图像信息对句子语义理解与表示的有效性验证与分析

张琨[1)] 吕广奕[1)] 吴乐[2)] 刘淇[1)] 陈恩红[1)]

[1)](中国科学技术大学计算机科学与技术学院 合肥 230027)

[2)](合肥工业大学计算机与信息学院 合肥 230601)

摘 要 现如今，图像文本建模研究已经成为自然语言处理领域一个非常重要的研究方向。图像常常被用于增强对句子语义的理解与表示。然而有研究人员质疑图像的必要性，因为文本已经能够提供一个强有力的先验知识，帮助模型取得非常好的效果；甚至在不使用图像的条件下就能得出正确的答案。因此研究图像文本建模需要首先回答一个问题：图像是否有助于句子语义的理解与表示？为了回答这个问题，本文选择一个典型的不包含图像的自然语言语义理解任务：自然语言推理，并将图像信息引入到该任务中用于验证图像信息的有效性。由于自然语言推理任务是一个单一的自然语言任务，在数据标注过程中没有考虑图像信息，因此选择该任务能够更客观的分析出图像信息对句子语义理解与表示的影响。具体而言，本文首先提出一种通用的即插即用框架（general plug and play framework）用于图像信息的整合。基于这个框架，本文选择五个目前最先进的自然语言推理模型，对比分析这些模型在使用图像信息前后的表现，以及使用不同图像处理模型与不同图像设置时的表现。模型在一个大规模公开数据集上进行了大量实验，实验结果证实图像作为额外知识，确实有助于句子语义的理解与表示。并且证实了不同的图像处理模型和使用方法对整个模型的表现也会造成不同的影响。

关键词 图像文本建模；句子语义理解与表示；图像信息；即插即用框架；自然语言推理

中图法分类号 TP301

Analysis of the Effectiveness of Additional Images for Sentence Semantics

Zhang Kun[1)] Lv Guangyi[1)] Wu Le[2)] Liu Qi[1)] Chen Enhong[1)]

[1)](School of Computer Science and Technology, University of Science and Technology of China, Hefei230027)

[2)](School of Computer Science and Technology, HeFei University of Technology, Hefei230039)

Abstract Recently, the Visual-to-Language (V2L) problem has attracted more and more attention and become an important research topic in natural language processing. By utilizing Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Attention Mechanism method, researchers have made full use of images and achieved much progress in V2L problem, especially in the area of natural language semantic understanding. In fact, images are often treated as the important auxiliary information for the enhancement of sentence semantic understanding. However, some researchers have begun to question the necessity of images. They argue that textual information has already provided a very strong prior to promise the good performance of most semantic understanding models, which are even capable of generating correct answers without the consideration of images. Thus, the very first problem of V2L research is that: is the image information really necessary and helpful for

本课题得到国家杰出青年科学基金(No. 61325010)、国家自然科学基金 (No. 61403358)资助。张琨，男，1990 年生，博士研究生，主要研究领域为自然语言理解与表示，E-mail: zhkun@gmail.com. 吕广奕，男，1990 年生，博士研究生，主要研究领域为自然语言处理，计算机视觉，E-mail: gylv@mail.ustc.edu.cn. 吴乐，女，1988 年生，博士，副教授，计算机学会(CCF)会员（40748G）,主要研究领域为教育数据挖掘及知识发现、推荐系统、社交网络，E-mail: lewu.ustc@gmail.com. 刘淇，男，1986 年生，博士，特任教授，计算机学会(CCF)会员（E200037367M）,主要研究领域为数据挖掘与知识发现、机器学习方法及其应用，E-mail: qiliuql@ustc.edu.cn. 陈恩红，男，1968 年生，工学博士，中国计算机学会（CCF）会士（09201F）和中国人工智能协会（CAAI）会士（E660910027F），安徽省计算机学会理事长，主要研究领域为数据挖掘，E-mail: cheneh@ustc.edu.cn

sentence semantic understanding and representation? To answer this question and investigate the effect of images, in this paper, we focus on a typical sentence semantic understanding task without images: Natural Language Inference (NLI), which requires an agent to determine the semantic relation between two sentences. Then, we incorporate images as the auxiliary information into the sentence pair to verify their effect and answer the question that has been raised before. Since it is originally a pure natural language task and images are not considered during the data annotation process and sentence semantic modeling process, choosing the NLI task for evaluation can help to assess the influence of image information on sentence semantic understanding and representing more objectively. To be specific, we first design a general plug and play framework for image utilization and integration, which consists of four general layers (i.e., Input Embedding layer, Contextual Encoding Layer, Interaction Layer, and Label Prediction Layer) and two plug and play layers (i.e., Fine-Grained Context-Enhanced Layer and Coarse-Grained Context-Enhanced Layer). Based on this plug and play framework, we then reproduce five state-of-the-art NLI models (i.e., Hierarchical BiLSTM Max Pooling model, Enhanced Sequential Inference model, Multiway Attention Network model, Stochastic Answer Networks model and Generalized Pooling method) with the same deep learning framework (i.e., TensorFlow). Next, we evaluate their performances with or without images on a large annotated NLI dataset (i.e., Stanford Natural Language Inference dataset). In order to better verify the role of images, we also compare the performances of models with different image processing methods (VGG19 and ResNet50) and different image utilization methods (Fine-grained method and Coarse-grained method). At last, extensive experimental results reveal that images, as the external knowledge, are really helpful for sentence semantic understanding. Furthermore, we have obtained some other conclusions: 1) Fine-grained image utilization method is capable of providing much more useful information. On the other hand, this kind of method has a greater influence on the sentence semantic understanding and representation of models; 2) As a more advanced method, ResNet50 can extract the important information from images more precisely than VGG19, which is able to provide much more comprehensive auxiliary information for sentence semantic understanding and representing models.

Key words Visual-to-Language; Sentence Semantic Understanding and Representing; Image Information; Plug and Play framework, Natural Language Inference.

1 引言

句子语义理解与表示是自然语言处理（NLP）中一个重要的研究内容。该研究要求一个模型能够利用给定的信息（图像或者文本）分析目标句子的语义，并且服务于其他具体的任务。例如：在复述识别（Paraphrase Identification，PI）任务中，模型需要将给定的两个句子作为彼此的情境信息，分析这两个句子是否表达同一个意思[1]；在自然语言推理（Natural Language Inference，NLI）任务中，模型需要以前提句子为情境信息，判断是否能从前提句子推理出假设句子的语义[2][3]。现如今，通过图像文本建模（Visual-to-Language, V2L）对句子语义进行理解与表示也受到越来越多的关注，例如视觉问答（Visual Question Answering，VQA）[4]，视觉推理（Visual Reasoning）[5][6]等。认知科学相关的研究也证实其他模态的信息（例如图像）对句子语义理解增强有着巨大的帮助[7][8]。

然而，虽然文本信息已经能够提供一个强有力的先验信息，帮助模型取得一个非常好的效果[9][10]，甚至在不使用图像的条件下就能得出正确的答案，但是在复杂情况下图像依然可以发挥重要的作用。图 1（A）展示这样的一个例子，该例子来自视觉问答数据集 VQA v1.0[11]，当提出“*Is the grass taller than the baby?*”这样一个问题时，大多数情况下答案都是“*Yes*”，因为“*baby*”是非常小的。此时图像信息似乎并没有那么大的用处。但当处理相对复杂的任务，例如自然语言推理时，图像信息又发挥着巨大的作用，例如图 1（B）给出的例子，该例子来自自然语言推理数据集 SNLI[2]，原始数据集中并不包含直接的图像信息。当判断两个句子之间的语义推理关系时，可以发现前提句子的语义是模糊的，无法判断前提句子中的天气是什么样的。尽管人们可以利用先验知识从“*outside market*”推理出“*sunny day*”，但这并不是确定无疑

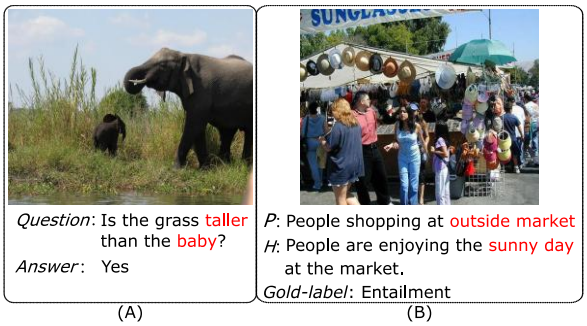


图 1：来自两个不同数据集中的语义理解例子

的，更别说连先验知识都没有的模型。但当引入图像信息增强对句子的语义理解时，就可以很轻松的判断出这两个句子之间的语义推理关系是*蕴涵*。

因此，为了验证视觉图像信息是否有助于对句子语义的理解与表示，本文选择一个典型的自然语言语义理解任务：自然语言推理，来验证图像信息的有效性。选择该任务的原因是因为原始的自然语言推理是一个纯文本理解任务，在整个数据标注的过程中并没有引入图像信息的影响，因此该任务能够更客观的展示图像信息的引入对句子语义理解的影响。具体而言，本文设计一种通用的即插即用框架（general plug and play framework），能够以多种不同的形式灵活的将图像信息整合到语义建模的过程中。基于该框架，本文复现五种目前最先进的方法，分别是 Hierarchical BiLSTM Max Pooling (Hbmp)[12]，Enhanced Sequential Inference model (ESIM)[13]，Multiway Attention Network (Mwan)[14]，Stochastic Answer Networks (SAN)[15]以及 Generalized Pooling method (GP)[16]。这些方法代表目前自然语言推理任务中两种最常用的框架：句子编码框架和词匹配框架。

