KDD2019

近年来，许多在线学习系统，如Khan Academy和LeetCode，由于其方便性和自主性，在从K12到大学，甚至成人的各个年龄段的学习者中获得了越来越多的普及。在这些平台中，有必要提前组织好如此丰富的问题。例如，我们需要根据难度属性对它们进行排序，或者根据它们的知识概念创建课程设计。这样的管理是非常必要的，因为它们可以帮助学生省力地找到目标培训和有效学习所需的问题。因此，找到一种有效的系统理解试题的方法至关重要。事实上，由于它是促进许多基于问题的应用的基本问题，例如难度估计、知识映射和分数预测，因此系统创建者和研究人员都非常关注它。

理解学习材料（如试题）是在线学习系统中的一个关键问题，它可以促进教育领域的许多应用。不幸的是，许多有监督的方法都面临着人类标记数据稀缺的问题，而大量的未标记资源却没有得到充分利用。为了缓解这个问题，一个有效的解决方案是使用预先训练好的表示法来理解问题。然而，NLP地区现有的预训练方法由于教育领域的一些特殊性，不可能学习试题表征。原因如下：

首先，问题通常由异构数据组成，包括内容文本、图像和旁白信息。

第二，既有基本的语言信息，也有领域逻辑和知识。

本文目标：提出了一种新的预训练方法（Pre-training），即QuesNet，用于全面学习问题表征。

1. 设计一个统一的框架，将问题信息及其异构输入聚合为一个综合向量。
2. 提出了一种两级分层的预训练算法，以便在无监督的情况下更好地理解试题。两级：
	1. Holed-language model objective: 提取底层语言特征，
	2. Domain-oriented objective: 学习高层逻辑和知识。
3. 说明QuesNet在许多基于问题的任务中具有良好的微调能力。我们在大规模的真实问题数据上进行了广泛的实验，实验结果清楚地证明了QuesNet对于问题理解的有效性及其优越的适用性。

异质问题例子：



理解问题的传统方式：

1. 基于规则的表示，学者们致力于设计许多细粒度规则或语法，并通过将问题文本解析为语义树或预定义特征来学习理解问题。然而，这些计划在很大程度上依赖于设计有效规则模式的专业知识，这显然是劳动密集型的。
2. 基于向量的表示，通过许多自然语言处理（NLP）技术，每个问题都可以作为潜在空间中的语义向量自动学习。最近，作为先前研究的延伸和结合，深度学习技术因其学习复杂语义的优越性而成为最先进的模型。例如，LSTM模型来捕捉疑问句的长期依赖性。卷积神经网络用于问题内容理解，目标是难度估计任务。问题：有监督的方法都存在标记数据稀缺的问题。也就是说，由于标签仅用于监督问题理解和任务建模的特定任务，因此对问题的理解非常有限，而大量未标记的问题数据库并未得到利用。此外，没有一项研究考虑了不同的问题输入形式，这导致了异构问题理解的信息丢失。

预训练（参考rql的报告吧…）：

1. 基于特征的方法，其中文本由某种特征提取器表示为固定向量。
2. 基于预训练的方法，其中模型参数在语料库上预训练，然后微调到特定任务。其中，最成功的模型是BERT。它利用Transformer以及与语言相关的培训前目标，以令人印象深刻的性能解决了许多NLP任务。

问题：很难直接应用于试题理解。首先，试题是异构的，其他输入形式中的许多信息都会被忽略。第二，测试问题包含许多领域逻辑和知识，需要理解和表示，而不仅仅是语言特征，这使得模型很难捕获。第三，由于模型修改的需要，这些方法中的许多方法很难应用，这在许多教育设置下是不方便的。

基于问题的应用领域：

1. 难度评估要求我们从问题的内容中评估问题的难度，而无需让一组学生准备对其进行测试。
2. 知识映射旨在将问题自动映射到相应的知识点。
3. 分数预测是一项预测学生在特定问题上的表现以及他们的锻炼历史的任务。

所有这些应用都有利于系统管理和服务，如个性化推荐。

QuesNet问题定义：

每个试题q是以异质形式给出，包括文本、图像、辅助信息meta-data（例如知识）。

输入：一个序列$x=\left\{x\_{t}\right\}, t\in \{1,…,T\}$，其中T是输入问题序列的长度。辅助信息用独热码m表示，$m\in R^{K}$，其中K是辅助信息中的类别数。每个输入项$x\_{t}$要么是词汇表中的一个单词（包括公式片段），要么是W×H的图像。

输出：内容表示向量，问题表示向量$v^{(s)}$。这里的每个向量维数均为N。

网络结构：

嵌入层、内容层、句子层



嵌入层（将异构内容嵌入到一个统一的形式中）：该层的目的是将异构输入内容投影到统一的空间，从而使我们的模型能够处理不同的输入形式。为了做到这一点，在第一层，我们设置了嵌入模块，将各种输入映射到固定长度的向量。

