# 课题组机器学习工作基础

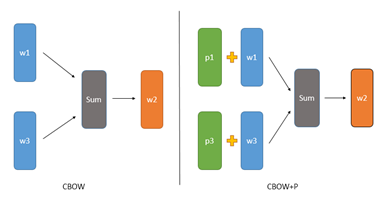
## 金鑫——语音强化的词向量模型

## 背景

用神经网络构建词表示，并借鉴语言学上关于语言符号的两个载体，即音载体和形载体，在表义上的关系的相关论述，将词的音信息也加入到词向量的建模中。我们以汉语、英语和西班牙语为例，在word2vec的基础上对词表示进行修正，使修正的词表示既包含词的形信息也包含词的音信息；与此同时，我们也以汉语为例，在CWE（character enhanced word embedding）模型的基础上对词表示进行修正。实验结果表明，在加入音信息后的词向量，不论是在哪个语言下，在词相似性和文本分类任务中都有提高。因此，我们提出的模型适用大部分语言，也能和很多现有的词向量模型相结合。

## 方法

从文字与发音的关系来说，汉语、英语和西班牙语是三种不同类型的语言。汉语中，文字与发音没有任何关联。在英语中，由于英语的复杂发音规则，我们只能从词的拼写中大概猜出词的发音。对于西班牙语来说，几乎每个字母的发音都是固定的，我们可以根据词直接得到发音。因此，本文以汉语、英语和西班牙语为例，在word2ec的基础上融入PWE思想。此外，本文也将以中文为例，在CWE（character enhanced word embedding）模型的基础上融入PWE思想。CWE模型考虑到一些语言的词是由更加基本的单位组成，如中文的词是由字组成，因此其将这些基本单位也融入到词向量的建模中。

1. 

Word2vec包含CBOW和Skip-gram，基于CBOW的PWE模型如上图所示，其利用词的发音信息对词向量进行修正。在模型中，我们首先对拼音或者音标向量化，得到音向量。修正的词向量就是将原先的词向量和音向量相加，使得修正的词向量既包含文字信息，也包含发音信息。

以word2vec为例，CWE模型也是在输入层对词表示进行修正，将词所包含的字向量也加入到原先的词向量中。因此，基于CWE的PWE模型是要将词的字信息和音信息一同加入到词向量中，使修正词向量既包含字信息，也包含音信息。

## 结论

我们以汉语、英语和西班牙语为例在词相似性和文本分类实验上对PWE模型进行了评估。汉语中，词相似性实验的数据集是wordsim-240和wordsim-296，文本分类的数据集是网易、搜狗和复旦新闻分类数据；英语中，我们用MTurk-771、MEN、WS-353-SIM和WS-353-REL为词相似性实验数据集，用20Newsgroups新闻分类语料库作为分类数据集；西班牙语中，我们用WS-353和RG-65作为词相似性实验数据，用TASS 2017作为文本分类数据，TASS 2017是源于推特的情感分类语料库。实验结果表明，不论是词相似性实验还是文本分类实验，基于PWE思想的模型普遍优于基准模型，说明加入音信息后，词向量确有改进。

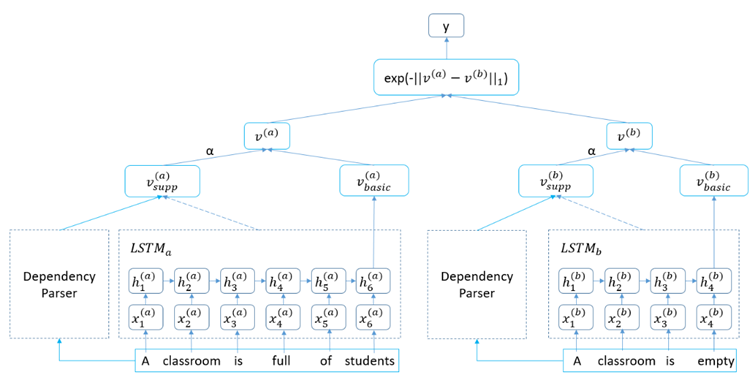
## 姚滕俊——基于依存句法分析的句子表示研究

（Dependency-based Siamese long short-term memory network for learning sentence representations）

