# MGNC-CNN：A Simple Approach to Exploiting Multiple Word Embeddings for Sentence Classification

**一、文章提出的问题：**

传统的分类模型仅仅使用一种词向量，但是不同的词向量模型着重语言的不同方面，我们并不知道哪种词向量模型适用于哪些场景。因此，本文提出一种新的、简单的CNN架构——MGNC CNN，这个架构可以利用多种词向量来进行句子的分类。MGCN CNN独立地从输入的向量集中提取特征然后在网络的倒数第二层将这些特征融合为一个最终的特征向量。然后采用组正则化策略，对从相应的词向量集合进行单独的正则化。

【现有：只考虑一种词向量模型；本文改进：考虑多种词向量模型，进行综合判断】

**二、背景介绍及现状综述：**

近年来神经网络在自然语言处理领域越来越多的获得关注。对于句子分类的任务，CNN已经取得了显著的效果（Kim,2014; Zhang,2015）。这些模型都是基于词向量的，词向量作为一种高密度、低维度的表示方法，它能够提取明显的语义和句法属性信息。因此，研究词向量的规范表示是一项重要工作。已经存在的观点有，（1）Kalchbrenner等人(2014)初始化词向量为随机的低维度向量，然后在训练中被优化。（2）Johnson等人(2014)对每个词使用固定的一次性编码。（3）Kim（2014）等人通过word2vec模型初始化词向量。word2vec是一个被谷歌训练了1000亿词的模型，且该模型被优化至今。这种分布式的词向量初始化使得传送学习到的分布语义成为可能。这也使得相对简单的CNN模型能够取得显著的效果。

现在我们可以很方便的在网络上找到许多已经训练好的词向量模型，然而不同的词向量表示方式可能注重语言的不同方面，比如基于词袋的模型（BOW）着重获取词语间的关联性（医生与医院），基于依赖的模型着重获取词语间的相似度（医生与外科医生）。因此，我们需要考虑怎样组合这些不同的词向量模型以改进NLP模型。

【现有多种词向量模型，每种词向量模型的侧重点不同】

**三、相关工作：**

现有的工作已经考虑结合语法和语义信息进行词向量的表示（VandeCruysetal，2011），以及针对一般的NLP任务引入多种模式的词向量组合Brunietal,2012）。最近，（1）Luo等人（2014）提出了一个框架，结合多个词向量来测量文本相似性，不过该项目重点不在于分类问题。（2）Yin等人（2015）提出了一个句子分类模型——MVCNN，这个基于CNN的模型能接收多个词向量的输入，每个词向量都是单独的通道，类似于图像中的RGB通道。过滤器同时作用在每个通道。MVCNN模型在多个句子分类任务中达到最先进的性能，但还是存在一些缺陷：比如，MVCNN要求输入的词向量是相同维度的；其次，该模型在实现和运行上都比较复杂，且训练时间较长。

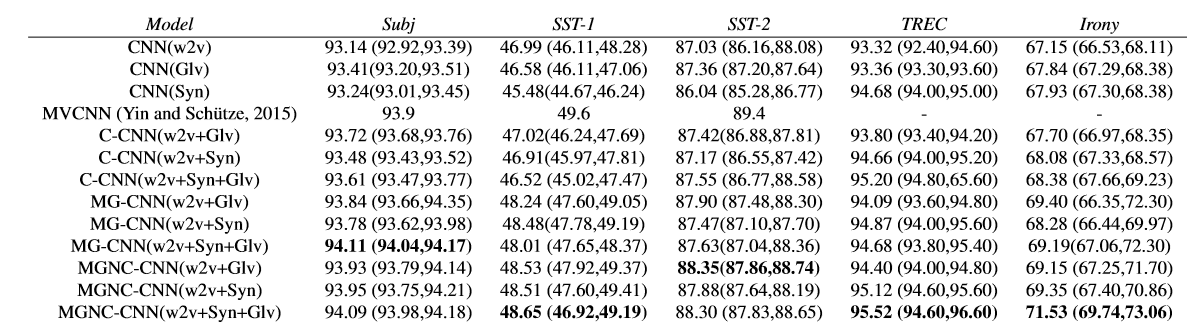
【Luo等人：考虑了多个词向量的文本相似度模型，但重点不在文本分类任务上；MVCNN：考虑了多种词向量的文本分类模型，但是有上述缺陷】

**四、本文模型：**

**MG-CNN**：假设有m种词向量，每种词向量对应的维度分别为d1,d2…dm，模型使用不同的过滤器组{w1},{w2}…{wm}独立处理每种词向量，和传统的CNN一样，对于每个{wl}，都可能有不同尺寸和个数的过滤器。然后在分类层，将过滤器通过每种词向量集合得到的特征向量进行简单拼接形成最终特征向量，最后输入softmax层。这种模型可以综合不同词向量的特征，因此被称为多组CNN（multiple group CNN）

**MGNC-CNN:**本文提出了一种MG-CNN的增强模型——MGNC-CNN（multiple group norm constraint CNN）,这个模型主要是对每种词向量模型单独进行正则化。具体来说，增加了分组正则化约束项，能独立地正则化每类词向量的参数，这些参数在训练过程中被自动优化。可见，该方法旨在更好地综合多种词向量的信息，利用词向量的特征。

**五、实验：**



* 对于任务一（二分类，客观的、主观的）：

MG-CNN（word2vec、GloVe、Syntactic）有着最优的效果。

* 对于任务二（五分类，非常消极、消极、中立、积极、非常积极）：

MGNC-CNN（word2vec、GloVe、Syntactic）有着最优的效果。

* 对于任务三（二分类，消极、积极）：

MGNC-CNN（word2vec、GloVe）有着最优的效果。

* 对于任务四（问题的六分类，缩写、实体、描述、人、位置、数字）：

MGNC-CNN（word2vec、GloVe、Syntactic）有着最优的效果。

* 对于任务五（二分类，讽刺的、不讽刺的）：  
  MGNC-CNN（word2vec、GloVe、Syntactic）有着最优的效果。

**六、论文结论：**

本文提出了一种新的、简单的、拓展性强的模型——MGNC CNN。该模型可以适应多种可变大小的词向量输入。本文的模型将不同类型的词向量看作不同的组，在每个组上单独使用CNN，从而产生相应的多个特征向量，然后再分类层级进行多个特征的连接。受到先前对NLP任务的规范化定义工作的启发，对不同的词向量集中产生的特征施加不同的权重。

本文的模型和现有模型相比，有如下优势：（1）更灵活，适应于不同维度的词向量输入；（2）更简单，不需要互相训练和预训练的过程；（3）训练时间更快。且通过实验可知，本文提出的MGNC-CNN和MG-CNN不管在什么分类任务下都是优于基准的（CNN），并且MGNC-CNN大多数情况好于MG-CNN

**七、小编总结：**

1. 智能客服项目可以根据该论文观点优化模型， 即使用多套词向量模型，并自动学习优化各个词向量模型的权重。
2. 自然科学基金申请是否可以参照论文的写法，进行不同CNN模型、不同词向量模型的性能对比。