1. 单词的映射emb-w：参数$θ\_{we}$，直接映射到$N\_{e}$。（例如word2vec）
2. 图片的映射emb-i：参数$θ\_{ie}$，三个conv和激活函数+max-pooling，映射到$N\_{e}$。
3. Meta-data（第一个元素）的映射emb-m：参数$θ\_{me}$，两层神经网络，映射到$N\_{e}$。

因此，我们从得到了，公式：



内容层（捕捉深层的语言关系和语境）：在这一层中，我们的目标是为每个输入项建模关系和上下文。像LSTM这样的现有方法只关注一侧的上下文，而在Transformer中，上下文和关系建模依赖于**位置嵌入**，这会丢失一些局部性。因此，使用上述嵌入式向量序列作为输入，我们结合了多层双向LSTM结构，该结构更能获取上下文信息。在这里，我们选择Bi-LSTM是因为它可以从前后两个方向充分利用疑问句的上下文内容信息。

使用L层的Bi-LSTM：



在每一层上都要合并一次隐藏向量（文章只是说这样做是有好处的，但是代码里没有这样写，只是在计算loss的时候做了拼接）。



句子层（聚合成一个具有高级逻辑和知识的向量）：在我们对较低层次（上下文）的语言特征建模之后，我们仍然必须以一种关注长期和全局复杂关系的方式来聚合这些信息，以便能够捕获领域逻辑和知识。句子层是一个自我注意模块，用于将项目表示向量$v^{(i)}$聚合为句子表示$v^{(s)}$。与关注上下文的LSTM相比，注意力机制更能建模长期逻辑（即使信息在更远处，和上一篇stock是一样的）和全局信息。

Q、K、V都是$v\_{+}^{(i)}$：



Pe表示位置向量。

预训练过程：

设计了一种新的分层预训练算法。我们首先分别对每个嵌入模块进行预训练。然后在主要的预培训过程中，我们提出了两个层次的目标。

1. 在低水平的预训练中，我们提出了一种新的Holed-language model objective作为学习低水平语言特征的目标。
2. 在高级学习中，增加了一个Domain-oriented objective，用于学习高级领域逻辑和知识。

两个级别的目标在一个pre-trained过程中一起学习。



验证嵌入的正确性：

1. 单词：word2vec的正确性。
2. 图片：制作一个解码器$\tilde{emb}\_{i}$，使得loss最小（l是距离函数，实验中使用MSE）。



因此，$θ\_{ie}$应当是使得l最大的值，作为初始化的weights。



1. Meta-data：同上，略。

Holed Language Model：

低水平的预训练目标是从大型语料库中学习语言特征。传统模型缺点：

1. 语言模型（LM）作为最常用的无监督语言特征学习目标，受到其单向性的限制。
2. ELMo中的双向LM实现，在ELMo[23]中，双方的上下文分别进行训练，没有任何交互。
3. BERT依赖于令牌的随机掩蔽，而不是对采样的单词进行学习。

在本文中，我们提出了HLM，它联合了双方的上下文。直观地说，HLM的目标是用左右两侧的上下文填充每个单词。即最大化每个单词在其位置出现的概率：



对于后面的序列来说，可以利用LSTM中的隐向量进行分析，即：



对于不同的输入，需要不同的loss建模方式：

1. 单词：一个全连接层out-w，交叉熵损失函数。输出层以$h\_{¬t}$为输出，生成一个经过softmax的向量，建模每个单词在位置t上的出现概率。
2. 图像：在decoder后使用一个全连接层out-i，MSE损失函数。
3. Meta-data：同上，略。

损失函数公式：

 

Domain-Oriented Objective：

低水平的HLM损失只帮助模型学习语言特征，如关系和语境。然而，领域逻辑和知识仍然被忽视。为了在最终表达中也包含这些信息，在本节中，我们设计了一个面向领域的高级预培训目标。

例子：异质问题例子中的内容和选项之间的关系包含许多领域特定的逻辑和知识

解决方式：设定预训练任务，在给定选项的情况下，模型应该输出是否正确的答案。



其中，D是一个1维的全连接神经网络（其实就是加权平均）。因此，我们的目标是让选项出现的概率更高，损失函数如下：



最终目标：



综上，QuesNet问题表示应该能够捕获低级语言特征和高级领域逻辑和知识，并将对问题的理解转移到教育领域的下游任务中。

Fine-tuning：

教育领域的下游任务往往相当复杂。例如：

1. 知识映射。在这个多标签问题上使用了细粒度模型，这需要对每个输入内容进行表示。
2. 分数预测。每个练习（测试问题）都表示为一个向量，然后作为序列模型的输入。

正如我们所看到的，不同的任务需要不同的问题表示。要将QuesNet表示应用于特定任务，我们只需提供所需的表示，以替换下游模型的等效部分，从而将模型修改的成本降至最低。此外，在每个下游任务中，只需要对QuesNet进行一些微调，这将导致更快的训练速度和更好的结果。