除了文本处理方法，在图像处理方法上本文选择目前最常用的两种方法：VGG19[17]和 Resnet50[18]。同时，为了更好的验证图像信息的影响，本文选择两种不同的图像特征表示方法：*粗粒度方法*：选择两种图像模型倒数第二层的全连接层的输出作为图像信息的向量表示，并将该向量表示整合到通用框架中的匹配层；*细粒度方法*：选择图像模型的最后一个卷积层的输出结果作为图像特征的矩阵表示，并将矩阵表示整合到通用框架中的情境信息增强层。最后本文设计一系列的实验验证图像信息对句子语义理解的影响，并进行深入分析，最终得到一些发现：1）图像信息确实有助于理解与表示句子语义；2）细粒度的图像使用方法可以提供更多有用的信息，对文本语义理解与表示的影响更大；3）Resnet50[18]相对于 VGG19[17]模型能够抽取更准确的图像特征信息，为句子语义理解提供更全面的辅助信息。

2 相关工作

本文的相关工作可以分为三个部分：1）自然语言推理：主要介绍利用文本信息判断两个句子之间的语义推理关系的相关工作；2）图像文本建模：主要介绍通过图像信息辅助自然语言理解的相关工作；3）视觉自然语言推理：主要介绍利用图像增强句子语义表示与推理的相关工作。

2.1 自然语言推理

随着大规模数据集，例如SNLI[2]，SCITAIL[19]等的不断发布，以及各种各样神经网络技术，例如CNN[20]，LSTM[21]和注意力机制[22][23]等迅速发展，大量理解与表示自然语言句子语义的方法被提出来用于解决自然语言推理问题。这些方法主要分为两种框架：句子语义编码框架与词匹配框架。

句子语义编码框架通过生成固定长度的句子语义表示向量，利用这些句子语义表示向量来预测两个句子之间的语义关系。目前已有大量基于该框架的方法被提出来，例如 TBCNN[24]，CAFE[25]，和 DRCN[26]。这些方法的核心是通过从不同角度编码句子语义增强对句子语义的理解与表示。特别地，注意力机制能够根据实际需求为输出选择最合适的输入，因此注意力机制被大量应用于句子语义表示方法中。例如：Liu et al.[27]提出内部注意力来模仿人类在阅读时更关注那些重要词的行为，接下来，他们使用平均池化生成句子语义的向量表示。Chen et al.[16]将注意力机制扩展为多头形式，并生成多个不同的句子语义表示向量用于表示句子句子不同方面的语义。除此之外，层次化结构，例如层次化的 BiLSTM 和最大池化操作[12]，也被用来从多个不同角度构建句子语义的向量表示。

第二种框架更多地关注句子之间的词语义对齐以及词级别的句子间语义交互。例如：Rocktäschel et al.[28]提出词级别的注意力机制用于获取词与句子之间的的注意力分部信息。Chen et al.[13]利用互注意力建模词级别的局部推理关系。除此之外，他们还使用词级别的启发式匹配方法以一个更细粒度的方式建模句子之间的语义关系。Tan et al.[14]采用多个不同的注意力计算方法，从词级别匹配两

个句子之间的语义关系。他们声称不同的注意力计算方法能够帮助识别不同的关系类型。除此之外，一些研究人员将额外的先验知识引入到推理过程中，Chen et al.[29]将两个句子之间的同义词，反义词，上下位词等先验信息显式编码，然后引入到注意力计算，局部推理收集以及推理关系整合等模块，实现了语义推理关系的准确识别。然而，大多数的这类方法都更关注于句子文本本身，通过不同的方式从多个角度建模分析句子语义，但他们并没有考虑句子文本以外的信息（例如图像信息）对句子语义理解的辅助和增强作用。也就是说这方面的研究仍有很大的进步空间。

2.2 图像文本建模

近些年，将图像信息和文本信息联合起来进行图像文本建模已经成为一个非常热门的研究方向，大量图像文本相关的问题也不断的被提出来，例如图像描述生成（Image Captioning[30][31][32]）,视觉问答（Visual Question Answering[33]）,视觉对话（Visual Dialog[34]），视觉推理（Visual Reasoning[5][6]）等。

目前最好的图像文本联合建模方法通常会分别选择一个 CNN 和一个 RNN 作为图像和句子的编码器，用于生成图像与句子的特征表示。同时为了更加有效的整合这两种不同类型的信息，注意力机制一般也会被考虑进来。例如 Mao et al.[35]在每一步利用前一个词和图像的 CNN 特征结果来估计下一次词的概率分布，从而更好的生成图像的描述句子；Ma et al.[36]使用了不同的 CNN 同时处理图像特征和句子特征。接下来，他们将这些特征融合起来，用于生成输入问题的答案。更进一步，为了更好评估图像信息的影响，Zhang et al.[10]将二值视觉问答问题转换为图像区域有效性验证问题。他们试图回答句子的语义信息是否能够在图像中找到对应的视觉内容。通过这样的实验，图像对句子的抽象语义表达就能够被更好的验证。然而，相对于文本建模，图像建模所需的计算和存储开销更大，与此同时，在某些条件下，引入图像信息所带来的的文本建模效果提升并不明显，例如图 1 中的例子（A），在不考虑图像信息的情况下依然能得到正确的答案，这样的例子在 VQA v1.0[11]数据集中还有很多。因此仍然有一些研究人员认为句子文本本身的信息就能够提供强有力的先验，并保证模型能够取得非常好的表现，甚至可以在不考虑图像信息的情况下生成正确的答案[10]。因此，图像信息是否有助于句子语义的理解与表示仍然不清楚，需要在对句子语义理解要求更高的场景下进行更为深入的研究来验证图像信息对模型理解与表示句子语义的影响。

2.3 视觉自然语言推理

受图像文本建模的快速发展所启发，研究人员提出许多利用图像信息增强对句子语义的理解与表示的工作。特别地，有研究人员将图像信息引入到自然语言推理任务中，利用图像信息辅助对句子对的语义推理关系的判断。例如：Kun et al.[37]利用图像特征生成词语义的另一种表示，并将其与词的原始表示整合起来，用于增强对词与句子的语义表示。除此之外，他们还提出多层次的结构用于更全面的建模句子语义以及句子之间的语义关系。Xie et al.[38]提出一个新的自然语言推理数据集（V-NLI），在该数据集中，前提句子被对应的图像信息所替换。他们试图利用这个数据集验证细粒度的图像理解与表示。虽然目前有很多将图像信息整合到语义理解过程中的工作[39]。但这些方法大多是通过同时引入图像信息和设计不同的网络结构实现最终效果的提升，图像对句子语义理解是否有增强作用以及对模型效果的提升程度仍然不明确。为此，本文设计了一种通用的推理框架，通过在此框架下对比多个当前先进的自然语言推理方法在仅改变图像信息利用方式的条件下的表现，实现对图像信息的增强作用的准确验证与分析。这也是本文的主要研究内容。

3 问题定义与通用推理框架

3.1 问题定义

作为一个有监督分类问题，自然语言推理任务输入为前提句子的表示 $s^a=\{w_1^a, w_2^a, ..., w_{l_a}^a\}$ 和假设句子的表示 $s^b=\{w_1^b, w_2^b, ..., w_{l_b}^b\}$，目标是训练一个分类器，能够准确识别两个句子之间的语义关系 $y=\xi(s^a, s^b)$。其中，w_i^a 和 w_j^b 是前提句子中第 i 个词和假设句子中第 j 个词的 one-hot 向量表示，l_a 和 l_b 是前提句子和假设句子的句子长度，待预测的语义关系主要有蕴涵（Entailment, E），矛盾（Contradiction, C），中立（Neutral, N）。

为了验证图像信息对模型理解与表示句子语义的影响，本文将图像信息引入到自然语言推理过程中。因此，和传统的自然语言推理任务相比，本

文将图像信息 I 作为额外的一个输入，因此当预测

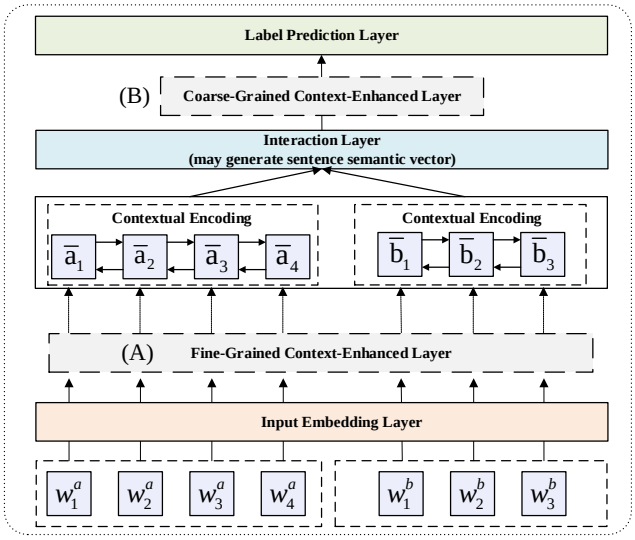


图 2 针对自然语言推理的通用即插即用框架

两个句子之间的语义关系时，模型形式化定义为 $y=\xi(s^a,s^b,I)$。

3.2 针对自然语言推理的通用框架

作为自然语言理解中一个基础但十分重要的任务，自然语言推理受到越来越多的关注[40]。研究人员提出大量的工作用于解决这个问题。这些方法大致分为两类框架：句子语义编码框架：将句子语义编码为一个向量表示，在此基础上利用各种不同的方法进行语义推理关系的判断[12][16]；词匹配框架：更关注句子中的细粒度词对齐以及细粒度的语义交互[13][14]。如图 2 实线框所示，本文首先将这两类框架统一到一个基本框架下，在此基础上介绍本文提出的通用的即插即用框架（general plug and play framework）。

