自然语言处理综述

【摘要】 本文主要介绍了当前自然语言处理在国内外的前沿技术。自然语言处理涉及深度学习，知识图谱等技术，本文主要介绍深度学习方法中最具代表性的几个方法。当下最火热的技术要属BERT，当前的研究有不少是基于BERT的变种，另一种存在更久的流派是RNN系列方法,其中以LSTM为代表。本文将以embedding，预训练任务，深度学习模型三个方面的发展展开。

【关键词】 自然语言处理；embedding；预训练任务；自然语言处理模型

Survey of NLP

【**Abstract**】 This paper mainly introduces the current frontier technology of natural language processing at home and abroad. Natural language processing involves deep learning, knowledge graphs and other technologies. This article mainly introduces the most representative methods in deep learning methods. The hottest technology at the moment is BERT. Many current researches are based on BERT variants. Another genre that has existed for a long time is the RNN series of methods, which is represented by LSTM. This article will focus on the development of embedding, pre-training tasks, and deep learning models.

【**Key Words**】 NLP; embedding; pre-train; NLP model

第 1 章 自然语言处理的发展进程

1.1 基于词级表示的 word2vec

Word2vec 模型采用 CBOW 和 SG 算法，首先将所有词用 one-hot 编码，然后将目标词汇在文本中的前后若干个词一起送入由高维向低维转化的全连接层，将一个词的 one-hot 表示转化成由其上下文表示维度大幅降低的词向量表示，而后用 SG 算法预测下一个词。与之前的算法相比，该模型的贡献在于引入了对词语上下文的考虑，简化了隐藏层的复杂程度，大幅的降低了 one-hot 带来的稀疏表示的问题，实现了词向量的低维稠密表示，在蕴含更多信息的同时减少了计算开销。但其不足在于无法克服一词多义带来的问题，且准确率远不如后来出现的模型，在文章整体理解上效果非常差。

1.2 基于文档级表示的 RNN 系列

RNN 系列可以通过循环来增加网络对句子整体的把握，在其发展过程中注意到了记忆消失，精度不够等问题，这些都在其发展的过程中不断被克服，但循环有个致命的弊端就是无法并行计算，在这个 GPU 焕发活力的时代，RNN 就显得差强人意了，但在很长一段时间里，RNN 都是主流的模型，直到 BERT 的出现，一度刷新了各个竞赛的排名。

1.2.1 RNN

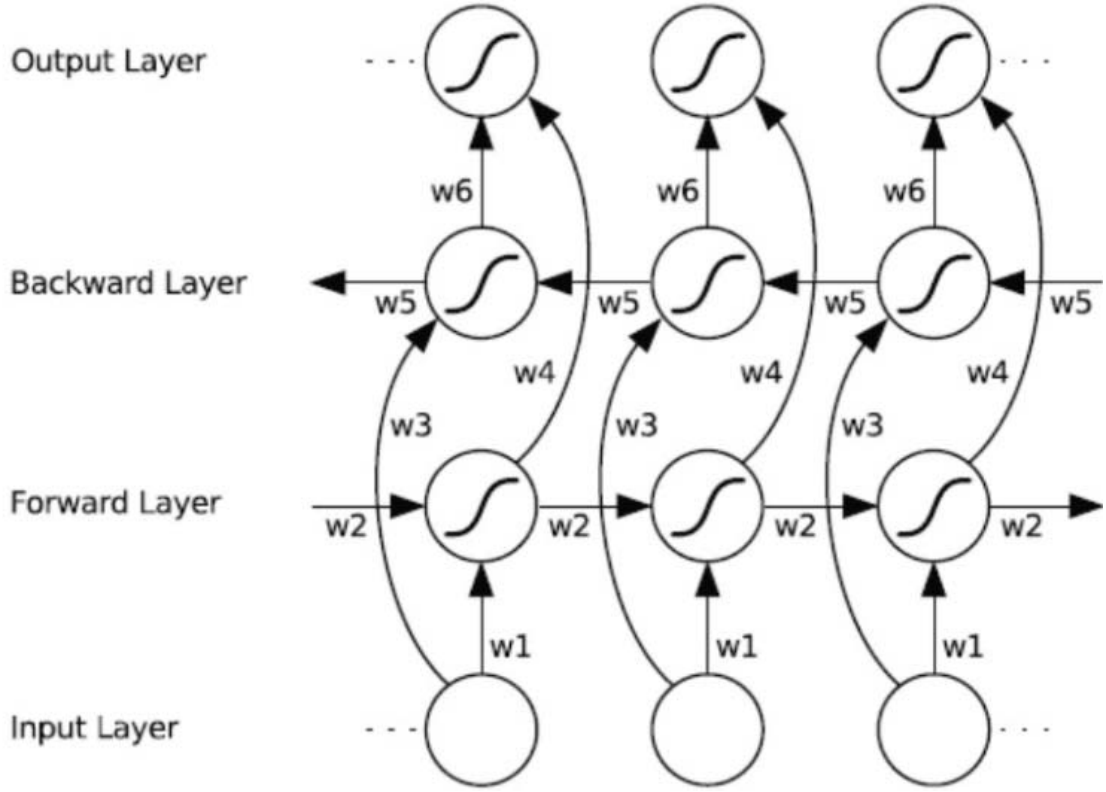
RNN 采用不断的将单词按序输入，来训练同一个模型，使得该模型能够通过上一个词预测下一个词，并为下一个词的预测提供依据。通过这种方式，似乎每一个单词的分类都是根据前序所有词为依据给出的，但实际上因为类似梯度消失的问题，随着递归次数的增强，前序词语逐个被遗忘，导致 RNN 的效果下降。

