走近 NLP

Bigluo

Chapter 1

NLP 团队总览

1.1 HuggingFace 抱抱脸

Hugging face 是一家总部位于纽约的聊天机器人初创服务商，开发的应用在青少年中颇受欢迎，相比于其他公司，Hugging Face 更加注重产品带来的情感以及环境因素。但更令它广为人知的是 Hugging Face 专注于 NLP 技术，拥有大型的开源社区。尤其是在 github 上开源的自然语言处理，预训练模型库 Transformers，已被下载超过一百万次，HiggingFace repo

- 2019 Oct, *Distill Bert* [1]

1.2 DeepMind

1.3 Google AI 谷歌人工智能

- 2017 Jun, *Transformer* [2]
- 2018 Apr, *Adafactor* [3]
- 2018 Oct, *Bert* [4]
- 2019 Sep, *Albert* [5]
- 2019 Oct, *T5* [6]
- 2020 Jan, *Reformer* [7]
- 2020 Feb, *GLU-Variants* [8] 各种门控线性单元变体
- 2020 Feb, *SimCLR* [9]
- 2021 Aug, *Sentence T5 (ST5)* [10]

1.4 FAIR facebook 人工智能研究院

- 2019 Apr, *Poly-Encoders* [11]
- 2019 Jul, *Roberta* [12]

- 2019 Aug, *Vilbert* [13]
- 2019 Nov, *MoCo* [14]
- 2020 Mar, *MoCo-2* [15]

1.5 OpenAI

- 2018, *GPT-1* [16]
- 2019, *GPT-2* [17]
- 2019 Apr, *Sparse Transformer* [18]
- 2020 May, *GPT-3* [19]
- 2021 Feb, *CLIP* [20]

1.6 Microsoft Research 微软研究院

- 2017 Dec, *LightGBM* [21]
- 2019 Jan, *Multi-Task DNN* [22]
- 2019 May, *UniLM* [23]

1.7 AllenAI

- 2016 Feb, *Weight Normalization* [24]
- 2020 May, *UnifiedQA* [25]
- 2021 Mar, *Unicorn* [26]

1.8 Tsinghua

- 2020 Dec, *CPM-1* [27]
- 2021 Jun, *CPM-2* [28]
- 2021 Mar, *M6* [29]
- 2021 Apr, *SimCSE* [30]

1.9 BAAI 北京智源人工智能研究院

- 2021 Jun, *CPM-2* [28]

1.10 Alibaba 阿里巴巴

- 2021 Mar, *M6* [29]

1.11 Baidu AIG 百度人工智能研究院

- 2019 Apr, *ERNIE-1.0* [31]
- 2019 Jul, *ERNIE-2.0* [32]
- 2019 Oct, *PLATO-1* [33]
- 2020 Jun, *PLATO-2* [34]
- 2020 Jun, *ERNIE-vil* [35]
- 2021 Feb, *Knover* [36]

1.12 IFlyTek 科大讯飞

- 2019 Jun, *Bert-wwm* [37]
- 2020 Mar, *Electra* [38]
- 2020 Apr, *MacBert* [39]