自然语言推理的基本框架主要包含四层：1）输入编码层（Input Embedding Layer）；2）情境感知编码层（Contextual Encoding Layer）；3）交互层（Interaction Layer）；4）预测层（Label Prediction Layer）。接下来将详细介绍每一层的具体作用。

输入编码层（Input Embedding Layer）：这一层的输入为前提句子和假设句子中每个词的 one-hot 向量表示 $s^a=\{w_1^a,w_2^a,...,w_{l_a}^a\}$ 和 $s^b=\{w_1^b,w_2^b,...,w_{l_b}^b\}$。该层可以使用多种不同的方法编码每个词。为了充分利用大规模文本信息，研究人员会选择在大规模语料上训练好的词向量，例如 Word2Vec[41]，Glove[42]等。为了让词的语义表示更具有任务相关性，研究人员还会将字符级别的词向量[43]或者 ELMo[44]加入到该层中。同时一些语法特征[45][46]也会被用来增强每个词的语义表示。最终，该层的输出是每个词的丰富特征表示 $\{a_i\mid i=1,2,...,l_a\}$ 和 $\{b_j\mid j=1,2,...,l_b\}$。

情境感知编码层（Contextual Encoding Layer）：本层将上一层的输出 $\{a_i\mid i=1,2,...,l_a\}$ 和 $\{b_j\mid j=1,2,...,l_b\}$ 作为输入，通过整合句子内部的情境信息和序列信息生成句子中每个词更全面的语义表示。因此高速网络（Highway Network）[47], LSTM[21], CNN[48],或者 Transformer[23]经常会被用来作为句子语义表示的生成模块。同时，一些额外的先验知识也会被融入到该层中。该层的输出是句子中每个词的情境化向量表示 $\{\bar{a}_i\mid i=1,2,...,l_a\}$ 和 $\{\bar{b}_j\mid j=1,2,...,l_b\}$。

交互层（Interaction Layer）：该层的输入是句子中每个词的情境化向量表示 $\{\bar{a}_i\mid i=1,2,...,l_a\}$ 和 $\{\bar{b}_j\mid j=1,2,...,l_b\}$。为了建模分析两个句子之间的语义交互，本层通常选择注意力机制[42]建模句子之间的语义交互。对于词匹配方法，本层主要完成两个句子中的词对齐以及语义相似度和交互分析；对于句子语义编码方法，本层更关注于句子语义的表示。具体而言，本层会根据具体任务的不同选择不同的注意力计算方法，例如互注意力[26],多头注意力[43]和自注意力[23]等。需要说明的是，句子语义编码框架会在该层生成句子语义的向量表示。

预测层（Label Prediction Layer）：该层主要利用交互层的结果进行句子之间的匹配和分类。具体而言，本层利用启发式的匹配方法[13]建模两个句子之间的语义推理关系。然后使用多层感知机（MLP）和 softmax(·)函数进行最终的分类。

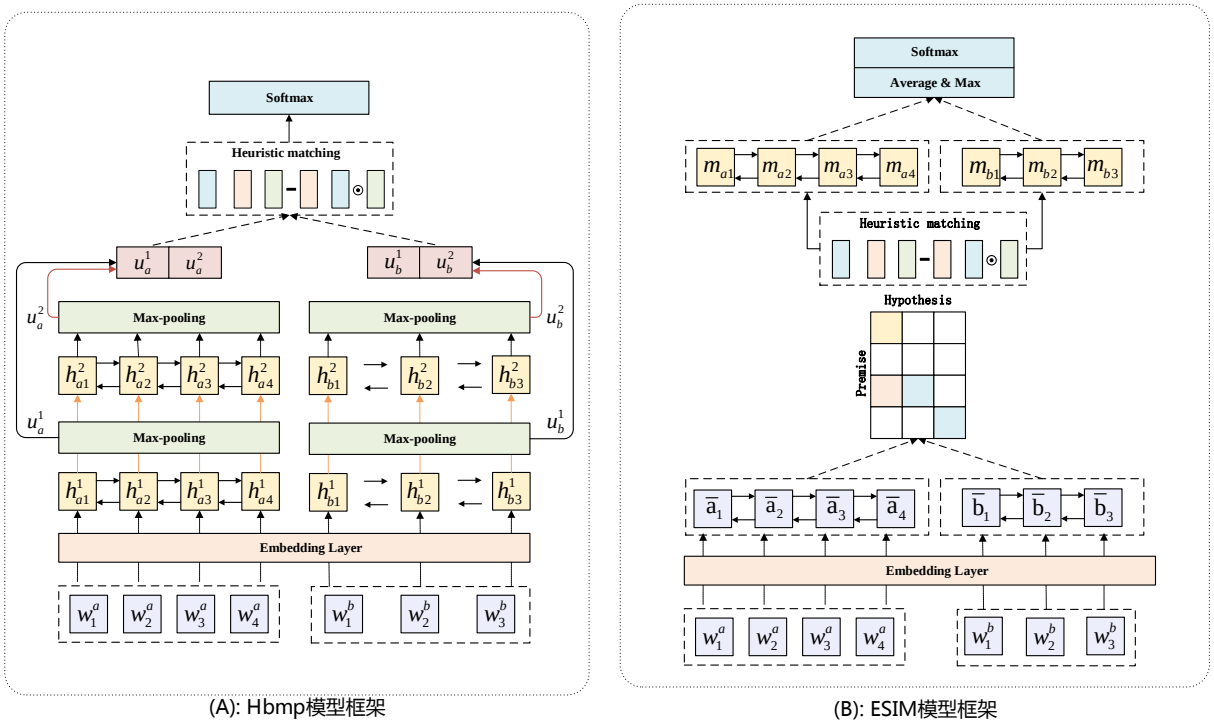
以上就是传统的自然语言推理方法的基本框架。为了更灵活的利用不同类型的情境信息，本文提出一种通用的即插即用框架（general plug and play framework），如图 2 所示，该框架主要增加了两层即插即用层：（A）细粒度的情境信息增强层（Fine-Grained Context-Enhanced Layer）；（B）粗粒度的情境信息增强层（Coarse-Grained Context-Enhanced Layer）。与其它层相比，这两层能够灵活运用各种不同的情境信息（例如图像信息或者知识图谱信息）增强对句子语义的理解。并且这两层能够灵活的从整个模型中加入或者删除，因此本文将该框架称之为即插即用框架。在接下来的描述中，本文以图像信息为例，具体介绍这两层。

细粒度的情境信息增强层（Fine-Grained

Context-Enhanced Layer）：为了更充分的利用情境信息，同时以一个细粒度的方式利用情境信息增强对句子语义的理解与表示，本文设计细粒度的情境信息增强层，利用图像情境信息对词级别的语义进行增强。如图 2 中的虚线框（A）所示，该层的文本输入为输入编码层的输出 $\{a_i \mid i=1,2,...,l_a\}$ 和 $\{b_j \mid j=1,2,...,l_b\}$，图像输入为细粒度的图像特征表示 $C=[c_1,c_2,...,c_{l_c}]$（例如 VGG19 模型的倒数第二个全连接层的输出结果），由于文本输入和图像输入均为矩阵表示，本文通过使用不同的融合方法（例如互注意力机制），从图像情境信息中抽取对每个词的语义表示最重要信息，利用这些信息从另一角度增强对词级别的语义理解，从而保证词的语义能够更为全面的建模。为了保证这一层的灵活性与即插即用特点，该层的输出和输入编码层的输出十分类似，依然是词级别的语义表示。

粗粒度的情境信息增强层（Coarse-Grained Context-Enhanced Layer）：除了细粒度的图像情境信息表示方法，图像情境信息也可以用一个单独的向量 c 表示，相对于细粒度的矩阵表示方法，该方法可以称之为粗粒度的表示方法。为了将这种表示整合到整个框架中，本文考虑将图像的单独向量表示与文本的单独向量表示进行整合。因此本文设计了粗粒度的情境信息增强层，用于整合情境信息的粗粒度表示。图 2 中的虚线框（B）展示该层的具体位置。考虑到文本特征表示与图像特征表示均为向量形式，对两个向量采用注意力机制进行建模意义不大，同时由于交互层已经整合两个句子之间的语义交互信息，本文直接将图像情境信息的向量表示拼接到对应的输出结果上，并将得到的结果输入多层感知机进行最后的分类。

正如前文所述，本文的目标是验证图像信息是否有助于理解与表示句子语义。因此，本文在实验验证过程中同样选择图像作为情境信息。通过添加删除图像信息，或者使用不同设置的图像信息验证



