深度学习常用文本分类方法

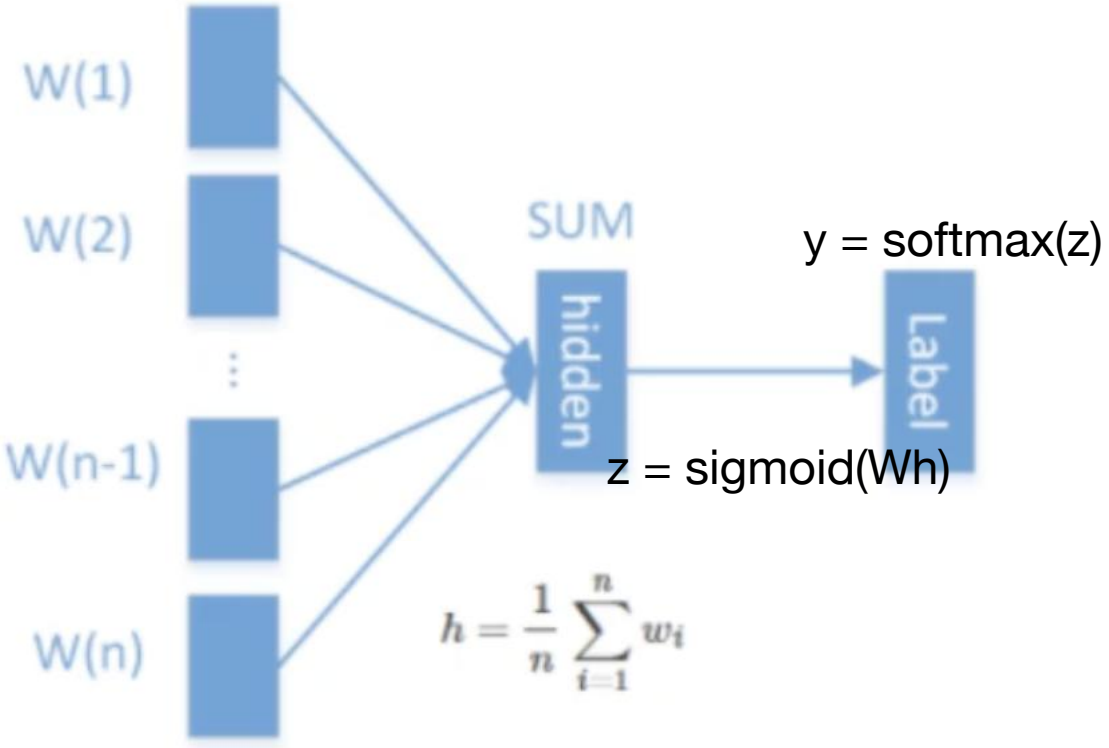
常用文本分类模型

- Fasttext（分类较为明显）
- TextCNN（短文本）
- DPCNN
- BiLSTM+Attention（通吃）
- HAN（长文本，对 BiLSTM+Attention 的改进）
- BERT
- Capsule
- TextGCN

开箱即用源码：https://github.com/649453932/Chinese-Text-Classification-Pytorch

Fasttext

- 通过 n-gram 分词；
- 一个简单的神经网络层；
- Fasttext 使用了基于霍夫曼编码树的分级 softmax，使训练的时间复杂度降为 O(hlog2(k))，使 Fasttext 速度快；
- fastText的核心思想是：将整篇文档的词及n-gram向量叠加平均得到文档向量，然后使用文档向量做 softmax 多分类得到其所属的类别 label。
- 源码：https://github.com/facebookresearch/fastText



Joulin A, Grave E, Bojanowski P, et al. Bag of tricks for efficient text classification[J]. arXiv preprint arXiv:1607.01759, 2016.