Chapter 2

基于 Encoder 的模型

Chapter 3

基于 Decoder 的模型

Chapter 4

基于 Encoder-Decoder 的模型

4.1 Transformer

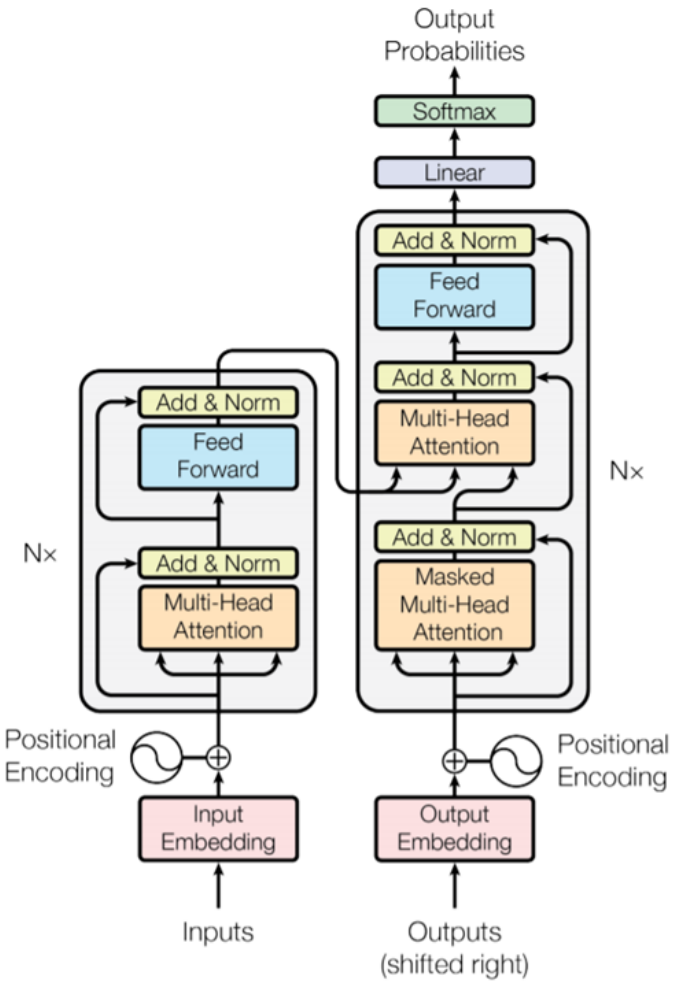

图 4.1: Transformer 模型架构

- Input Embedding: 输入的 token sequence embedding
- Positional Encoding 位置编码满足下列特性 1) 每个位置有唯一的位置编码

$$PE(pos, 2i) = sin(\frac{pos}{1000^{\frac{2i}{d_{model}}}}) \tag{4.1}$$

$$PE(pos, 2i+1) = cos(\frac{pos}{1000^{\frac{2i}{d_{model}}}}) \tag{4.2}$$

2) 由三角函数加法公式可求得相对距离 $|pos_1 - pos_2|$, 但无法确定两个位置谁处于前面

$$cos((pos_1 - pos_2) * i) = cos(pos_1 * i)cos(pos_2 * i) + sin(pos_1 * i)sin(pos_2 * i) \tag{4.3}$$

- Multi-Head Attention, 多头注意力模型
  1. $Q_t = XW_Q, K_t = YW_K, V_t = ZW_V$; 一般 Y, Z 相同, 若 X 也相同则称为自注意力, <bs, seq, d>
  2. Q_t.reshape(bs, n, seq, d/n), K_t.reshape(bs, n, seq, d/n), V_t.reshape(bs, n, seq, d/n); n=#heads
  3. Attention(Q_t,K_t,V_t)=softmax($\frac{Q_t K_t^T + mask}{\sqrt{d_k}}$)$V_t$; softmax 用于归一化权重, 分母用于归一化协方差, mask 矩阵中负无穷元素值表示 masked token 或 pad;
  4. Attention.reshape(bs, seq, d)
- FNN 前向回馈网络: $activate(xW_1 + b_1)W_2 + b_2$
  - ▷ $W_1, < d, d_{ff} >$
  - ▷ $W_2, < d_{ff}, d >$
- Output Embedding: 同语种时与 Input Embedding 一致, 不同语种时 (如处理翻译任务) 则是另一 Learnable Embedding

4.2 T5

4.3 Sentence T5 (ST5)

Chapter 5

对比学习

Chapter 6

多模态

Chapter 7

基本组件

7.1 Attention 注意力机制

Neural machine translation by jointly learning to align and translate [40] 是由德国不来梅雅各布大学 Jacobs University Bremen、加拿大蒙特利尔大学 Université de Montréal 联合发表的工作。该论文第一次将注意力机制引入了 NLP 领域 (论文中工作为基于 Seq2Seq 的 NMT 任务)

7.2 Normalization 归一化

7.2.1 WeightNormalization 权重归一化

7.2.2 BatchNormalization (BN) 批量归一化

7.2.3 LayerNormalization (LN) 层归一化

Chapter 8

提升技巧

Chapter 9

Optimizer 优化器

Chapter 10

集成学习

附录

.1 Acronyms 缩略词

A

B

BN Batch Normalization, 批量归一化

C

D

E

F

G

H

I

J

K

L

LN Layer Normalization, 层归一化

M

N

NLP Natural Language Processing, 自然语言处理

NMT Neural Machine Translation, 神经机器翻译

O

P

Q

R

S

T

U

V

W

X

Y

Z

.2 Glossary 词汇表

参考文献

[1] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, pp. 5998–6008, 2017.

[3] N. Shazeer and M. Stern, "Adafactor: Adaptive learning rates with sublinear memory cost," in *International Conference on Machine Learning*, pp. 4596–4604, PMLR, 2018.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[5] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2019.

[6] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *arXiv preprint arXiv:1910.10683*, 2019.

[7] N. Kitaev, Ł. Kaiser, and A. Levskaya, "Reformer: The efficient transformer," *arXiv preprint arXiv:2001.04451*, 2020.

[8] N. Shazeer, "Glu variants improve transformer," *arXiv preprint arXiv:2002.05202*, 2020.

[9] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, pp. 1597–1607, PMLR, 2020.

[10] J. Ni, N. Constant, J. Ma, K. B. Hall, D. Cer, Y. Yang, *et al.*, "Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models," *arXiv preprint arXiv:2108.08877*, 2021.

[11] S. Humeau, K. Shuster, M.-A. Lachaux, and J. Weston, "Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring," *arXiv preprint arXiv:1905.01969*, 2019.