1.2.2 Deep RNN

Deep RNN 在 RNN 的基础上，对 RNN 单元进行了堆叠，即将前序 RNN 的输出作为后序 RNN 的输入，通过这种方式堆叠深度较深的模型，这种做法强化了 RNN 的拟合能力，有效的提高了精度，但仍存在 RNN 原有的所有缺陷。

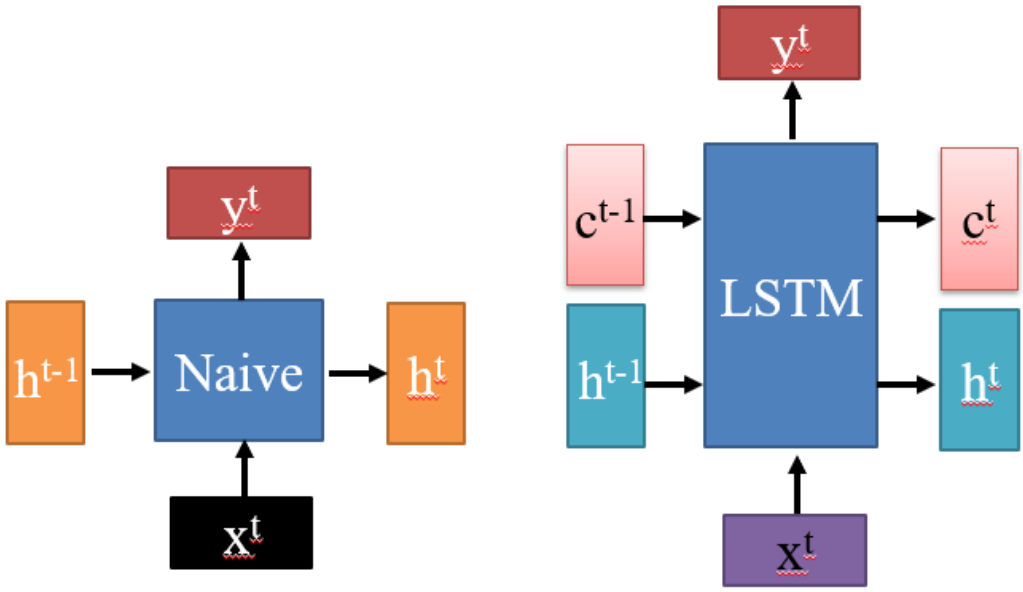
1.2.3 Bidirectional RNN

BRNN 主要解决之前 RNN 每一个词的预测都忽略了后续词句对单词理解的影响，他引入了两个 RNN 分别以正序和逆序的文本作为输入，将他们的计算共同作为预测的依据来预测结果。该模型的感受野更全面所以精度上也得到了提高。