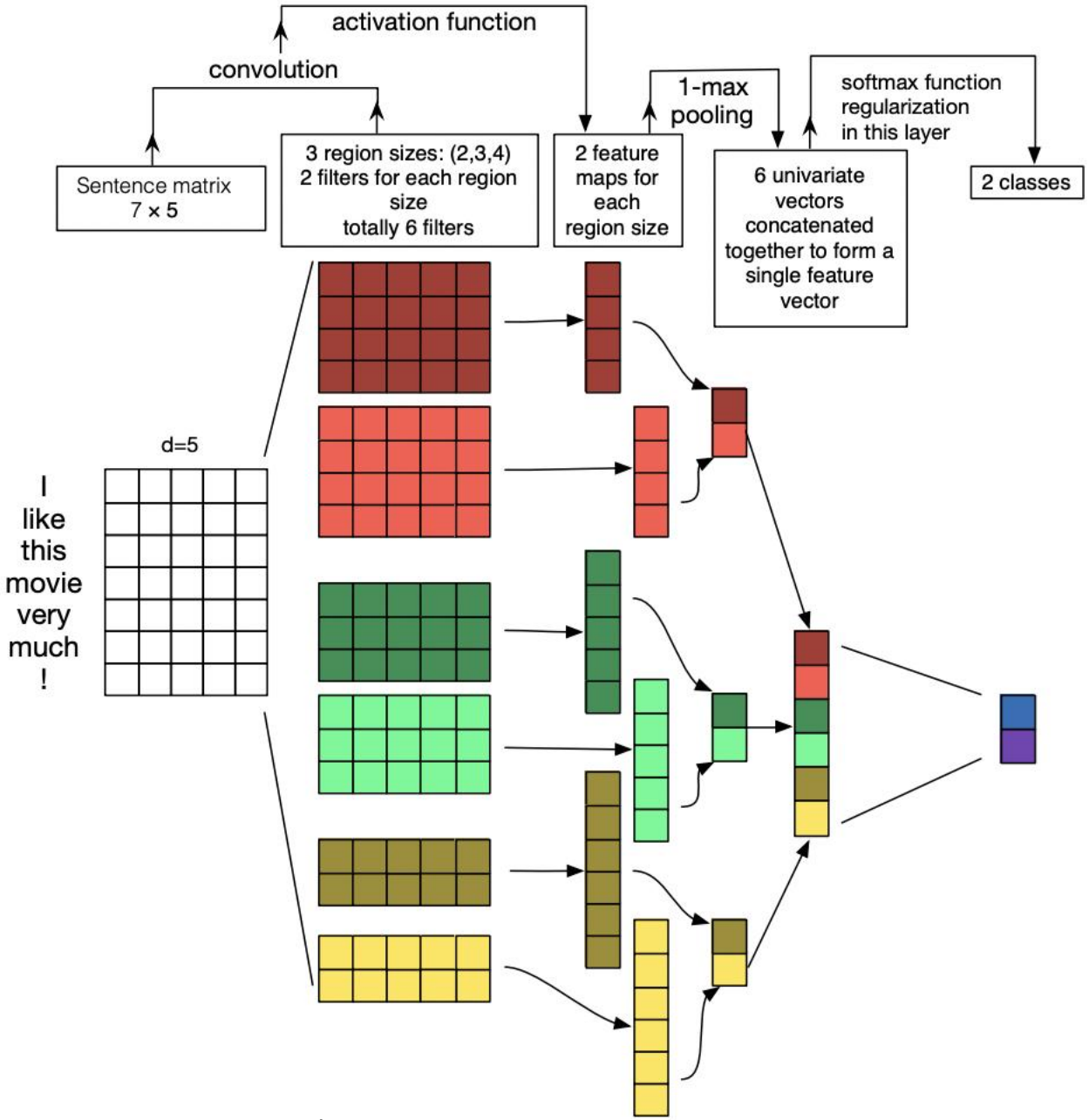
TextCNN

采用 CNN 应用于文本处理，可更好地捕捉局部相关性。

更适用于短文本分类

类似于提取N-Gram信息，
而且只有一层网络，难以捕捉长距离特征，不能通过卷积获得文本的长距离关系依赖。

参考 tf 的 pytorch 版实现：
https://github.com/XqFeng-Josie/TextCNN



Zhang Y, Wallace B. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification[J]. arXiv preprint arXiv:1510.03820, 2015.