[12] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[13] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *arXiv preprint arXiv:1908.02265*, 2019.

[14] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

[15] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.

[16] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018.

[17] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[18] R. Child, S. Gray, A. Radford, and I. Sutskever, "Generating long sequences with sparse transformers," *arXiv preprint arXiv:1904.10509*, 2019.

[19] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[20] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," *arXiv preprint arXiv:2103.00020*, 2021.

[21] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, vol. 30, pp. 3146–3154, 2017.

[22] X. Liu, P. He, W. Chen, and J. Gao, "Multi-task deep neural networks for natural language understanding," *arXiv preprint arXiv:1901.11504*, 2019.

[23] L. Dong, N. Yang, W. Wang, F. Wei, X. Liu, Y. Wang, J. Gao, M. Zhou, and H.-W. Hon, "Unified language model pre-training for natural language understanding and generation," *arXiv preprint arXiv:1905.03197*, 2019.

[24] T. Salimans and D. P. Kingma, "Weight normalization: A simple reparameterization to accelerate training of deep neural networks," *Advances in neural information processing systems*, vol. 29, pp. 901–909, 2016.

[25] D. Khashabi, S. Min, T. Khot, A. Sabharwal, O. Tafjord, P. Clark, and H. Hajishirzi, "Unifiedqa: Crossing format boundaries with a single qa system," *arXiv preprint arXiv:2005.00700*, 2020.

[26] N. Lourie, R. L. Bras, C. Bhagavatula, and Y. Choi, "Unicorn on rainbow: A universal commonsense reasoning model on a new multitask benchmark," *arXiv preprint arXiv:2103.13009*, 2021.

[27] Z. Zhang, X. Han, H. Zhou, P. Ke, Y. Gu, D. Ye, Y. Qin, Y. Su, H. Ji, J. Guan, *et al.*, "Cpm: A large-scale generative chinese pre-trained language model," *AI Open*, vol. 2, pp. 93–99, 2021.

[28] Z. Zhang, Y. Gu, X. Han, S. Chen, C. Xiao, Z. Sun, Y. Yao, F. Qi, J. Guan, P. Ke, *et al.*, "Cpm-2: Large-scale cost-effective pre-trained language models," *arXiv preprint arXiv:2106.10715*, 2021.

[29] J. Lin, R. Men, A. Yang, C. Zhou, M. Ding, Y. Zhang, P. Wang, A. Wang, L. Jiang, X. Jia, *et al.*, "M6: A chinese multimodal pretrainer," *arXiv preprint arXiv:2103.00823*, 2021.

[30] T. Gao, X. Yao, and D. Chen, "Simcse: Simple contrastive learning of sentence embeddings," *arXiv preprint arXiv:2104.08821*, 2021.

[31] Y. Sun, S. Wang, Y. Li, S. Feng, X. Chen, H. Zhang, X. Tian, D. Zhu, H. Tian, and H. Wu, "Ernie: Enhanced representation through knowledge integration," *arXiv preprint arXiv:1904.09223*, 2019.

[32] Y. Sun, S. Wang, Y. Li, S. Feng, H. Tian, H. Wu, and H. Wang, "Ernie 2.0: A continual pre-training framework for language understanding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 8968–8975, 2020.

[33] S. Bao, H. He, F. Wang, H. Wu, and H. Wang, "Plato: Pre-trained dialogue generation model with discrete latent variable," *arXiv preprint arXiv:1910.07931*, 2019.

[34] S. Bao, H. He, F. Wang, H. Wu, H. Wang, W. Wu, Z. Guo, Z. Liu, and X. Xu, "Plato-2: Towards building an open-domain chatbot via curriculum learning," *arXiv preprint arXiv:2006.16779*, 2020.

[35] F. Yu, J. Tang, W. Yin, Y. Sun, H. Tian, H. Wu, and H. Wang, "Ernie-vil: Knowledge enhanced vision-language representations through scene graphs," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 3208–3216, 2021.

[36] H. He, H. Lu, S. Bao, F. Wang, H. Wu, Z. Niu, and H. Wang, "Learning to select external knowledge with multi-scale negative sampling," *arXiv preprint arXiv:2102.02096*, 2021.

[37] Y. Cui, W. Che, T. Liu, B. Qin, Z. Yang, S. Wang, and G. Hu, "Pre-training with whole word masking for chinese bert," *arXiv preprint arXiv:1906.08101*, 2019.

[38] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "Electra: Pre-training text encoders as discriminators rather than generators," *arXiv preprint arXiv:2003.10555*, 2020.

[39] Y. Cui, W. Che, T. Liu, B. Qin, S. Wang, and G. Hu, "Revisiting pre-trained models for chinese natural language processing," *arXiv preprint arXiv:2004.13922*, 2020.

[40] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.