(A): Hbmp模型框架　　(B): ESIM模型框架

图 3 Hbmp 模型和 ESIM 模型的框架

图像信息对模型理解与表示句子语义的影响。需要强调的是该通用即插即用框架不仅能够灵活的添加或者删除情境信息增强层，同时能够非常灵活的使用各种不同的情境信息。

4 模型与即插即用框架技术介绍

本节主要介绍在通用即插即用框架下复现的五种自然语言推理模型的相关实现技术细节。

4.1 Hbmp模型

Hbmp 模型[12]是一种典型的句子语义编码方法。模型结构如图 3（A）所示。该模型利用层次化的结构进行句子语义编码。具体而言，Hbmp 模型使用层次化的双向LSTM和最大池化操作取得非常好的效果。简单起见，本文以前提句子的处理为例。层次化处理过程可以形式化为如下公式：

$$\begin{aligned} x_i^l &= [h_i^{l-1}, \bar{a}_i], \\ h_i^l &= \text{BiLSTM}(h_{i-1}^l, x_i^l), \\ u^l &= \text{maxpooling}([h_1^l, h_2^l, ..., h_{l_a}^l]), \end{aligned} \tag{1}$$

其中，x_i^l 和 h_i^l 分别表示第 l 层双向 LSTM 的第 i 个输入和隐层状态。u^l 表示第 l 层的最大池化操作的输出，同时也是第 l 层的句子语义表示。在得到所有层的句子语义表示之后，Hbmp 模型将所有的结果拼接到一起，最后利用一个多层感知机进行最后标签的预测。

4.2 ESIM模型

图 3（B）展示了 ESIM 模型[8]的整体结构。该模型将情境感知的句子编码层的输出 $\{\bar{a}_i \mid i = 1, 2, ..., l_a\}$ 和 $\{\bar{b}_j \mid j = 1, 2, ..., l_b\}$ 作为注意力机制的输入，并使用互注意力[21]建模两个句子之间的局部推理关系：

$$\begin{aligned} e_{ij} &= \bar{a}_i^T \bar{b}_j, \\ \tilde{a}_i &= \sum_{j=1}^{l_b} \frac{\exp(e_{ij})}{\sum_{k=1}^{l_b} \exp(e_{ik})}, \forall i \in [1, 2, ..., l_a], \\ \tilde{b}_j &= \sum_{j=1}^{l_a} \frac{\exp(e_{ij})}{\sum_{k=1}^{l_a} \exp(e_{kj})}, \forall j \in [1, 2, ..., l_b], \end{aligned} \tag{2}$$

在此之后，ESIM 模型使用启发式的匹配方法对句子之间的局部推理关系进行增强分析，并使用基于树结构的 LSTM（Tree-LSTM）整合学习到的信息，该过程可以形式化为如下表示：

$$\begin{aligned} v_i^a &= TreeLSTM([\bar{a}_i; \tilde{a}_i; (\bar{a}_i - \tilde{a}_i); (\bar{a}_i \odot \tilde{a}_i)]), \\ v_j^b &= TreeLSTM([\bar{b}_j; \tilde{b}_j; (\bar{b}_j - \tilde{b}_j); (\bar{b}_j \odot \tilde{b}_j)]). \end{aligned} \tag{3}$$

最后，使用平均池化和最大池化处理这些信息，并将得到的结果拼接起来生成语义推理关系表示向量 v，并通过一个多层感知机进行最后的分类。

4.3 Mwan模型

Mwan 模型[14]的整体结构如图 4（A）所示。这是一个词匹配方法。该模型使用多种不同的注意力计算方法对句子语义进行匹配。具体而言，Mwan 模型设计四种不同的注意力计算方法（拼接注意力（concat attention），双线注意力（bilinear attention），点乘注意力（dot attention），相减注意力（minus atention））匹配句子对中词之间的语义关系，为了方便描述，本文仅展示为归一化之前的四种注意力权重计算方法：

$$\begin{aligned} e_{ij}^c &= v_c^T \tanh(W_c \bar{a}_i + U_c \bar{b}_j), \\ e_{ij}^b &= \bar{a}_i W_b \bar{b}_j, \\ e_{ij}^d &= v_d^T \tanh(W_d (\bar{a}_i \odot \bar{b}_j)), \\ e_{ij}^m &= v_m^T \tanh(W_m (\bar{a}_i - \bar{b}_j)), \end{aligned} \tag{4}$$

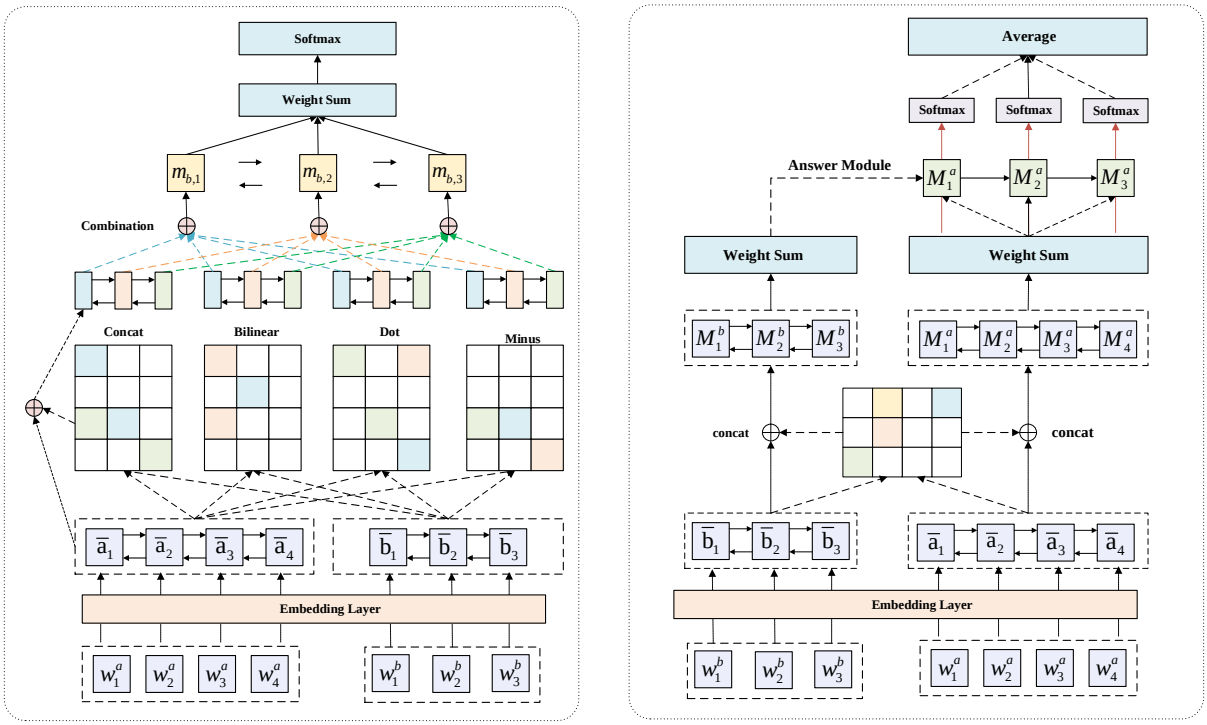
其中 $e_{ij}^c, e_{ij}^b, e_{ij}^d$ 和 e_{ij}^m 表示前提句子中第 i 个词和假设句子中第 j 个词之间未归一化的注意力权重值。c, b, d, m 分别表示四种不同的注意力计算方法。在得到未归一化的权重之后，Mwan 方法使用加权和的方法整合这些学习到的匹配信息，该过程可以由公式 5 形式化表示：

$$\begin{aligned} \tilde{a}_i^k &= \sum_{j=1}^{l_b} \frac{\exp(e_{ij}^k)}{\sum_{n=1}^{l_b} \exp(e_{in}^k)} \bar{b}_j, \forall i = 1, 2, ..., l_a, \\ \tilde{b}_j^k &= \sum_{i=1}^{l_a} \frac{\exp(e_{ij}^k)}{\sum_{n=1}^{l_a} \exp(e_{nj}^k)} \bar{a}_i, \forall j = 1, 2, ..., l_b, \end{aligned} \tag{5}$$

其中 $k \in [c, b, d, m]$。由于这些匹配信息仍然关注的是词级别，Mwan 模型使用残差连接和 GRU 整合来自四种不同注意力计算方法的匹配信息，然后使用加权和整合所有信息的信息。最后利用多层感知机来预测每个标签出现的概率。

4.4 SAN模型

图 4（B）展示了 SAN 模型[15]的整体结构。和其他的自然语言推理方法相比，SAN 模型更多的关注在决策过程。首先，SAN 模型和 ESIM 模型一样，利用互注意力[26]建模句子之间的语义交互，然后，SAN 模型使用双向 LSTM 针对两个句子生成一个工作记忆状态：



(A): Mwan模型框架　　(B): SAN模型框架

图 4 Mwan 模型和 SAN 模型的框架图

$$
\begin{aligned}
&\hat{a}_i = \text{concat}(\overleftarrow{a}_i, \overrightarrow{a}_i), \forall i \in [1,2,...,l_a],\\
&\hat{b}_j = \text{concat}(\overleftarrow{b}_j, \overrightarrow{b}_j), \forall j \in [1,2,...,l_b],\\
&M_i^a = \text{BiLSTM}([\hat{a}_i, \overrightarrow{a}_i], M_{i-1}^a),\\
&M_j^b = \text{BiLSTM}([\hat{b}_i, \overrightarrow{b}_i], M_{i-1}^b),
\end{aligned} \tag{6}
$$