DPCNN

- 论文中提出了一种基于word-level(以单词为语义单位)级别的深层金字塔卷积网(DPCNN),通过不断加深网络，可以抽取长距离的文本依赖关系。
- 实验证明在不增加太多计算成本的情况下，增加网络深度就可以获得最佳的准确率。

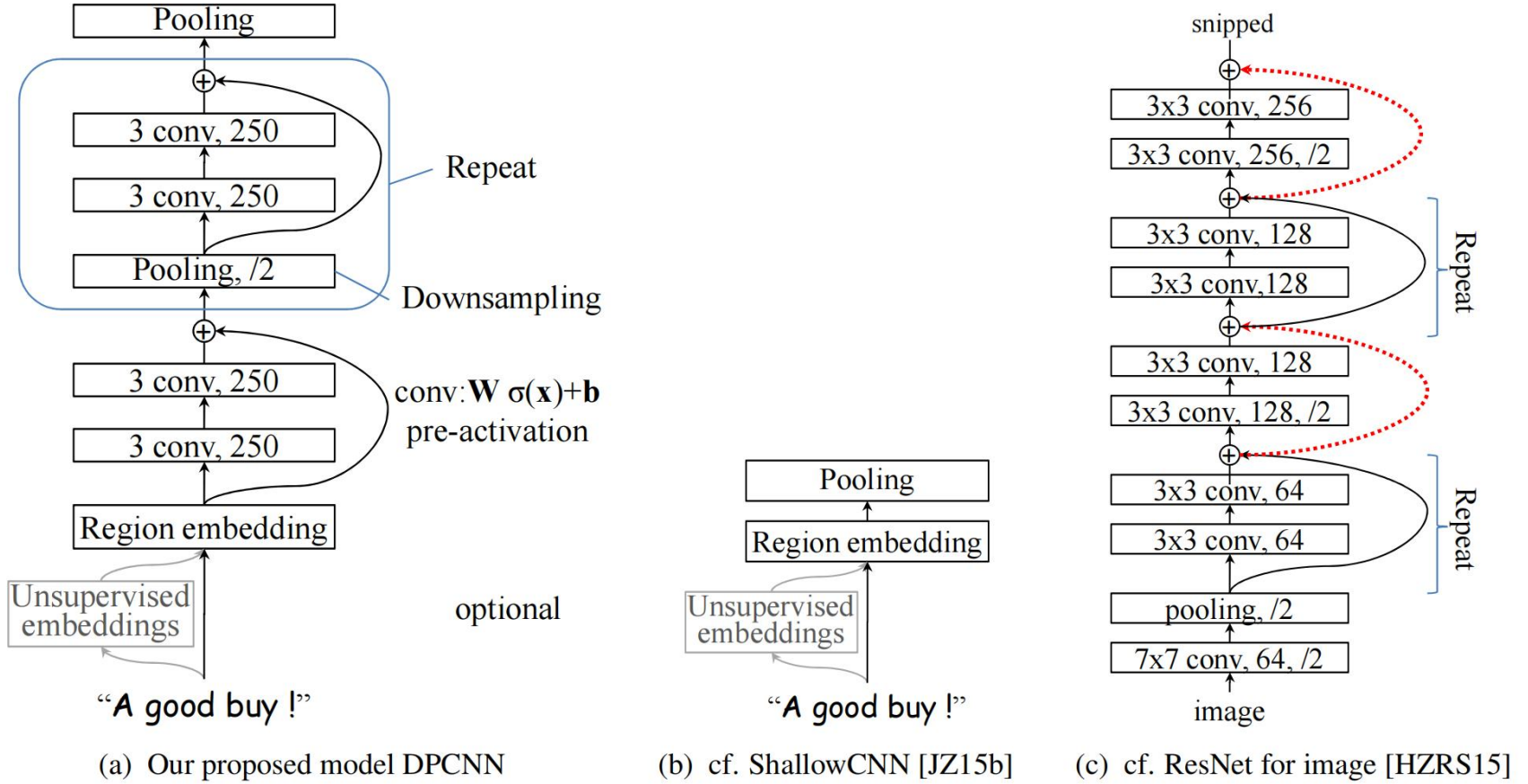
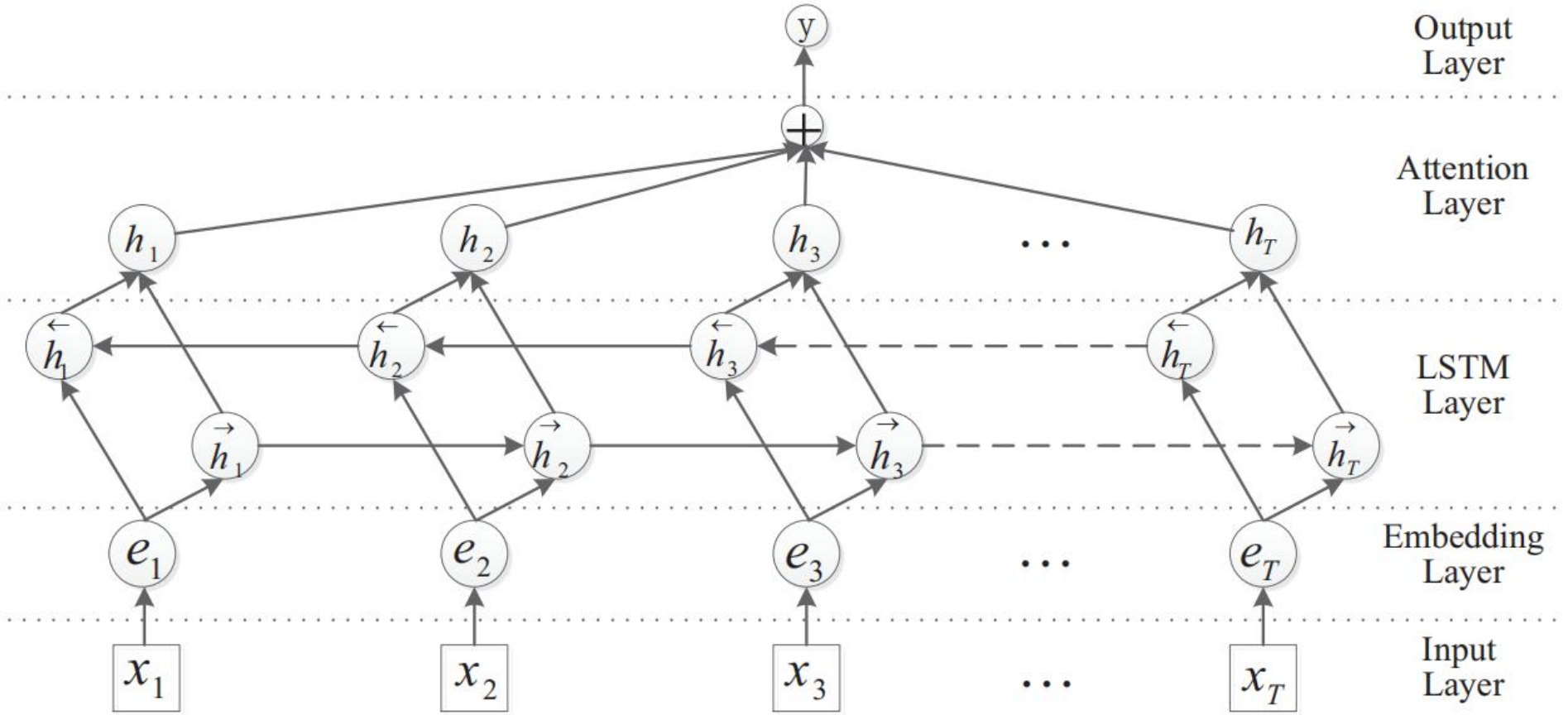