1.2.4 LSTM

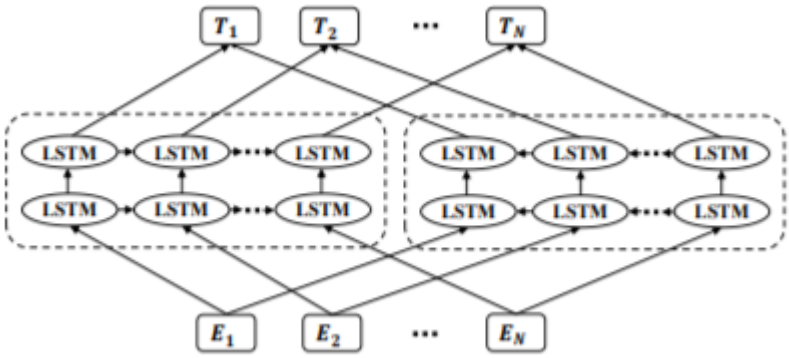
LSTM 的主要贡献在于引入了长期记忆这一概念，之前的 RNN 只输出一个代表短期记忆的向量供后续词语的预测，LSTM 则通过设置一个学习率较低的参数，代表长期记忆。在 LSTM 每一次循环则输出两个向量 c 和 h，分别代表长期记忆和短期记忆。



这有效解决了在多次递归中出现的遗忘问题，但是对后文的考虑就有所欠缺，这时 ELMO 就应运而生。

1.2.5 ELMO

ELMO 其实就是把 BRNN 和 LSTM 结合起来，形成双向 LSTM。但事实上其对于上下文的理解还是存在局限性。因为他是通过叠加正向和反向单向的结果得到的，这一分裂的行为导致其无法同时理解上下文的信息，使其对上下文的理解存在偏差。