其中，M_i^a 和 M_j^b 表示针对前提句子的第 i 个工作记忆状态和针对假设句子的第 j 个工作记忆状态。接下来，SAN 模型设计一个答案模块在 T 个记忆步上进行标签预测。首先，该模块利用 M^b 的加权和作为答案模块的初始状态：

$$
\begin{aligned}
&\alpha_j = \frac{\exp(\omega_b M_j^b)}{\sum_k \exp(\omega_b M_k^b)},\\
&s_0 = \sum_{j=1}^{l_b} \alpha_j M_j^b,
\end{aligned} \tag{7}
$$

ω_b 是模型在训练过程中的参数。接下来，SAN 模型采用一种 GRU 的变种来处理这些工作记忆状态，该过程可以形式化为如下：

$$
\begin{aligned}
&\beta_j = \text{softmax}(s_{t-1} W M^a), \forall t \in [1,2,...,(T-1)],\\
&x_t = \sum_{i=1}^{l_a} \beta_j M^a,\\
&s_t = \text{GRU}(s_{t-1}, x_t),
\end{aligned} \tag{8}
$$

其中 W 是模型需要训练的参数，T 是决策步数。在得到每一步的隐层状态之后，SAN 模型使用一个单层的分类器计算每一步得到的标签概率，然后使用平均池化整合所有的概率，从而得到最终的答案：

$$
\begin{aligned}
&P_t = \text{softmax}(U[s_t; x_t; (s_t - x_t); (s_t \odot x_t)]),\\
&P = \text{avg}([P_0, P_1, ..., P_{T-1}]).
\end{aligned} \tag{9}
$$

4.5 GP模型

GP 模型[16]也是一种典型的句子语义编码方法，正如第 3.2 小节中描述的，句子语义编码方法通常会在交互层生成一个句子语义的表示向量。但这是一个固定的向量，GP 模型认为“*该向量表示仅仅关注于句子的某个方面*”[16]。因此，它将交互层拓展为多头形式，并且生成多个句子语义的向量表示。这样句子语义的不同方面都能够被表示出来，这点和本文在第 2 章中的第一个方法是十分类似的。接下来，GP 模型通过一个拼接操作生成句子语义的最终表示，并使用多层感知机进行语义推

理关系的分类。

除此之外，GP 模型还设计三种不同的约束，用于减少多头操作的重复性，保证最终结果具有多样性，能够更全面描述句子语义的不同方面。首先，GP 模型在整个模型的参数矩阵上进行约束，保证不同的句子语义计算矩阵有着不同的参数，因此，GP 模型以如下形式最大化任意两个参数矩阵的 Frobenius norm：

$$L=\mu\sum_{i=1}^{l_h}\sum_{j=i+1}^{l_h}\max(\lambda-\|W^i-W^j\|_F^2,0), \quad (10)$$

这里 μ,λ 都是预先设定好的超参，W^i,W^j 是多头注意力计算中的不同矩阵。l_h 表示不同注意力计算的个数。

其次，GP 模型提出对注意力值的矩阵进行约束，用于保证注意力值的多样性，与文章[49]中通过添加约束 $\|AA^T\|_F^2$ 保证注意力值的矩阵的标量值的多样性相比，GP 模型使用如下方式保证注意力值的矩阵的向量值的多样性，其中，A^i,A^j 是多头注意力计算得到的不同注意力值的矩阵：

$$L=\mu\sum_{i=1}^{l_h}\sum_{j=i+1}^{l_h}\max(\lambda-\|A^i-A^j\|_F^2,0). \quad (11)$$

最后，GP 模型也对句子语义的向量表示添加约束，因为多头注意力操作能够能够生成多个句子语义的向量表示，所以 GP 模型提出在句子的向量表示上添加如下形式的约束：

$$L=\mu\sum_{i=1}^{l_h}\sum_{j=i+1}^{l_h}\max(\lambda-\|v^i-v^j\|_F^2,0). \quad (12)$$

4.6 基于通用的即插即用框架的解决方案

本节主要介绍通用即插即用框架中细粒度的情境信息增强层与粗粒度的情境信息增强层的相关技术细节。正如第 3.2 小节所描述的，本文提出的通用即插即用框架主要关注两个额外的层：（A）细粒度的情境信息增强层；（B）粗粒度的情境信息增强层。并且在本文中，图像信息被选作额外的情境信息对模型理解与表示句子语义进行增强。

4.6.1. 细粒度的情境信息增强层

如第 3.2 小节描述的，输入编码层的输出是每个词的丰富特征表示 $\{a_i\mid i=1,2,...,l_a\}$ 和 $\{b_j\mid j=1,2,...,l_b\}$。为了能够利用图像的细粒度特征表示 $C=[c_1,c_2,...,c_{l_c}]$ 增强句子中的每个词的语义表示，本文提出一种互注意力的变种用于将图像信息融入到词的语义表示中。由于互注意力[26]能够从细粒度的角度建模两种特征之间的交互，该操作能够选出对句子语义重要的信息增强对语义的理解与表示。为了简单起见，本文以前提句子的处理过程为例介绍相关细节：

$$\begin{aligned}
&e_{ij}=v_f^T\tanh(a_iW_fc_j),\\
&\hat{a}_i=\sum_{j=1}^{l_c}\frac{\exp(e_{ij})}{\sum_{k=1}^{l_c}\exp(e_{ik})}c_j,\forall i\in[1,2,...,l_a],\\
&\delta_i=\sigma(u_f^Ta_i),\\
&a_i:=[\delta_ia_i;(1-\delta_i)\hat{a}_i],
\end{aligned} \quad (13)$$

其中 v_f,W_f,u_f 是注意力训练过程中的参数。$\hat{a}_i$ 是前提句子中第 i 个词的候选表示。δ_i 是更新门，用于决定第 i 个词的输入表示有多少被保留。为了保证整个框架的一致性，在这里使用赋值操作 $:=$ 表示对每个词的丰富特征表示进行更新。在该方法中，注意力计算能够从图像中选择出重要的信息对句子语义表示进行增强，更新门保证模型始终关注在原始表示与图像增强表示中最重要的部分。这样图像信息就能够高效准确的融合到句子中每个词的语义表示中。

4.6.2. 粗粒度的情境信息增强层

与细粒度的情境信息增强层相比，粗粒度的情境信息增强层主要有两处不同。首先，该层使用一个向量 c 表示图像的特征信息；其次，该层直接将图像特征 c 和通用即插即用框架的标签预测层的多层感知机的输入进行拼接。接下来通过多层感知机预测每个标签所占的概率并决定最后输入的标签。假设多层感知机的输入为 v，该过程可以形式化为：

$$P(y\mid s^a,s^b,I)=MLP([v;c]), \quad (14)$$

其中多层感知机包含两层带非线性激活函数的 tanh(·)的变换层，以及一层 softmax(·)分类层。由于预训练的图像模型能够生成单一的图像特征向量表示，该方法希望验证是否粗粒度的图像特征表示也有助于对句子语义理解的增强。

5 实验

本节首先介绍新数据集构建以及对应的基本统计信息。然后，本文将介绍五种模型的复现细节以及相关实验结果的详细分析。

表 1 不同测试集的基本统计情况

数据集划分	数据规模			平均句子长度	
	蕴涵	矛盾	中立	前提句子	假设句子
Train	182,167	181,938	181,535	13.05	7.26
Dev	3,329	3,278	3,235	14.19	7.35
test	3,368	3,237	3,219	13.91	7.48
Hard	1,058	1,135	1,068	13.81	7.71
Lexical	782	5,164	43	11.42	11.60

5.1 数据集介绍

为了更好的对比在不同图像利用条件下模型的表现，本文选择 SNLI 数据集[2]作为基本数据集，并为每条数据添加一张对应的图像信息。选择该数据集基于以下两个原因：1）**客观性**：SNLI 数据集在生成过程中并没有考虑图像信息，标注员在人工标注时并不会受图像信息的影响。因此使用该数据集验证加入图像信息前后的模型效果更客观；2）**图像关系**：本文的研究重点是为了验证图像信息是否有助于句子语义的理解与表示，因此需要为原始不包含图像信息的数据集中添加图像信息，并且这些添加的图像信息需要和句子对之间有对应关系。正如[4]描述的，SNLI 数据集中的前提句子均来自于图像描述数据集 Flick30K[50]，每个前提句子都是一幅图像的描述句子，SNLI 数据集中也标注出每个句子对在 Flick30K 数据集中对应的图像名字，因此 SNLI 数据集非常适合这个任务。

具体而言，在数据预处理阶段，本文根据 SNLI 数据集中每个句子对中标注的图像名字，从 Flick30K 数据集[50]中抽取对应的图像信息，将其添加句子对中，用于构建新的包含图像信息的自然语言推理数据集，需要说明的是原始的 SNLI 数据集中有一部分数据无法找到对应的图像信息，因此本文将这部分数据删除。同时为了更好验证模型效果以及图像信息的作用，本文同时选择更具挑战的 Hard test[46]和 Lexical test[51]作为额外的测试集。加入图像信息之后的新的数据集的基本统计信息如表 1 所示。其中 Hard 和 Lexical 分别表示 Hard test[46]和 Lexical test[51]测试集。