## 背景

在自然语言处理领域中，语言表示是至关重要的。一个优秀的语言模型能够大大提升文本理解和信息抽取等自然语言处理任务的效率。目前短文本表示(如单词或者短语)工作已经取得了巨大进步，诸如word2vec等短文本建模方法已经相当成熟可靠，并且被广泛用于各种自然语言处理任务。然而，长文本(句子、段落或者篇章)表示的研究由于句子等长文本相对于单词和短语有着更为复杂的结构，仍存在巨大挑战。由于LSTM在处理序列上的优势，它经常被用于对句子的建模，但是我们认为标准的LSTM模型获取的句子表示对句子的各个组成部分都“一视同仁”，没有对句子的结构信息(主谓宾)给予足够的重视。而这些被忽略的结构信息可能对句子表示有着更加重要的影响。基于这样的思想，我们提出了基于依存句法分析的D-LSTM（Dependency-based Siamese long short-term memory）。

## 方法



该模型通过一个预训练好的依存句法分析器对句子进行依存关系分析，并结合原始句子输入得到句子的辅助表示v\_supp，然后通过LSTM得到句子的基本表示v\_basic，最终通过引入权重调节因子α来调整辅助表示和基本表示在句子表示中的重要程度，进而获得完整的句子表示。与传统的LSTM得到的句子表示相比，通过D-LSTM得到的句子表示能更加突出句子的主要语义。

## 结论

D-LSTM的性能评估在SICK（a sentence involving a compositional knowledge）数据集上进行，该数据集9927个句子对，其中5000对用于训练，4927对用于测试。每个句子对都有一个相似度评分以及一个类别标签（共有3类）。我们在2个自然语言处理任务中对模型进行了评估，分别是语义相关度评分（semantic relatedness scoring）和蕴含关系分类（entailment classification）。语义相关度评分任务的评价指标为皮尔逊相关性系数，蕴含关系分类的评价指标为分类准确率。实验表明，D-LSTM在上述两个自然语言处理任务中都比SICK公布的最优模型的性能好。s

## 徐永林——材料计算

## 背景

如何提高新材料研发设计的效率，成为材料科学工作者的首要目标。人们先后采用了一系列手段来筛选具有期望属性值的材料：传统实验合成试错法，基于计算模拟的高通量计算方法和基于历史数据的统计学习预测方法。

传统实验合成试错法得到的实验结果真实可靠，但是存在着成本高、周期长、需要借助多年的人工经验这些显而易见的缺点。在材料理论模型与计算材料科学的不断发展的条件下，人们开始采用基于密度泛函理论（DFT）的第一性原理方法、分子动力学方法、蒙特卡洛方法等计算模拟方法的高通量计算手段。

高通量计算方法作为材料计算科学的主要手段极大地提高了材料的发现与发展的效率，然而，典型的高通量计算任务设计会面临材料空间大、计算耗时长的问题，因此，如何提高高通量计算的效率成为材料研究人员越来越感兴趣的话题。

随着高通量计算技术的广泛应用，出现了大量的可获得计算材料或预测材料属性的数据库（如ICSD、NOMAD、AFLOWLIB、CSD等）.受数据驱动原理启发，机器学习作为一种well-established的统计学技术逐渐被引入到材料数据的挖掘中，获取有价值的信息，来减少材料高通量计算的任务量，并发现了一些新的材料。

