W_H}) \odot T(x, \boldsymbol{W_T}) + x \odot C(x, \boldsymbol{W_C})$
2. The dimensionality of $x, y, H(x, \boldsymbol{W}_H)$ and $T(x, W_T)$ must be the same.

$$y = H(x, \boldsymbol{W_H}) \odot T(x, \boldsymbol{W_T}) + x \odot (1 - T(x, \boldsymbol{W_T})) \tag{3}$$

- Sub-sampling or zero-padding.
- Use plain layer to change dimensionality.

Sequence-to-Sequence Model (Bahdanau et al., 2014)



1. **Encoder:** Transform the word into dense representations.
2. **Attention:** Select the most important source words when generating different target words.
3. **Decoder:** Generate target words sequentially.

Introduction to Reinforcement Learning

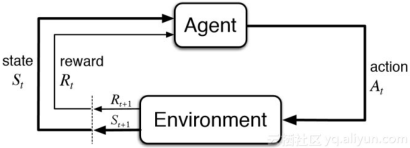
Reinforcement Learning:

1. Different from supervised, unsupervised and semi-supervised learning.
2. **Characteristics:**
    - Don't need labeled data.
    - Still be able to achieve the goal (reward).

Introduction to Reinforcement Learning

Reinforcement Learning:

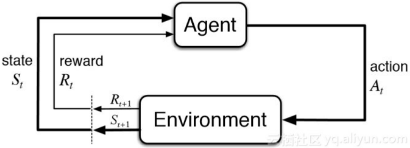
1. **Key concepts:**
   - Agent, state, action, environment, reward.



Introduction to Reinforcement Learning

Reinforcement Learning:

1. **Key concepts:**
   - Agent, state, action, environment, reward.



2. **Training:** maximize the expected cumulative reward.

$$L(\theta) = E(r_1 + \gamma r_2 + \gamma^2 r_3 + \cdots | \pi(, \theta)) \quad (4)$$

Policy Gradient in Text Generation (Ranzato et al., 2015)

Motivation:

- View the text generation task as a sequence decision process.

Model:

1. **Agent:** generative model / decoder.
2. **State:** the current produced words $(y_1, \cdots, y_{t-1})$.
3. **Action:** the word y_t to be generated.
4. **Reward:** the sentence-level reward (BLEU, Rouge).

Policy Gradient in Text Generation (Ranzato et al., 2015)

Motivation:

- View the text generation task as a sequence decision process.

Model:

1. **Agent:** generative model / decoder.
2. **State:** the current produced words $(y_1, \cdots, y_{t-1})$.
3. **Action:** the word y_t to be generated.
4. **Reward:** the sentence-level reward (BLEU, Rouge).

Training:

$$L(\theta) = -\mathbb{E}_{\boldsymbol{y}^s \sim p_\theta}[r(\boldsymbol{y}^s)] \tag{5}$$

$$\nabla_\theta L(\theta) = -\mathbb{E}_{\boldsymbol{y}^s \sim p_\theta}[r(\boldsymbol{y}^s)\nabla_\theta \log(p_\theta(\boldsymbol{y}^s))] \tag{6}$$

Self-Critical Training (Rennie et al., 2016)

Motivation:

$$\nabla_\theta L(\theta) = -\mathbb{E}_{\boldsymbol{y}^s \sim p_\theta}[(r(\boldsymbol{y}^s) - b)\nabla_\theta \log(p_\theta(\boldsymbol{y}^s))] \tag{7}$$

1. Although the estimation is unbiased, but it suffers high variance.
2. Use the baseline b to reduce variance.

Self-Critical Training (Rennie et al., 2016)

Motivation:

$$\nabla_\theta L(\theta) = -\mathbb{E}_{\boldsymbol{y}^s \sim p_\theta}[(r(\boldsymbol{y}^s) - b)\nabla_\theta \log(p_\theta(\boldsymbol{y}^s))] \tag{7}$$

1. Although the estimation is unbiased, but it suffers high variance.
2. Use the baseline b to reduce variance.
3. How to calculate the baseline?
    - Hyper-parameter.
    - MLP.
    - Self-critical

Self-Critical Training (Rennie et al., 2016)

Self-critical: use the reward of the generated sequence through testing algorithm.

1. **Advantages:**
    - Reduce variance.
    - Enhance the consistency of the model training and testing, which can effectively avoid exposure bias.

Self-Critical Training (Rennie et al., 2016)

Self-critical: use the reward of the generated sequence through testing algorithm.

1. **Advantages:**
   - Reduce variance.
   - Enhance the consistency of the model training and testing, which can effectively avoid exposure bias.
2. **Conclusion:**
   - Lower variance and better performance.

SeqGAN (Yu et al., 2017)

Motivation:

1. GAN has enjoyed considerable success in generating images.
2. Limitations on text generation task.
   - Discrete outputs from the generative model make it difficult to pass the gradient.
   - Discriminative model can only assess a complete sequence.

SeqGAN (Yu et al., 2017)

Proposal:

1. The reward r is defined as the probability that the discriminator D considers the sample to be true.
2. Use Monte-Carlo search method to evaluate a complete sequence.