5.2 模型实现

本文在开发集上选择所有的最优超参数，同时为了更客观的评价在通用框架下复现的所有模型，本文使用相同的参数训练所有的模型，因此在部分参数的设定会与原始模型的实现有不同，相关参数设定如表 2 所示。

表 2 模型超参数设定

超参数	参数值
VGG19 细粒度特征维度	512
VGG19 粗粒度特征维度	4,096
Resnet50 细粒度特征维度	2,048
Resnet50 粗粒度特征维度	2,048
预训练词向量维度	300
字符级别词向量维度	100
GRU/LSTM 隐层单元数	300
注意力单元数	200
多层感知机隐层单元数	200 和 100
初始学习率	10^{-4}

对于图像信息的表示，本文使用以 Tensorflow 为后端的 Keras 工具包实现的 VGG19 模型[17]和 ResNet50 模型[18]处理所有的图像，并使用最后一个卷积层的输出结果和倒数第二个全连接层的输出结果作为图像的细粒度特征表示与粗粒度特征表示。对于五个模型的复现，本文将预训练的词向量维度设为 300，字符级别的词向量维度为 100，dropout 设为 0.6，词向量使用的是预训练的 840B GloVe 词向量[42]。双向 LSTM 或者 GRU 的隐层单元数为 300。互注意力或者自注意力计算的隐层单元数设为 200，标签预测层的多层感知机的隐层单元数为 200 和 100。受[4]启发，模型中所有的权重使用范围在 $-\sqrt{6/(nin+nout)}$ 和 $\sqrt{6/(nin+nout)}$ 之间的均匀分布进行初始化[52]，nin 和 $nout$ 分别表示模型中权重参数所在层的输入和输出维度。所有的偏置初始化为 0。本文使用 Adam 优化器优化模型，初始学习率为 10^{-4}。

5.3 实验结果概述

本小节详细分析实验结果，需要说明的是，本文选择在不同测试集上的准确率作为模型的评价标准。

原始结果 v.s. 复现结果：表 3 展示了模型原始结果与在通用框架下复现的结果。从表中可以看出，本文复现的结果要稍低于模型的原始结果。本文总结出现这种情况的原因如下：1）本文使用统一的通用即插即用框架，并且为了更客观的评价模型效果，一些共有的超参数本文使用相同的参数设定；2）本文主要研究的是模型仅在使用文本信息以及使用文本和图像信息之间的表现对比，因此针

表 3 使用不同图像设置的模型复现效果（准确率）

图像使用方法	模型	原始结果			复现结果			VGG19 处理结果			Resnet50 处理结果		
		Full	Hard	Lexical	Full	Hard	Lexical	Full	Hard	Lexical	Full	Hard	Lexical
粗粒度方法	Hbmp	86.6%	-	-	85.2%	66.9%	68.2%	85.4%	67.3%	66.9%	85.6%	67.2%	69.6%
	GP	86.6%	-		83.9%	67.6%	69.1%	84.3%	66.6%	67.4%	84.5%	67.1%	69.7%
	Mwan	88.1%	-	-	86.6%	68.2%	65.3%	87.4%	69.3%	68.8%	87.5%	68.8%	70.0%
	SAN	88.5%	-	-	84.5%	64.3%	56.9%	86.2%	64.4%	58.7%	86.4%	65.3%	64.8%
	ESIM	88.0%	-	-	87.1%	71.3%	69.4%	87.9%	72.1%	70.6%	88.1%	72.6%	69.9%
细粒度方法	Hbmp	86.6%	-	-	85.2%	66.9%	68.2%	87.1%	69.5%	65.7%	86.8%	68.5%	67.4%
	GP	86.6%	-	-	83.9%	67.6%	69.1%	85.2%	70.4%	69.8%	85.0%	68.7%	70.6%
	Mwan	88.1%	-	-	86.6%	68.2%	65.3%	87.9%	69.8%	66.2%	88.3%	69.7%	70.7%
	SAN	88.5%	-	-	84.5%	64.3%	56.9%	85.8%	65.7%	56.8%	85.9%	64.1%	58.5%
	ESIM	88.0%	-	-	87.1%	71.3%	69.4%	88.2%	72.5%	68.7%	88.5%	71.8%	70.2%
整合粗粒度细粒度方法	Hbmp	86.6%	-	-	85.2%	66.9%	68.2%	87.1%	69.4%	65.9%	86.9%	68.5%	67.5%
	GP	86.6%	-	-	83.9%	67.6%	69.1%	85.3%	70.2%	69.7%	85.1%	68.8%	70.6%
	Mwan	88.1%	-	-	86.6%	68.2%	65.3%	88.0%	69.6%	66.4%	88.4%	69.7%	70.8%
	SAN	88.5%	-	-	84.5%	64.3%	56.9%	85.6%	66.1%	56.8%	85.7%	64.3%	58.2%
	ESIM	88.0%	-	-	87.1%	71.3%	69.4%	88.3%	72.7%	68.7%	88.5%	71.8%	70.3%

对每个模型，本文并没有进行额外的超参数调整。除此之外，本文还使用不同的深度学习框架，也可能会对模型有一些影响。例如原始的 ESIM 模型是使用 Theano 实现的，而本文是使用 Tensorflow 实现的。这些原因都会造成模型最终效果的不同。在接下来的章节中，本文将复现的结果作为基准结果，然后将该结果与使用不同图像信息设置得到的结果进行对比分析。

不使用图像结果 v.s. 使用图像结果：表 3 展示了在使用不同设置之后的模型结果。从表 3 展示的结果可以看出，使用图像信息之后，所有模型的表现均有不同程度的提升，一些模型的表现甚至要高于其原始文章中的结果，例如 Hbmp（87.1%）和 ESIM（88.2%）。与此同时，当比较模型在更具挑战的测试集（Hard test 和 Lexical test）上的表现时，可以发现大多数模型的表现均优于不使用图像信息的基准结果。这些现象都说明图像信息确实有助于模型对文本语义的理解与表示，并能提升其在下游任务上的表现。

细粒度方法 v.s. 粗粒度方法：在第 3.2 小节中，本文提出两个图像使用方法：细粒度图像特征表示与粗粒度图像特征表示。在本节中，本文对比这两种方法以及整合粗粒度和细粒度方法对模型表现的影响，相关实验结果如表 3 所示。首先，细粒度的图像特征能够更全面的表示图像的不同特征，因此模型能够利用注意力机制从这些特征表示中选择出合适的信息辅助对句子语义的理解与表示，从而取得更好的效果。于此相比，粗粒度的图像特征表示的原始目的是为了解决图像分类问题，因此它并不能像细粒度的特征那样为模型提供更准确的信息表示。整合粗粒度和细粒度方法的实验结果也证实了这一点，从实验结果中可以看出，模型在整合粗粒度和细粒度的图像特征表示条件下的表现和只使用细粒度特征条件下的表现相差不大，说明了细粒度的图像特征对模型的影响更大一些。其次，从表 3 中可以看出句子语义编码方法（Hbmp，GP）的效果提升力度要大于词匹配方法（Mwan，SAN，ESIM），甚至一些词匹配方法在使用细粒度的图像特征之后模型表现会出现一定程度的下降。对不同推理框架进一步分析之后发现，句子语义匹配方法更多的关注句子语义的向量表示生成，因此它可以从图像信息中选择合适的信息用于句子语义的增强表示，而下游任务最终使用的就是句子语义的向量表示，因此额外的信息并不会损害模型在具体任务上的表现；而词匹配方法更多的关注细粒度的词匹配以及词之间的语义交互，当加入额外的

细粒度图像特征信息时，一方面这些信息有助于增强理解句子中每个词的语义，但另一方面细粒度的图像特征信息可能会引入不相关的信息，并且误导模型错误匹配某些词之间的语义，最终造成模型在具体任务上的提升程度出现可能低于句子语义匹配方法的情况。

VGG19 模型 v.s. Resnet50 模型：作为典型的图像分类模型，VGG19 模型[17]和 Resnet50[18]在图像分类任务上已经取得令人瞩目的成绩，并且逐渐成为多种图像文本任务中必不可少的一部分，包括计算机视觉和自然语言处理。因此本文选择这两种模型来作为图像有效性验证实验中的图像处理模型。表 3 的结果中可以看出，使用 Resnet50 作为图像处理模型时，大多数自然语言推理模型均取得比较大的提升。特别的，可以发现当选择粗粒度的图像特征表示时，选择 Resnet50 进行图像处理，五个模型的表现要明显优于使用 VGG19 处理图像时的表现。同时，当检查两种模型生成的粗粒度图像特征表示的维度时，可以发现 Resnet50 生成的图像特征维度时 2,048 维，是 VGG19 生成的向量的维度（4,096）的一半。因此，可以总结出 Resnet50 生成的图像信息表示的质量要远高于 VGG19，同时图像信息表示的质量比图像信息表示的数量更为重要一些。Resnet50 相对于 VGG19 而言拥有更多的隐层，因此它能够生成质量更高的图像特征表示。

