**文本分类国内外发展现状**

文本分类是自然语言处理的基本任务之一。目标是为文本分配标签。它具有广泛的应用，包括主题标签，情感分类和垃圾邮件检测，情感分析，问答应答或对话管理等。如何捕获不同文本单元的特征，例如短语，句子和文档是文本分类的挑战。

针对文本分类，研究者已经提出了各种类型的模型。最近的方法使用了深度学习，如卷积神经网络（Blunsom等人，2014）和循环神经网络基于长期短期内存（LSTM）（Hochreiter和Schmidhuber，1997）来学习文本表示。基于深度学习的神经网络模型已经对文本分类任务取得了巨大的进步。这些模型通常包括将文本的单词映射到向量的投影层。然后将矢量与不同的神经网络组合，以形成一个固定长度的表示。根据结构，它们可以分为四类：递归神经网络（RecNN1），RNN，CNN和其他神经网络。

**递归神经网络：**RecNN由递归树结构定义。在递归模型的类型中，来自树的叶节点及其内部节点的信息以自下而上的方式组合。引入了递归神经张量网络，通过基于解析树组合邻域成分来构建短语和句子的表示。 Irsoy和Cardie（2014）提出了深度递归神经网络，它是通过将多个递归层叠加在一起而构建的。

**RNN：**RNN开发了目标相关的长期短期内存网络（LSTM（Hochreiter和Schmidhuber，1997）），其中目标信息被自动考虑。Tai 等人（2015）将LSTM推广到Tree-LSTM，其中每个LSTM单元从其子单元获得信息。Zhou 等人（2016）介绍了BLSTM与注意机制，以自动选择对分类具有决定性影响的功能。Yang 等人（2016）引入了具有两个关注机制的层次网络，即文字分类和文字注意。

RNN有着复现结构优势，非常适合于处理可变长度文本。RNN可以通过首先将包括每个文本的令牌转换为形成矩阵的向量，从而来利用词的分布式表示。 该矩阵包括两个维度：时间步长维度和特征向量维度，并且将在学习特征表示的过程中更新。然后，RNN使用1D 最大池操作或基于注意的操作，其提取最大值或在矩阵的时间步长维度上生成加权表示，以获得固定长度向量。

**卷积神经网络：**CNN（LeCun等人，1998）是具有2D卷积层和2D池层的前馈神经网络，最初是为图像处理而开发的。然后CNN应用于NLP任务，如句子分类（Kalchbrenner等，2014; Kim，2014）和关系分类（Zeng等，2014）。区别在于NLP任务中的公共CNN由1D卷积层和1D池层构成。 Kim（2014）定义了一个具有两个通道的CNN架构。 Kalchbrenner 等人（2014）提出了一种用于句子建模的动态k-max合并机制。（Zhang和Wallace，2015）进行了一层CNN的敏感性分析，以探讨架构组件对模型性能的影响。Yin和Sch（2016）介绍了多通道嵌入和无监督预训练以提高分类精度。（Zhang和Wallace，2015）进行了一层CNN的敏感性分析，以探讨架构组件对模型性能的影响。

CNN利用1D卷积来执行特征映射，并且在时间步长上应用1D最大合并操作以获得固定长度输出。但是，由于RNN通过词读取句子词，可以有效地将矩阵视为“图像”。与NLP不同的是，图像处理任务中的CNN应用2D卷积和2D池操作来获得输入的表示。利用2D卷积和2D池来在时间步长维和特征向量维上对更有意义的特征进行文本分类是一个很好的选择。

**其他神经网络：**在现有的CNN和RNN模型的基础上进行改进，如Iyyer等人 （2015）引入了一个深度平均网络，在分类之前通过多个隐藏层提供了一个未加权平均的字嵌入。 Zhou等人（2015）使用CNN提取一系列更高级的短语表示，然后将这些表示送入LSTM以获得句子表示。