Figure 1: (a) Our proposed model DPCNN. (b,c) Previous models for comparison. ⊕ indicates addition. The dotted red shortcuts in (c) perform dimension matching. DPCNN is dimension-matching free.

Johnson R, Zhang T. Deep pyramid convolutional neural networks for text categorization[C] //Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017: 562-570.

BiLSTM+Attention



Zhou P, Shi W, Tian J, et al. Attention-based bidirectional long short-term memory networks for relation classification[C] //Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers). 2016: 207-212.

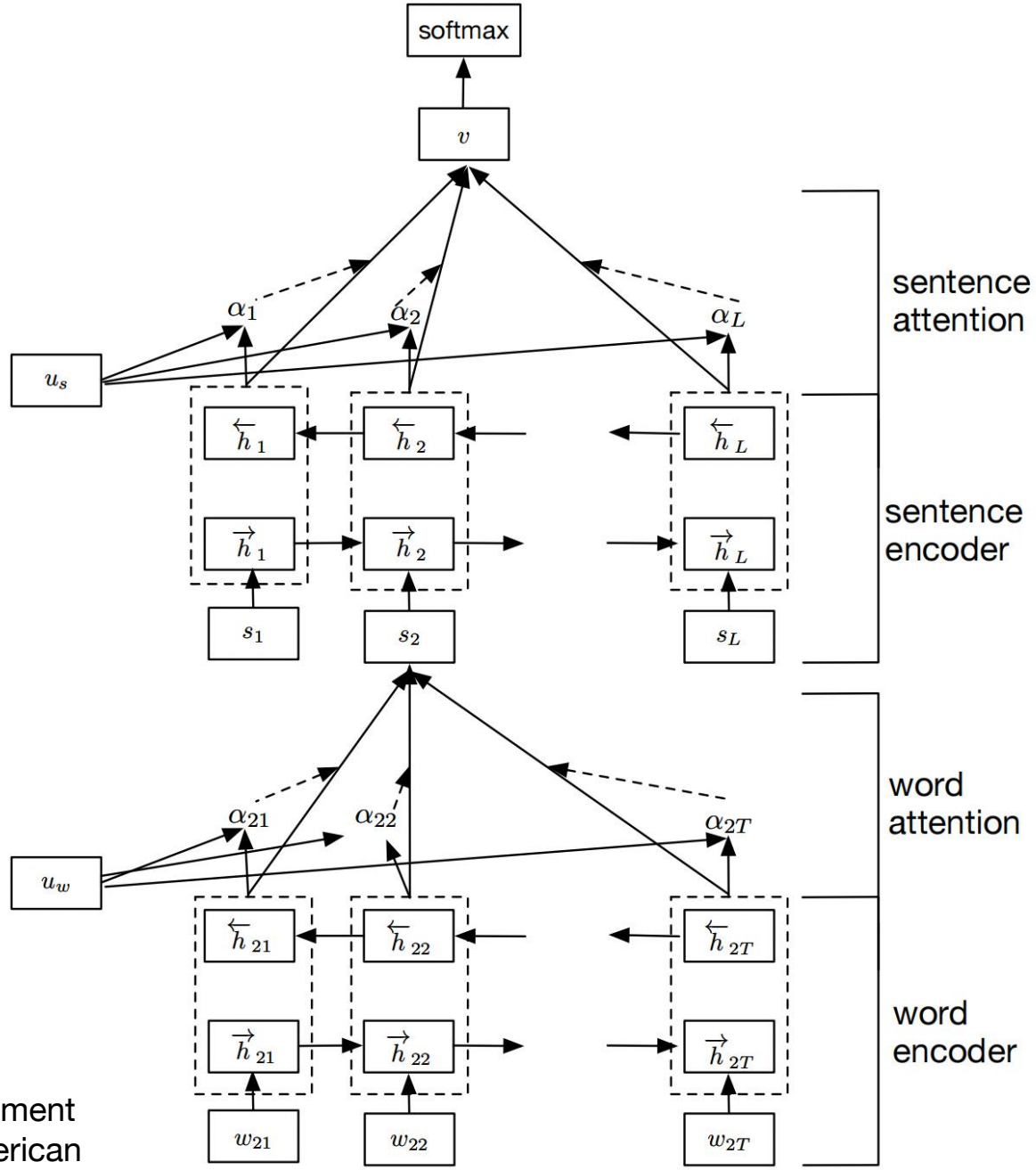
HAN-层次注意力网络

字编码器（word encoder）： 词向量嵌入+双向GRU；

单词注意机制（word attention）： 引入注意机制来提取对句子含义重要的词，并汇总那些信息词的表示以形成句子向量；

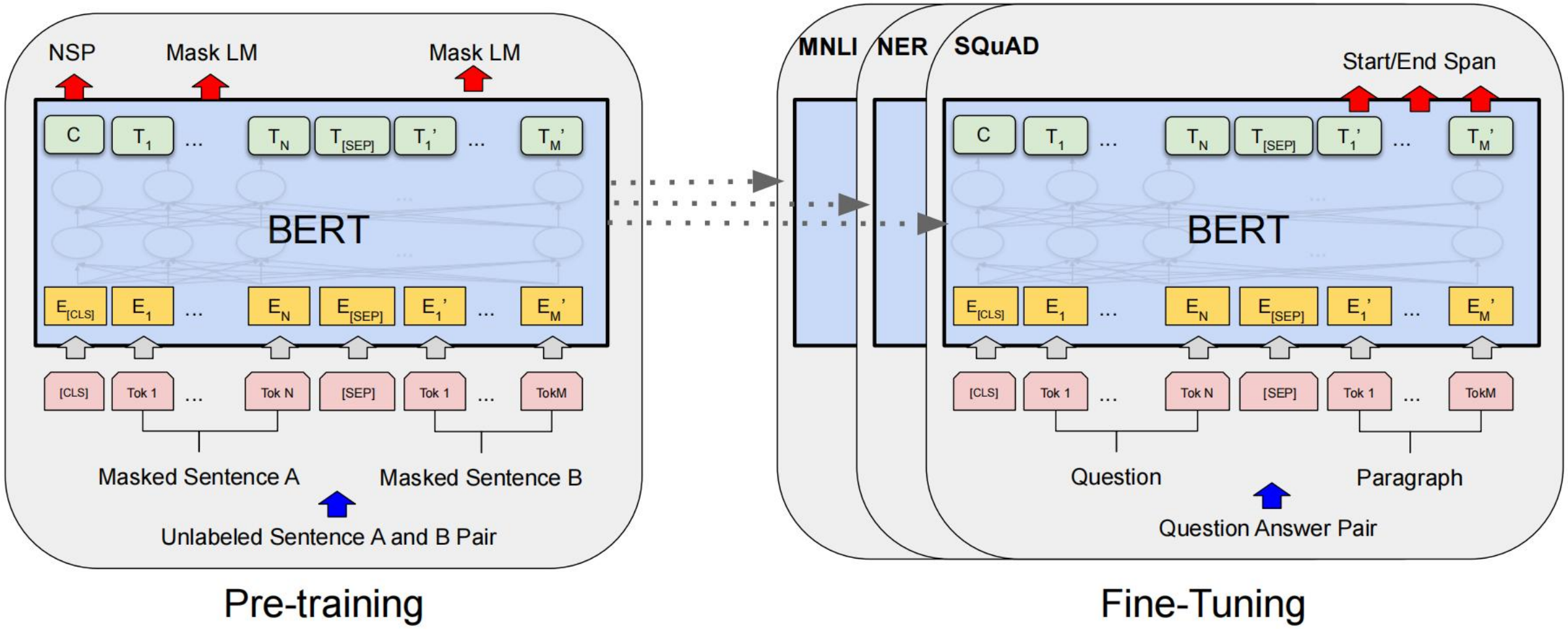
句子编码器（sentence encoder）： 类似于字编码器，获得文档向量；