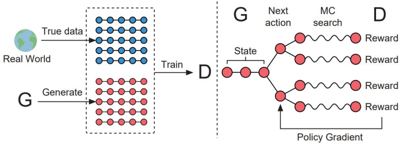


Figure 1: The illustration of SeqGAN. Left: D is trained over the real data and the generated data by G. Right: G is trained by policy gradient where the final reward signal is provided by D and is passed back to the intermediate action value via Monte Carlo search.

SeqGAN (Yu et al., 2017)

Proposal:

1. The reward r is defined as the probability that the discriminator D considers the sample to be true.
2. Use Monte-Carlo search method to evaluate a complete sequence.

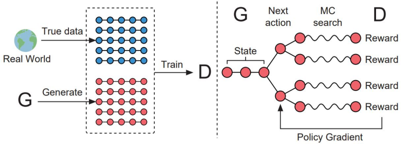


Figure 1: The illustration of SeqGAN. Left: D is trained over the real data and the generated data by G. Right: G is trained by policy gradient where the final reward signal is provided by D and is passed back to the intermediate action value via Monte Carlo search.

Conclusion:

1. Performs better than MLE and policy gradient with sentence-level reward.

What is multi-label classification?

1. **Definition:**
   - Assign multiple labels to each sample in the dataset.

What is multi-label classification?

1. **Definition:**
   - Assign multiple labels to each sample in the dataset.
2. **Applications:**
   - Text categorization.
   - Tag recommendation.
   - Information retrieval.

Example

[1] **arXiv:1807.01996 [pdf]**

A Formal Ontology-Based Classification of Lexemes and its Applications
Sreekavitha Parupalli, Navjyoti Singh
Comments: Accepted as Oral Presentation at Second Edition of Widening Natural Language Processing (WiNLP) workshop in 16th Anr
substantial text overlap with arXiv:1804.02186
Subjects: **Computation and Language (cs.CL)**

[2] **arXiv:1807.01956 [pdf, ps, other]**

Neural Language Codes for Multilingual Acoustic Models
Markus Müller, Sebastian Stüker, Alex Waibel
Comments: 5 pages, 3 figures, accepted at Interspeech 2018
Subjects: **Computation and Language (cs.CL)**; Machine Learning (cs.LG); Sound (cs.SD); Audio and Speech Processing (eess.AS)

[3] **arXiv:1807.01882 [pdf, ps, other]**

Chinese Lexical Analysis with Deep Bi-GRU-CRF Network
Zhenyu Jiao, Shuqi Sun, Ke Sun
Comments: 10 pages, 1 figure, 4 tables
Subjects: **Computation and Language (cs.CL)**

[4] **arXiv:1807.01855 [pdf]**

Zipf's law in 50 languages: its structural pattern, linguistic interpretation, and cognitive mo
Shuiyuan Yu, Chunshan Xu, Haitao Liu
Comments: 18 pages, 3 figures
Subjects: **Computation and Language (cs.CL)**

[5] **arXiv:1807.01763 [pdf, other]**

Seq2RDF: An end-to-end application for deriving Triples from Natural Language Text
Yue Liu, Tongtao Zhang, Zhicheng Liang, Heng Ji, Deborah L. McGuinness
Comments: Proceedings of the ISWC 2018 Posters & Demonstrations
Subjects: **Computation and Language (cs.CL)**; Artificial Intelligence (cs.AI)

Background

Mainstream models:

1. Problem transformation methods.
2. Algorithm adaptation methods.
3. Ensemble methods.
4. Neural network models.

Background

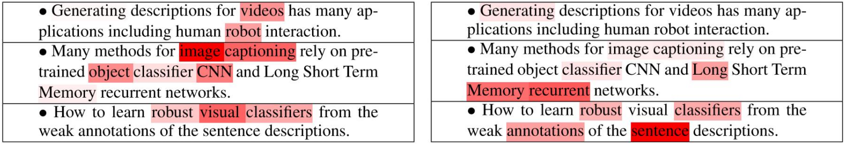
Previous work:

1. Can't capture label correlations very well or is computationally intractable..
   - **Label correlations:** Some labels are closely correlated.
   - **Example:** (green, leaf) or (green, dog).

Background

Previous work:

1. Can't capture label correlations very well or is computationally intractable..
   - **Label correlations:** Some labels are closely correlated.
   - **Example:** (green, leaf) or (green, dog).
2. Ignore differences in the contributions of textual content when predicting labels.



(a) Visual analysis when the SGM model predicts "*CV*".

(b) Visual analysis when the SGM model predicts "*CL*".

Figure 1: Visualization of attention.

Sequence Generation Model

Transform classification task into generation task.

1. **Key idea:**
   - View the text as the source language and label as target language.
   - Base on sequence-to-sequence model.

Sequence Generation Model

Transform classification task into generation task.

1. **Key idea:**
    - View the text as the source language and label as target language.
    - Base on sequence-to-sequence model.
2. **Advantages:**
    - **Capture label correlations:** Generate labels sequentially, and predicts the next label based on its previously generated labels.
    - **Model differences in contributions of textual content:** Apply the attention mechanism.