## 方法

MIP材料研究团队在材料信息学方面做了深入研究，在机器学习方面，探索了类金刚石的性能、群速度，能带与材料原子属性、化合物属性方面的关联关系。

另一方面，针对具体高通量计算上，我们发现典型的材料高通量计算任务计算周期比较长，而现有的材料机器学习应用仅仅是简单的筛选出具有期望属性的化合物，减少执行完整高通量计算的条目数，无法真正提高单个高通量计算任务的效率。考虑到一般的材料高通量计算任务往往需要经历若干个计算步骤，如结构驰豫优化，静态自洽，能带计算，输运计算等。这些计算步骤明显的具有顺序性、阶段性、有针对性的的特点，且在计算过程中，会产生大量的计算属性信息，如bader、bandgap等材料属性，以及执行计算任务本身带来的信息，如计算时间，收敛次数等。因此可以将机器学习技术引入到高通量计算的整个过程中，对高通量计算任务本身进行优化与指导，则有望可以直接为单个高通量计算任务减负。

## 结论

目前，我们已实现了材料大数据查重算法，过滤掉重复的材料体系，并实现了单组元单组分的材料数据掺杂、置换的技术。在机器学习方面，我们实现了能带，类金刚石化合群速度相关属性的机器学习预测工具，以提高材料人员的研究效率。此外，我们设计了一种自评价式高通量计算框架—SEHC，该框架为高通量计算架构引入了self-evaluation机制，可使得在高通量的计算过程中，通过预判断设定的指标信息来提前停止不符合期望的材料计算任务，减少无效的计算量。我们在254个化合物空间中寻找具有较高群速度的热电材料时，设计了具有两个带预测单元的Stage的Pipeline模型，并实现了基于SEHC的高通量计算框架原型。经试验验证，该新型高通量计算框架相对于普通高通量计算系统，加入了自评估方法，如机器学习方法，可以优先筛选出符合期望的体系进行计算，此外，其在使用了Pipeline模型以后，较普通机器学习预测方法，可以提高11%的预测准确率。通过在已有pipeline中替换新设计的high-throughput stage可以快速的实现新的研究目的的高通量计算系统原型，提高研发了高通量计算任务设计效率。

## 胡冠男——基于k-最近邻训练的深度Q学习

(Deep Q-learning based on Neural K-nearest Neighbourhood Training)

## 背景

使用深度神经网络解决增强学习问题时，往往存在着这些问题：（1）模型需要通过激励信息学习agent在一段时间内的总激励，而通常激励信号具有稀疏、有噪声和延迟的特点，即采取行为和产生激励信号之间可能间隔多步，与传统意义上的监督学习不同；（2）深度学习的预设条件是样本数据是独立的，增强学习问题中的样本不独立，特别是相邻序列之间的状态是高度相关的；（3）在增强学习中，样本数据的分布随着采取动作的不同而发生变化，深度学习则假设样本的分布不变；为了缓解样本高度相关和分布变化带来的模型的不稳定的情况，引入经验回访机制，将所产生的样本数据放入memory中，随机选择一定数量的样本进行训练，在一定程度上缓解模型的不稳定性。但是深度学习应用的过程中学习速度慢的弊病：（1）应用于深度神经网络优化的随机梯度下降方法存在着优化速度慢的特点。随机梯度下降方法需要使用较小的学习率，使得模型收敛速度慢；（2）样本数据的高度不平衡，低激励的样本数据数量往往大大超过高激励的样本数据，使得模型很难学习；为解决上述问题，提出基于K-最近邻训练深度Q学习方法，提高样本数据的利用效率，提高模型的训练效率。

## 方法

Q-learning估计：, 成本函数,更新方法

基于K-最近邻训练深度Q学习方法，设置一个KND（K-nearest Neighbourhood Dict）,KND可以查询和写，在通过状态查询得到k-最近邻的激励集合和状态集合，则状态的一个激励，是的第i个最近邻在对应动作的激励，，是的第i个最近邻对应的状态,是向量和向量的和函数。如高斯核函数.

## 结论

将算法应用在Atari 2600的游戏上，结果表明方法相较于Prioritised Replay和Retrace()回放机制，回放机制KND将相似状态的样本数据合并，得到一个更加合理的激励，基于K-最近邻的回放机制具有更高的样本数据效率，提高了模型的训练效率。