迁移学习在自然语言处理领域的应用

原创：小左 CodeInHand 2018-10-31

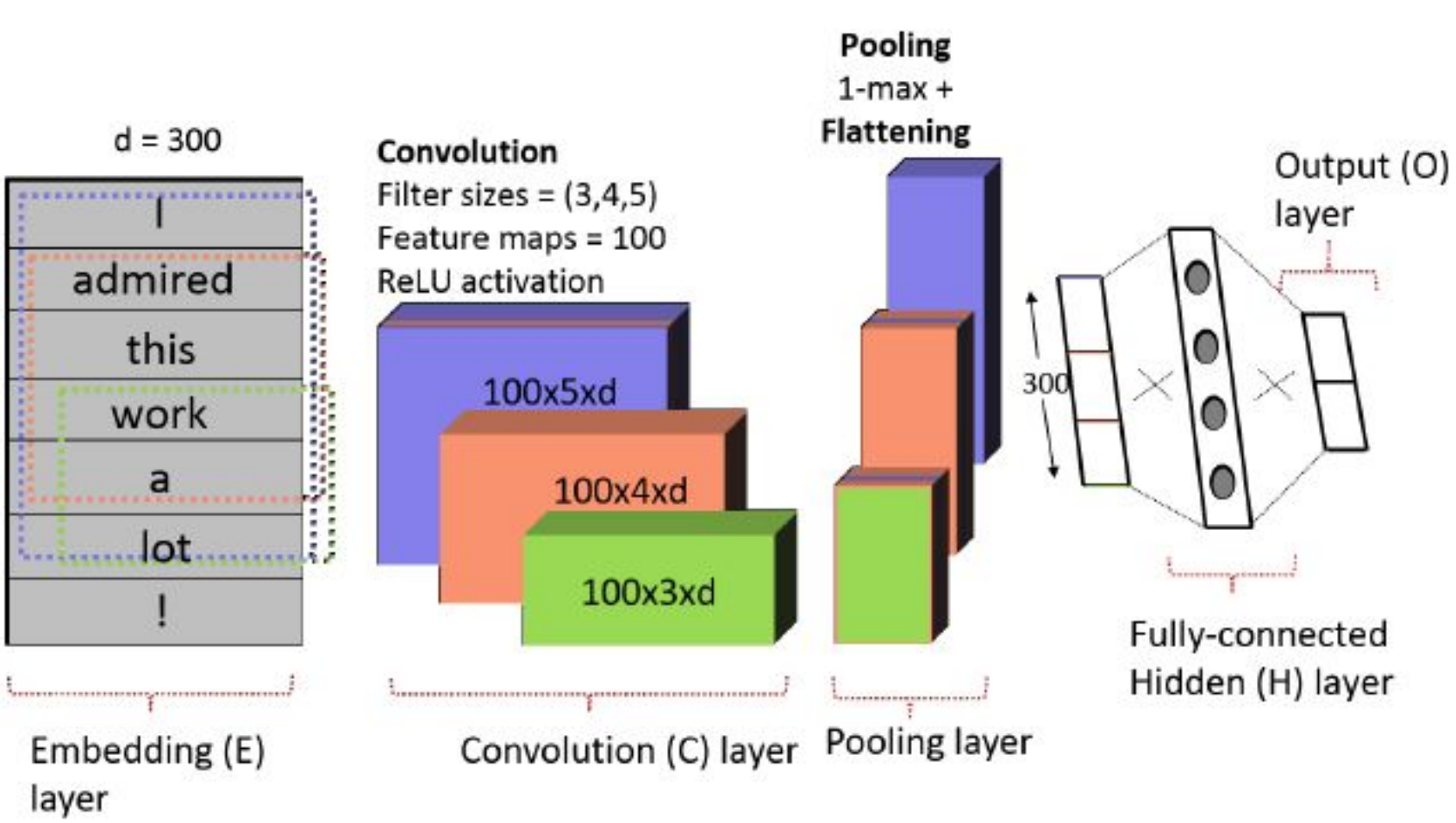
迁移学习

迁移学习近年来在图形领域中得到了快速的发展，主要在于某些特定的领域不具备足够的数据，不能让深度模型学习的很好，需要从其它领域训练好的模型迁移过来，再使用该模型进行微调，使得该模型能很好地拟合少量数据的同时又具备较好的泛化能力（不过拟合）。

在迁移学习任务中，需要事先定义一组源数据集合，使用该集合训练得到预训练好的模型，该模型具备了源数据集合中的一些知识，在目标数据集合上微调该预训练的模型，使得模型能够很好地完成目标数据集合定义的任务，即完成了迁移学习。

由于深度学习模型结构复杂，在NLP领域中迁移学习不够成熟，不知道如何进行迁移、迁移模型的哪个结构部分、源数据集合与目标数据集合之间需要满足怎样的关系。本文以CNN文本分类任务为例进行描述，总结一下迁移学习在NLP领域文本分类任务中的一些经验。

CNN文本分类模型框架



如上图为CNN文本分类的模型结构图，总的模型结构来说可以分为四层：Embedding层、卷积层（含池化层）、全连接隐层、输出层。Embedding层主要将词语映射为词向量表示、卷积层主要对词语矩阵进行卷积操作得到句子的抽象表示、全连接隐层一般是进行维度压缩、输出层是进行分类（对应类别的数量）。

在文本分类任务中的迁移学习，例如源数据集合为新闻文本的分类（数据量大），目标数据集合为

短视频标题分类（标注的数据少），通过预先训练的新闻分类模型，在短视频标题分类任务上进行模型（Embedding层、卷积层、全连接隐层、输出层）的微调，使得模型既能完成对少量有监督数据的拟合，又具备相应的泛化能力。下边将针对CNN文本分类任务进行经验总结。

经验与建议

经验

（1）目标数据集合与源数据集合在语义上太相似，反而会影响迁移学习的效果，部分相似效果最好；

（2）源数据集合的词典大小越大、OOV比例越小，迁移效果越好；

（3）对于Embedding层的迁移，无论是固定不变、还是微调效果都挺好；

（4）对于卷积层和隐层，若模型参数固定不变，很难提高迁移学习的效果，除非目标数据集合与源数据集合语义上非常相似、很少的OOV、具备很大的词典；

（5）输出层参数的迁移效果很差；

（6）源数据集合上训练的模型最好不加非线性激活函数，目标数据集上再添加

（7）dropout rate设置在0.5-0.7之间比较好

建议

（1）选择源数据集合时，尽量保证数据量大、OOV少、词典大，语义上与目标数据集合部分相似就行（不要太像）；

（2）最好迁移Embedding层；

（3）如果考虑迁移卷积层和隐层，尽量考虑微调，不要使用固定参数。

（4）如果分类类别数量不相同，尽可能不要迁移隐层；

（5）不要试图迁移输出层，除非是在线学习，使用少量数据进行微调（源数据与目标数据基本一致）

参考文献

[1] Semwal T, Mathur G, Yenigalla P, et al. A Practitioners' Guide to Transfer Learning for Text Classification using Convolutional Neural Networks[J]. 2018.