总结：

首先，它为异构问题提供了统一且普遍适用的表示。

第二，它不仅能够学习低层次的语言特征，如关系和语境，还能够学习高级领域逻辑和知识。

第三，它很容易应用于下游任务并进行微调。

数据集：<http://www.zhixue.com>

所有数据都是从高中数学测试和考试中收集的。



数据集显然是异构的，如表所示。约25%的问题包含图像内容，约72%的问题包含side information。知识映射（KM）任务中使用的辅助信息是一些其他问题元数据（等级，表1中的问题数量）。在所有其他任务（难度估计DE、学生表现SP）中，知识的概念用作辅助信息。

问题平均包含约60个单词，但所包含的信息大于60，因为问题文本中有大量以latex表示的公式。

任务：

知识映射：知识映射任务的主要目标是将给定问题映射到相应的知识。这是一个多标签任务，其中大约13000个问题被标记（仅占整个未标记问题数据集的1.98%）。为了说明表示方法如何缓解这个标记较少的问题以及它如何执行此任务，我们选择了最先进的知识映射模型，并将其问题表示部分替换为我们想要比较的每个表示模型（包括QuesNet）。经过微调后，我们使用了一些常用的多标签问题评估指标，包括准确性（ACC）、精确性、召回率和F-1分数。

难度预测：一个更难的回归任务。打上标签的数据更少，因为只有0.37%的问题已经贴上标签。同时，任务需要更多的领域逻辑和知识作为指导，以获得更好的性能，因为对练习难度的评估需要更深入地理解问题。此任务的数据集仅包含2400个问题。指标有MAE、DOA和PCC。

分数预测：非常更加难的问题。主要目标是按顺序预测学生在他们练习的每个测试问题上的表现[27]。虽然学生记录规模很大，但这项任务的问题数量非常有限，只有2.22%左右。为了更好地为学生练习序列建模，一些最先进的模型将问题内容与问题表示结合到建模中。我们用我们的比较方法替换该模块，并使用MAE、RMSE、ACC和AUC评估性能。

网络细节：

1. 嵌入模块的所有输出向量大小均为128。
2. 图像嵌入模块和相关解码器分别实现为4层卷积神经网络和转置卷积神经网络。每层的卷积大小分别为16、32、32、64。
3. Meta嵌入模块和相关解码器被设置为两层完全连接的神经网络。隐藏层的大小设置为256。
4. 对于QuesNet的主要部分，我们使用4层Bi-LSTM和1层自注意力。这些模块中隐藏状态的大小设置为256。
5. 为了防止过度拟合，我们还在每层之间引入了dropout=0.2。
6. 使用Xavier初始化策略初始化层（即每一层输出的方差应该尽量相等）。
7. 在预训练过程中，通过Adam优化算法更新参数。
8. 对于每个任务中的优化器，我们都遵循相应论文中描述的设置。
9. Linux集群测试，并针对Tesla K20m GPU进行了优化。

Baseline：

Original：没有预训练，直接学。

ELMo：LSTM的预训练，仅对文本表示有用，因此忽略图像与meta。

BERT：同样的，忽略图像与meta。

H-BERT：适用于异质数据，使用与QuesNet一样的嵌入维数。

结果与结论：



1. QuesNet对问题有了更好的理解，并且能够更有效地从大型未标记语料库转移到小型语料库。然而，本表中还有更多需要解释的内容。
2. 支持异构输入的模型比没有异构输入的类似结构具有更好的结果，这证明在理解问题时处理异构输入是至关重要的。
3. 由于所有方法都经过调整以包含相似数量的参数，因此QuesNet是最有效的方法。
4. 基于Transformer的方法的结果略低于其他基于LSTM的方法。这可能是因为在BERT中使用的掩蔽语言模型预训练策略的样本效率较低，而在ELMo中使用的双向语言模型的性能更好。但是不管怎样，在QuesNet中使用的新型holed language model的性能优于两者。

消融实验：



QN-T、QN-I、QN Mrefer to QuesNet仅包含文本、图像或侧边信息，分别用于预训练过程。

QN-TI、QN-TM、QN-IM是指不同类型输入的组合。

QN-L指的是仅包括低水平预培训目标（holed language model）的QuesNet。

QN-H指的是仅包括高级领域目标的QuesNet。

结论：

1. 模型包含的信息越多，性能越好，这符合直觉。
2. 如果我们关注QN-H和QN-L之间的比较，我们会注意到它们在不同的任务中获得不同的效果。在更注重低层次特征（如知识映射）的任务上，QN-L略优于QN-H，而在其他更面向领域的高层次任务（难度估计和学生成绩预测）上，QN-H表现稍好。这清楚地表明了这两个目标关注的不同方面。当然，QuesNet优于QN-L和QN-H。
3. QN-T的性能比QN-I好，QN-I比QN-M好，这表明文本在问题中承载了大部分信息，然后是图像，然后是meta。但是，由于省略这两种输入都会导致性能损失，因此异构输入中的所有信息对于理解好问题至关重要。

未来方向：

1. 我们可能会研究一些特定于领域的模型架构，以更细粒度的方式对问题中的模型逻辑进行研究。
2. 对QuesNet的理解是不可理解的，在未来，我们将致力于理解和解释，以产生更具说服力的表述。
3. 我们的方法的总体思想适用于更异构的场景，我们希望进一步探索我们在其他异构数据和任务上的工作的可能性。