句子注意机制（sentence attention）： 使用注意机制并引入句子级别的上下文向量，并使用向量来衡量句子的重要性。



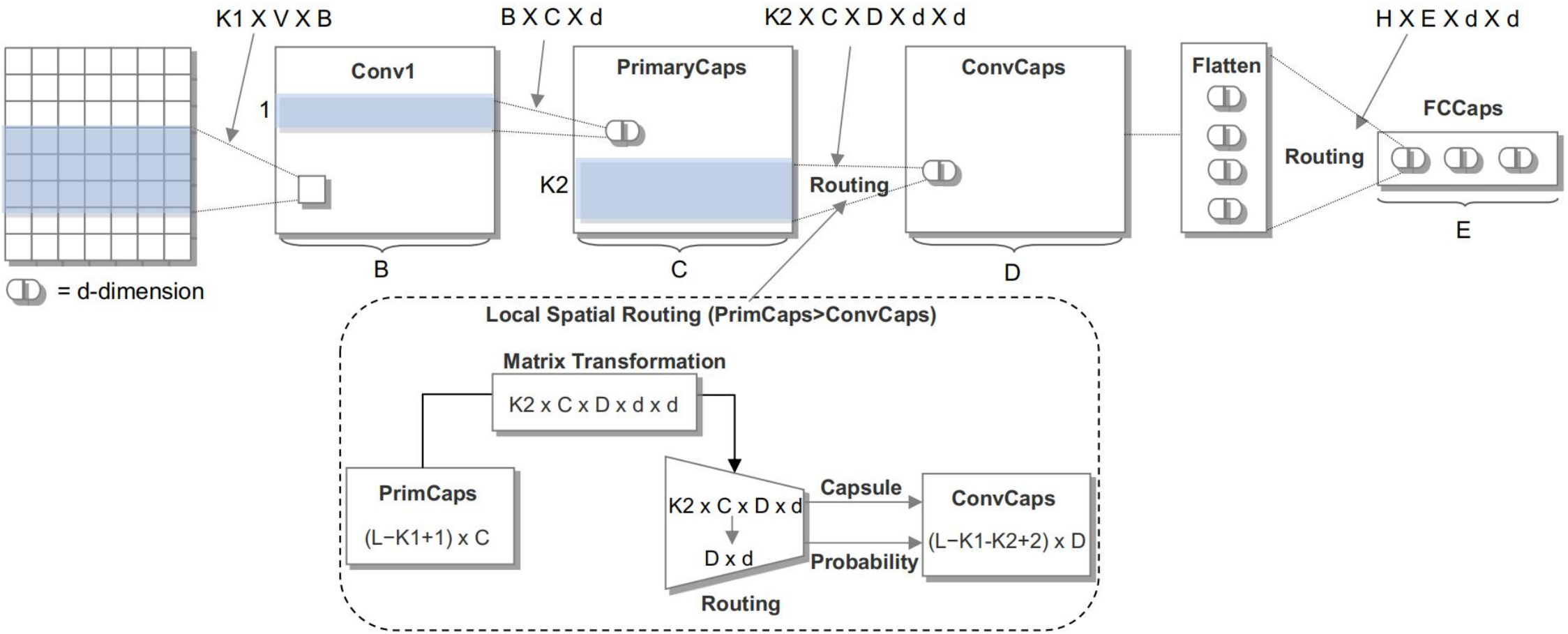
Yang Z, Yang D, Dyer C, et al. Hierarchical attention networks for document classification[C]//Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies. 2016: 1480-1489.

BERT-预训练语言模型



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.

Capsule Network（胶囊网络）



Zhao W, Ye J, Yang M, et al. Investigating capsule networks with dynamic routing for text classification[J]. arXiv preprint arXiv:1804.00538, 2018.

TextGCN-图卷积网络模型

- 主要思路为基于词语共现以及文本单词之间的关系构建语料库中文本的Graph，然后将GCN学习文本的表示用于文本分类；
- 通过多个基准数据集实验表明，Text-GCN无需额外的单词嵌入或者先验知识就能够取得优于最新的文本分类方法；
- Text-GCN还能够学习和预测词语与文档的嵌入表示；

图神经网络介绍：https://distill.pub/2021/gnn-intro/

Yao L, Mao C, Luo Y. Graph convolutional networks for text classification[C]
//Proceedings of the AAAI conference on artificial intelligence. 2019, 33(01): 7370-7377.

code: https://github.com/yao8839836/text_gcn

趋势

1. 现有模型的改进
2. 预训练语言模型的优化，得到更强大的语言模型
3. 融合迁移学习
4. 基于Transformer架构
5. 融合对抗学习
6. 结合可解释性分析