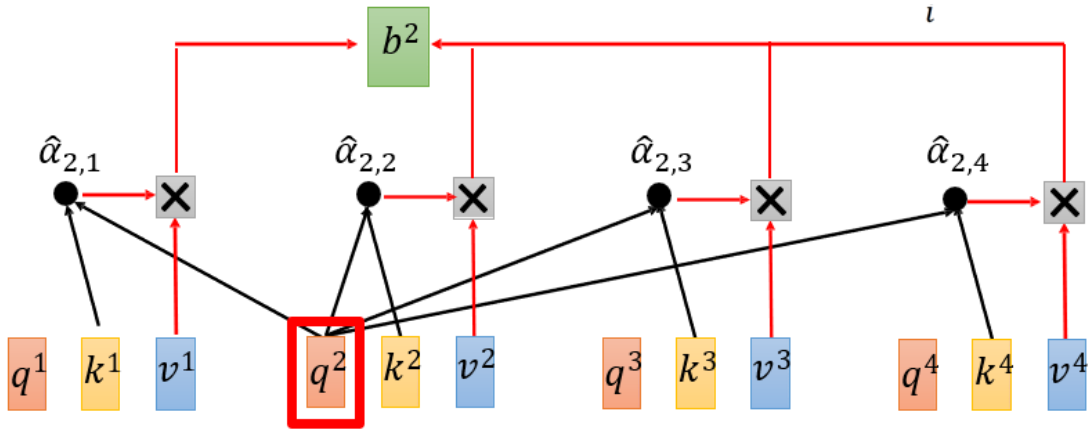
1.3 基于全局文本信息表示的模型

随着 Attention 机制的引入和计算资源的增加，深度学习模型能够从更高的维

度理解文本。

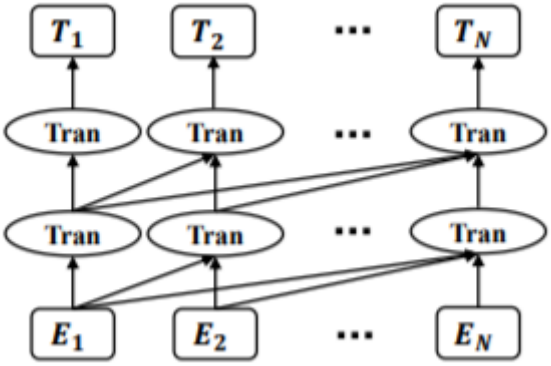
1.3.1 Transformer

Transformer 引入了 Self-attention 机制，即首先将一个单词转化为 q，k，v 表示，然后用目标词的 q 与其他所有词的 k 进行点积计算两者相关度，在利用这个相关度乘上每个单词对应的 v 之后进行叠加得到了对该词的预测结果。这种方式同步对上下文进行了参考了，解决了 LSTM 系列无法克服的无法并行和割裂看待上下文的问题，但同时也引入了大量的参数，需要更大的数据集和更多的计算资源驱动。为了增加 Transformer 的拟合能力，又引入了 Multi-head Self-attention，做多次的 self-attention。



1.3.2 GPT

GPT 则是在 Transformer 上进行了堆叠，但是在堆叠的过程中遇到了问题。不同于 LSTM，Transformer 是双向的，如果直接堆叠就会导致在第二层在做 self-attention 的时候使得单词“看到”了自己，这就使得模型提前知道了正确结果失去了堆叠的意义而当时又无法解决这一问题，其做法就是简单了砍去了 Self-attention 的一个方向，变成正序单向的，而后通过堆叠这一行为提高了拟合能力从而提高了精确度。



1.3.3 BERT

BERT 的出现是 NLP 领域的一个里程碑。BERT 在许多方面上都做出了改进。

首先是 embedding，BERT 同时对语义信息，句中位置信息和全局位置信息进行了编码，这使得 BERT 模型能够从多个维度分析问题，且可以适应各种各样的输入和任务。其次 BERT 克服了 GPT 无法实现双向的问题，其主要通过引入了 MLM 的预训练方法克服了词“看到”自己的问题，其次他还进行了 NSP 的预训练方法，使得他在文章层面有更好的效果。最后因为 BERT 是一个预训练的模型他并不面向任何特定的任务，这使得 BERT 具有非常好的迁移学习能力，对于不同的任务只需在 BERT 上进行 Fine-tune 就可以适应各种各样的任务，并且可以在 BERT 的基础上进行更强大的模型的开发。其缺点在于参数数量十分庞大达到 110~340M 个参数，需要大量的计算资源。

第 2 章 当前自然语言处理的发展

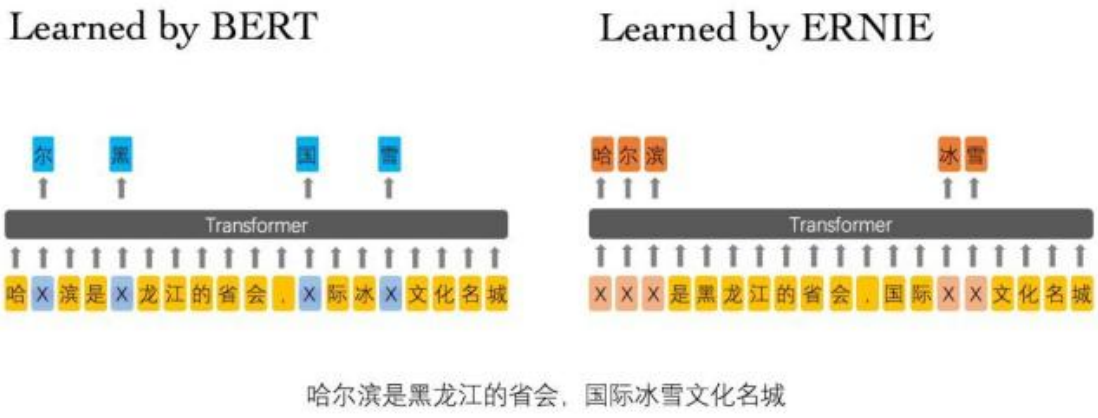
随着 BERT 带来了突破性的进展，近年来出现了大量对 BERT 进行改进的模型，其主要对 BERT 的 embedding，预训练任务，减少参数个数（包括 pre-train 和 fine-tune 等环节）进行优化。

2.1 Adaptor

Adaptor 主要简化了 Fine-tune 环节的开销，将 BERT 的参数锁定，在 BERT 的模型中穿插 Adapter 层，通过对 Adapter 层进行 Fine-tune，降低了 Fine-tune 的开销，且能够比之前更快的收敛。