5.4 原始标签改变

本文深入研究图像信息的引入对原始便签的影响，图像信息是否会造成使用原始标签评价模型变得无效化。正如[2]中描述的，SNLI 数据集在标注时，标注人员并没有考虑图形信息，因此引入图像信息之后，原始的标签可能会发生变化。为了评估这种变化的影响，本文分别从 SNLI 数据集的 Full test 测试集和 Hard test 测试集采样 500 条样本，然后邀请 10 位相关的研究人员在考虑图像信息的条件下重新标注这些样本。

受混淆矩阵（Confusion Matrix）启发，本文在表 4 中重新展示相关的结果。每一列（行）表示原始便签（重新标注的标签）的数量（其中 E,C,N,O 分别指代蕴涵（Entailment），矛盾（Contradiction），中立（Neutral）和合计（Overall））。例如：第一列表示由 153 个蕴含例子被重新标注为蕴涵，有 6 个蕴含例子被重新标注为矛盾，有 7 个蕴涵的例子被重新标注为中立。本文使用 kappa 系数验证重新标注的标签与原始标签之间的一致性。如果 kappa 系数在 0.61~0.80 之间，表明两种分布是具有高度一致性的。从表中可以得出，在 Full test 测试集上的 kappa 系数为 0.759，Hard test 测试集上的 kappa 系数为 0.699，在所有采样样本上的 kappa 系数为 0.729。这些结果充分说明重新标注的标签和原始标签之间是高度一致的，因此图像信息的引入并不会造成使用原始标签评价模型的无效化。

表 4 不同测试集的基本统计情况

	Full test				Hard test			
	蕴涵	矛盾	中立	合计	蕴涵	矛盾	中立	合计
E	153	4	32	**189**	147	2	45	**194**
C	6	160	26	**192**	7	156	23	**186**
N	7	5	107	**119**	26	11	97	**120**
O	166	169	165	500	166	169	165	500

5.5 不同图像设置结果对比

本文已经通过实验验证图像信息的确有助于增强模型对句子语义的理解与表示。但是，大多数的图像都是由点，线，形状，颜色等特征组成。这些特征同样能够为模型提供额外的参考信息。是否任何一张图像均有助于句子语义的理解与表示呢？为了回答这个问题，本文进行额外的实验验证，对比在不同设置下图像的表现。

具体而言，本文在训练阶段，使用和第 5.3 小节相同的设置，但是在测试阶段，本文选择不同的设置。具体而言，当使用测试集验证模型效果时，本文利用数据集中一张随机的无关的图片替换原始图片，标记为错误图像，然后使用该图像验证模型在所有测试集上的效果，结果如表 5 所示。接下来本文将从三个方面对实验结果进行分析。

原始图像结果 v.s. 错误图像结果：首先，本文对比使用原始图像的结果和使用错误图像的结果。对比表 3 和表 5，可以发现五个模型在使用错误图像时的表现均有不同程度的下降，有的甚至比不使用图像的效果还差。这些现象说明图像确实提供有意义的信息，不是简单的点，线，形状等信息。同时，本文还发现使用错误图像时模型在更具挑战的测试集（Hard test 和 Lexical test）上的表现下降的更多。由于这些测试集将那些容易分类的例子都删除，因此需要模型能准确的理解并表示句子语义。而不相关的错误图像信息会引入噪音，使得模型理解句子语义变得更加困难，因此模型的表现也会有

表 5 使用错误图像的复现模型效果（准确率）

图像使用方法	模型	VGG19 错误图像结果			Resnet50 错误图像结果		
		Full	Hard	Lexical	Full	Hard	Lexical
粗粒度方法	Hbmp	83.5% (-1.9)	67.3% (-3.6)	63.3% (-3.6)	82.9% (-2.7)	62.6% (-4.6)	65.1% (-4.5)
	GP	82.3% (-2.0)	63.1% (-3.5)	62.5% (-3.9)	81.9% (-2.6)	62.8% (-4.3)	64.7% (-5.0)
	Mwan	84.6% (-2.8)	64.6% (-4.7)	64.3% (-4.5)	84.3% (-3.2)	63.7% (-5.1)	66.1% (-3.9)
	SAN	82.5% (-3.7)	60.4% (-4.0)	55.7% (-3.0)	82.3% (-4.1)	60.8% (-4.5)	62.7% (-2.1)
	ESIM	83.6% (-4.3)	65.3% (-6.8)	62.9% (-7.7)	84.1% (-4.0)	62.2% (-10.4)	65.5% (-4.4)
细粒度方法	Hbmp	82.6% (-4.5)	57.8% (-11.7)	61.4% (-4.3)	82.3% (-4.5)	58.2% (-10.3)	62.9% (-4.5)
	GP	82.3% (-2.9)	63.7% (-6.7)	61.9% (-7.9)	82.5% (-2.5)	63.6% (-5.1)	61.5% (-9.1)
	Mwan	84.9% (-3.0)	63.8% (-6.0)	62.7% (-3.5)	83.9% (-4.4)	61.7% (-8.0)	63.8% (-6.9)
	SAN	82.1% (-3.7)	60.1% (-5.6)	53.5% (-3.5)	82.1% (-3.8)	60.2% (-3.9)	59.7% (+1.2)
	ESIM	83.0% (-5.2)	62.4% (-10.1)	59.9% (-8.8)	83.5% (-5.0)	61.7% (-10.1)	63.8% (-6.4)

很大的下降。除此之外，从表中还可以发现词匹配方法在使用错误图像的情况下表现会下降更多。正如 2.1 节和 3.1 节所描述的，词匹配的方法更多的关注于细粒度的词对齐以及词语义交互，当使用错误图像信息时，模型会被误导使用不正确的信息对词的语义理解进行增强，从而做出错误的判断。

粗粒度图像特征表示 v.s. 细粒度图像特征表示：和第 5.3 小节中的分析类似，在该小节中，本文也验证不同粒度的错误图像使用方法对模型的影响。首先从表 5 中可以发现，细粒度的图像特征使用方法对模型的表现影响更大。由于细粒度的图像特征是直接作用于词级别的语义增强，不正确的信息将会误导模型错误理解词级别的语义，更别说句子级别的语义。因此当使用错误图像时，细粒度的图像特征表示方法会导致模型表现有更大的下降。其次，可以发现细粒度的图像特征对词匹配模型在更具挑战的测试集上的表现影响更大。这个现象表明对图像信息利用的越充分，图像信息对模型的表现影响就更大。再者，本文还发现一些不正常的现象：部分模型使用细粒度的图像特征时，模型表现下降的程度要低于使用粗粒度的图像特征，这是本文最初的发现是相互矛盾的。在对实验设定以及训练测试过程详细分析之后发现，模型在训练过程中，使用的实验设置是正确的图像信息，但当在测试集上评价模型表现时，实验设置发生改变，使用随机的不相关的错误图片，模型在不同测试集上的表现也就很难预测，因此在该条件下，各个不同模型会出现一些不正常的现象。

VGG19 模型结果 v.s. Resnet50 模型结果：通过对比使用不同图像处理模型时各个模型的表现，本文发现更多的证据证明在之前章节得到的结论。Resnet50 模型[18]相对于 VGG19 模型[17]有更深的网络结构，因此它能够抽取更全面，更准确的图像特征表示，也因此对模型表现有更大的影响。因此无论是使用正确图像对模型效果进行提升还是使用错误图像导致模型效果降低，Resnet50 模型带来的影响都要大于 VGG19 模型。

6 结论和展望

本文设计一种通用的即插即用框架用于验证图像信息是否有助于理解与表示自然语言句子语义。借助该框架，本文能够从多个不同角度更为全面验证图像信息对模型理解与表示句子语义的影响。同时自然语言推理是一个单一的句子语义匹配问题，在数据标注过程中并没有考虑图像信息。因此将图像信息引入到该任务中能够更客观的评价图像信息对句子语义建模分析的影响。本文复现 5 个最先进的自然语言推理方法，深入对比引入图像信息前后模型的表现。实验结果表明使用合适的图像信息能够不同程度提升各个模型理解与表示句子语义的能力。除此之外，本文还深入分析不同图像使用方法，不同图像处理模型对自然语言推理模型最终效果的影响。大量实验证明图像信息利用的越充分，对模型造成的影响越大。因此，从实验中可以发现细粒度的图像使用方法与词匹配方法在所造成的影响更大。更进一步，本文还将每个句子对对应的原始图像随机替换为任意一张不相关的图像，更好的分析图像信息对模型理解与表示句子语义的影响。

在接下来的工作中，本文将以一种更具体，更全面的形式验证图像信息对自然语言语义表示的影响，探索更好的图像文本联合建模方法，并将本文提出的通用即插即用框架扩展到更多的句子语义理解任务中。

参考 文 献

[1]. William B Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases//Proceedings of the Third International Workshop on Paraphrasing (IWP2005). 2005,

[2]. Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 2015, 632–642.

[3]. Samuel R Bowman, Jon Gauthier, Abhinav Rastogi, Raghav Gupta, Christopher D Manning, and Christopher Potts. A Fast Unified Model for Parsing and Sentence Understanding//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin, Germany, 2016, 1466-1477.

[4]. Peng Wang, Qi Wu, Chunhua Shen, and Anton van den Hengel. The vqa-machine: Learning how to use existing vision algorithms to answer new questions//Proceedings of the IEEE conference on computer vision and pattern recognition, Hawaii, USA,2017, 1173-1182.