Sequence Generation Model

Motivations:

1. Capture label correlations.
2. Model differences in contributions of textual content.

Sequence Generation Model

Motivations:

1. Capture label correlations.
2. Model differences in contributions of textual content.

Difficulties and solutions:

1. **Repeated labels:**
   - Use the record module to smooth the probability distribution.

Sequence Generation Model

Motivations:

1. Capture label correlations.
2. Model differences in contributions of textual content.

Difficulties and solutions:

1. **Repeated labels:**
   - Use the record module to smooth the probability distribution.
2. **Exposure bias:**
   - Use highway network to introduce the global information of previous time-steps.

Sequence Generation Model

Motivations:

1. Capture label correlations.
2. Model differences in contributions of textual content.

Difficulties and solutions:

1. **Repeated labels:**
   - Use the record module to smooth the probability distribution.
2. **Exposure bias:**
   - Use highway network to introduce the global information of previous time-steps.
3. **Sequence order:**
   - Sort the label sequence according to the frequency.
   - Sequence-to-set model (another work).

SGM

Proposed model:

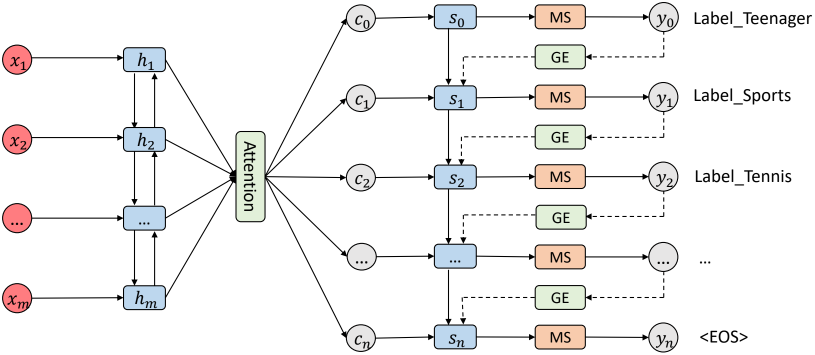


Figure 2: The overview of SGM. MS denotes the masked softmax layer. GE denotes the global embedding.

Sequence-to-Set Model

Motivation: Address the **order** problem of sequence generation.

1. The Seq2Seq model requires humans to predefine the order of output labels.
2. The output labels are **an unordered set rather than an ordered sequence (swapping-invariance property)**.

Sequence-to-Set Model

Motivation: Address the **order** problem of sequence generation.

1. The Seq2Seq model requires humans to predefine the order of output labels.
2. The output labels are **an unordered set rather than an ordered sequence (swapping-invariance property)**.
3. **Wrong penalty:**
   - Model may be wrongly penalized by the MLE method due to **inconsistent label order**.
   - Correct: [A, B, C]
     Wrong: [C, A, B]

Sequence-to-Set Model

Seq2Set model: By means of **reinforcement learning**.

1. **Key idea:**
   - **Policy gradient:** Reward r is independent of the label order.
   - **Bi-decoders:** Sequence-decoder fuses **human prior knowledge** and set-decoder satisfies the **swapping-invariance property** of the output label set.
2. **Overview of model:**

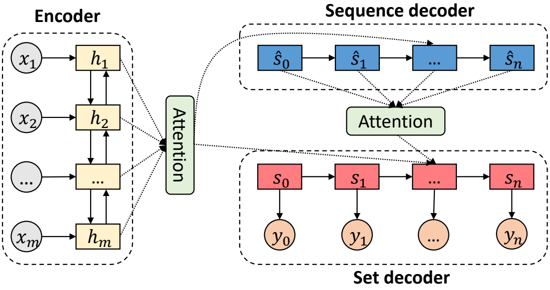


Figure 3: The overview of sequence-to-set model.

找准自己的定位

1. 职业发展
   - Research、Coding、Program Manager...
2. 研究方向
   - 基础理论、具体应用；图像、自然语言处理...

良好的编程基础

1. 熟练使用Python
2. 至少精通一门深度学习框架语言
    - Tensorflow, Pytorch, Keras...
3. 形成自己的基准代码体系
    - 可以基于自己的基准代码体系快速实现自己的idea.
    - 自己研究领域的经典模型和主流模型.

扎实的理论基础

1. AI的基础知识
   - 如CNN, RNN，Q-learning，Policy gradient等.
2. 了解自己研究方向上的前沿热点
   - 掌握自己研究领域的背景知识.
   - 把握自己研究领域的发展方向.
   - 深耕于自己的研究领域.

Inner Peace

1. 想不出idea, 发不出论文
   - 尝试做一个应用型任务, 学以致用.
   - 多读论文, 厚积薄发.
   - 重视小的创新, 循序渐进, 积少成多.
2. Idea成功
   - 能否进一步创新, 拉大与baseline的gap.
   - 成功的idea是否存在什么问题, 能否进一步解决.
   - 如何解释idea成功的原因?
   - 如何设计全方位的实验来验证你的motivation?

DEECAMP

THANK YOU