2.2 ERNIE

ERNIE 主要贡献在于引入了 WWM 预训练任务。该篇论文提到，对于强关联的词组，mask 单个词是没有用的，因为可以仅通过前后俩个词就能准确的预测出这个词，不利于对其在整句话或者整篇文章的维度理解他，所以要对文章进行分词，将词语一组一组的 mask 掉来用于预训练。



2.3 BART

BART 也是提出了一种对预训练任务的强化。在 BART 中除了对文本进行了 mask，还引入了 delete，rotate，permutation，infilling 等操作，其目的在于让 BERT 能够适应更多的任务。但事实表明，rotate 和 permutation 的收效甚微，而 infilling 起到了很好的效果。

2.4 ALBERT

ALBERT 则采用循环的方式，即让原来 BERT 的 12 层共用同一组参数，有效的缩减参数的数量，而且并不会影响 BERT 的并行，因为 BERT 本来也是按层序进行的。

2.5 Transformer-XL

该模型通过缩减 self-attention 的复杂度来达到模型加速，缩减的主要是深层次的 self-attention，层次越深缩减越多。

2.6 replaced token dection

Clark 等人提出将 BERT 中 MLM 进行替换，采用了 replaced token dection 任务进 行预训练，经过测试发现，相较 BERT 性能有 2.9 个百分 点的提高，并且性能和 RoBRETa 相差不多但是训练所 需时间缩短了 75%。

第 3 章 自然语言处理的展望

3.1 多模态

VideoBERT 能够将视频和文本数据联合训练得到一个能够促进视频和文本互相理解的模型，使模型能够借助文本理解视频，通过视频理解文本，让模型更接近人的理解方式，让 AI 更加智能。

3.2 模型轻量化

目前的模型虽然都做了不少的精简但还是非常庞大，单机难以开展，研究人员也无法从头训练模型，只能在现有的庞大模型中进行迁移学习，使得整个领域的研究受到阻碍。

3.3 可解释性

无法解释是深度学习模型的通病，这也是知识图谱等传统方法仍有用武之地的原因。不可解释性使得模型安全成了一个很重要的话题，只有把模型做安全了甚至可解释了，那么深度学习模型才能用于更加严格的领域。

结论

NLP 领域非常适合深度学习，在该邻域内有大量的可用数据和应用场景。当前 NLP 领域模型的兼容性十分强大，单个模型可以做许许多多各种各样的任务，甚至可以用 BERT 做视频处理，其在自监督领域上的成功值得其他 AI 领域借鉴和推广，故以 NLP 领域为基础向多模态发展大有可为，将 AI 多个领域联合起来或许就能实现强人工智能。

参考文献

[1] Shaoqing Ren, Kaiming He, Ross Girshick. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, IEEE,2017.

[2] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[3] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.

[4] Graves A, Mohamed A, Hinton G. Speech recognition with deep recurrent neural networks[C], 2013 IEEE international conference on acoustics, speech and signal processing. Ieee, 2013: 6645-6649.

[5] Schuster M, Paliwal K K. Bidirectional recurrent neural networks[J]. IEEE transactions on Signal Processing, 1997, 45(11): 2673-2681.

[6] Jaderberg M, Simonyan K, Zisserman A, et al. Spatial transformer networks[J]. arXiv preprint arXiv:1506.02025, 2015.

[7] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training[J]. 2018.

[8] Sun Y, Wang S, Li Y, et al. Ernie: Enhanced representation through knowledge integration[J]. arXiv preprint arXiv:1904.09223, 2019.

[9] Yang Z, Dai Z, Yang Y, et al. Xlnet: Generalized autoregressive pretraining for language understanding[J]. arXiv preprint arXiv:1906.08237, 2019.

[10] Lan Z, Chen M, Goodman S, et al. Albert: A lite bert for self-supervised learning of language representations[J]. arXiv preprint arXiv:1909.11942, 2019.

[11] Dai Z, Yang Z, Yang Y, et al. Transformer-xl: Attentive language models beyond a fixed-length context[J]. arXiv preprint arXiv:1901.02860, 2019.