[5]. Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA,2017. 1988–1997.

[6]. Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Vancouver, Canada. 2017, 217–223.

[7]. Mark Andrews, Gabriella Vigliocco, and David Vinson. Integrating experiential and distributional data to learn semantic representations. Psychological review, 2009, 116(3): 463.

[8]. Lawrence W Barsalou. Grounded cognition: Past, present, and future. Topics in cognitive science, 2010, 2(4): 716-724.

[9]. Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA,2017, 6904–6913.

[10]. Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and yang: Balancing and answering binary visual questions//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016, 5014–5022

[11]. Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering//Proceedings of the IEEE international conference on computer vision, Santiago, Chile, 2015, 2425–2433.

[12]. Aarne Talman, Anssi Yli-Jyrä, and Jörg Tiedemann. Natural language inference with hierarchical bilstm max pooling architecture. arXiv preprint arXiv:1808.08762, 2018.

[13]. Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, Hui Jiang, and Diana Inkpen. Enhanced LSTM for Natural Language Inference//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver, Canada, 2017, 1657–1668.

[14]. Chuanqi Tan, Furu Wei, Wenhui Wang, Weifeng Lv, and Ming Zhou. Multiway Attention Networks for Modeling Sentence Pairs//Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, Sweden, 2018, 4411–4417.

[15]. Xiaodong Liu, Kevin Duh, and Jianfeng Gao. Stochastic answer networks for natural language inference. arXiv preprint arXiv:1804.07888, 2018.

[16]. Qian Chen, Zhen-Hua Ling, and Xiaodan Zhu. Enhancing sentence embedding with generalized pooling. arXiv preprint arXiv:1806.09828 (2018).

[17]. Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv preprint arXiv:1409.1556, 2014.

[18]. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition//Proceedings of the IEEE conference on computer vision and pattern recognition. Copenhagen, Denmark, 2017, 770–778.

[19]. Tushar Khot, Ashish Sabharwal, and Peter Clark. SciTail: A textual entailment dataset from science question answering//Proceedings of Thirty-Second AAAI Conference on Artificial Intelligence. New Orleans,USA , 2018, 5189-5197

[20]. Wenpeng Yin, Hinrich Schütze, Bing Xiang, and Bowen Zhou. Abcnn: Attention-based convolutional neural network for modeling sentence pairs. Transactions of the Association for Computational Linguistics, 2016, 4: 259-272.

[21]. Jianpeng Cheng, Li Dong, and Mirella Lapata. Long Short-Term Memory Networks for Machine Reading// Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, USA, 2016, 551-561.

[22]. Ankur P. Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A Decomposable Attention Model for Natural Language Inference. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, USA, 2016, 2249–2255.

[23]. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Proceedings of the Advances in neural information processing systems. Long Beach, USA, 2017. 5998–6008.

[24]. Lili Mou, Rui Men, Ge Li, Yan Xu, Lu Zhang, Rui Yan, and Zhi Jin. Natural Language Inference by Tree-Based Convolution and Heuristic Matching//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Berlin, Germany, 2016, 130–136.

[25]. Yi Tay, Luu Anh Tuan, and Siu Cheung Hui. A Compare-Propagate Architecture with Alignment Factorization for Natural Language Inference. arXiv preprint arXiv:1801.00102, 2017.

[26]. Seonhoon Kim, Jin-Hyuk Hong, Inho Kang, and Nojun Kwak. Semantic Sentence Matching with Densely-connected Recurrent and Co-attentive Information//Proceedings of Thirty-third AAAI Conference on Artificial Intelligence. Hawaii, USA,2019, 6586-6593

[27]. Yang Liu, Chengjie Sun, Lei Lin, and Xiaolong Wang. Learning Natural Language Inference using Bidirectional LSTM model and Inner-Attention. arXiv preprint arXiv:1605.09090, 2016.

[28]. Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomás Kociský, and Phil Blunsom. Reasoning about Entailment with Neural Attention. arXiv preprint arXiv:1509.06664, 2015.

[29]. Chen, Qian and Zhu, Xiaodan and Ling, Zhen-Hua and Inkpen, Diana and Wei, Si, Neural natural language inference models enhanced with external knowledge//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 2018, 2406-2417.

[30]. Hao Fang, Saurabh Gupta, Forrest N. Iandola, Rupesh Kumar Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. From captions to visual concepts and back//Proceedings of the IEEE conference on computer vision and pattern recognition, Boston, USA, 2015, 1473–1482.

[31]. Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. Jointly Modeling Embedding and Translation to Bridge Video and Language//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016, 4594–4602.

[32]. Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator//Proceedings of the IEEE conference on computer vision and pattern recognition, Boston, USA, 2015, 3156–3164.

[33]. Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, and Anton van

den Hengel. What value do explicit high level concepts have in vision to language problems?//Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, USA, 2016, 203–212.

[34]. Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii,USA, 2017, 326-335.

[35]. Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L. Yuille. Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN). arXiv preprint arXiv:1412.6632, 2014.

[36]. Lin Ma, Zhengdong Lu, and Hang Li. Learning to Answer Questions from Image Using Convolutional Neural Network//Proceedings of Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, USA, 2016, 3567-3573.

[37]. Zhang Kun, Lv Guangyi, Wu Le, Chen Enhong, Liu Qi, and Wu Han. Image-Enhanced Multi-Level Sentence Representation Net for Natural Language Inference//Proceedings of IEEE International Conference on Data Mining (ICDM). Singapore, 2018, 747-756.

[38]. Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual Entailment: A Novel Task for Fine-Grained Image Understanding. arXiv preprint arXiv:1901.06706, 2019.

[39]. Kun Zhang, Guangyi Lv, Enhong Chen, Le Wu, Qi Liu, and CL Philip Chen. Context-Aware Dual-Attention Network for Natural Language Inference//Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, Macau, China, 2019, 185–198.

[40]. Adina Williams, Nikita Nangia, and Samuel Bowman. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). New Orleans, USA, 2018, 1112–1122.

[41]. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. Proceedings of the Advances in neural information processing systems. Lake Tahoe, USA, 2013, 3111–3119.

[42]. Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 2014, 1532–1543.

[43]. Yichen Gong, Heng Luo, and Jian Zhang. Natural Language Inference over Interaction Space. arXiv preprint arXiv:1709.04348, 2017

[44]. Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. arXiv preprint arXiv:1802.05365, 2018.

[45]. Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading Wikipedia to Answer Open-Domain Questions// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver, Canada, 2017, 1870–1879.

[46]. Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation Artifacts in Natural Language Inference Data//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). New Orleans, USA, 2018, 107–112.

[47]. Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway Networks. arXiv preprint arXiv:1505.00387, 2015.

[48]. Hua He, Kevin Gimpel, and Jimmy Lin. Multi-perspective sentence similarity modeling with convolutional neural networks// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal, 2015, 1576–1586.

[49]. Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. arXiv preprint arXiv:1703.03130, 2017.

[50]. Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics, 2014, 2: 67-78.

[51]. Max Glockner, Vered Shwartz, and Yoav Goldberg. Breaking NLI Systems with Sentences that Require Simple Lexical Inferences.//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Melbourne, Australia, 2018, 650-655.

[52]. Genevieve B Orr and Klaus-Robert Müller. Neural networks: tricks of the trade. Springer, 2003

Zhang Kun, born in 1990, Ph.D. candidate. His research interests include natural language processing and deep learning.

Lv Guangyi, born in 1990, Ph.D. candidate. His research interests include natural language understanding.

Wu Le, born in 1988, Ph. D. Associate Professor. Her research interests include Educational Data Mining and Knowledge Discovery in Database, Recommender System, Social Network.

Liu Qi, born in 1986, Ph. D. Professor. His research interests include Data Mining and Knowledge Discovery in Database, Machine Learning Method and Application.

Chen Enhong, born in 1968, Ph. D. Professor. His research interests include data mining and machine learning, social network analysis, and recommender systems.

Background

This paper focuses on leveraging image information to enhance the sentence semantic understanding and representing. Recently, Visual-to-Language (V2L) has become a hot topic and attracted more and more attention. It takes the images into consideration for the understanding and representing of sentence semantic and has broad applications, such as Image Captioning, Visual Question Answering, Visual Dialog, as well as Visual Reasoning. Current methods usually employee a CNN and an RNN as “encoders” for image and sentence semantic representations, respectively. In order to integrate these two types of information, attention mechanism is often utilized for the final decision. With the development of representation methods, such as transformer, Bert and GPT-2, researchers also try to leverage the transformer to model the images and sentences simultaneously. These Cognitive scientists have also advocated that other modalities (e.g., images) are quite helpful for semantic understanding enhancement.

In our work, we try to figure our whether image information can help to understand and represent sentence semantic. First, we focus on Natural Language Inference (NLI),

a typical sentence semantic understanding task, and introduce images as extra information to verify the effect. Then, we propose a general plug and play framework for flexible image utilization. Based on this framework, we re-implement five state-of-the-art NLI models and compare their performances with different image settings on a large annotated NLI dataset (SNLI). Finally, we present a series of findings with quantitative measurements and in-depth analyses.

This research was partially supported by grants from the National Science Foundation for Distinguished Young Scholars of China (Grant No. 61325010), the Natural Science Foundation of China (Grant No. 61403358)