自然语言推理数据实现通用句子表达的监督学习

一、 摘要

很多 NLP 系统都将 word embedding 作为 base feature，word embedding 通常是以无监督方式在大型语料库中训练得到的。但是，对于更长的文本（例如句子）的编码仍然存在很多困难。好几种无监督学习的句子表达都没有达到很好的性能，从而无法被广泛的使用。在这篇论文中，作者展示了在 Standford Natural Language Inference dataset 上训练得到通用 sentence representation 可以取得比一些非监督方法(例如 SkipThought）学习得到的 sentence representation 在很多任务上更好的效果。就像在计算机视觉领域，利用从 imagenet 数据上学到的特征来帮助完成其他任务，我们也希望利用自然语言推理来帮助完成其他 nlp 任务。我们的 encoder 是公开的。

二、 简介

单词的分布式表示被证明在计算机视觉和自然语言处理的多项任务中能够提供有效的特征。针对词嵌入的有用性以及如何通过学习得到词嵌入向量，有共识认为对于具有完整句子意义的陈述而言这两点尚不清楚。也就是说，如何在单个向量中捕获多个单词和短语之间的关系仍然

是需要解决的问题。

在这篇文章中，我们研究‘句子通用表达学习’这一任务，比如在大型语料库上训练得到的编码模型用于其他任务。为了得到编码模型需要解决两个问题:哪种神经网络架构最好；在对应任务上怎么训练这种神经网络。跟随目前在词嵌入学习方面的工作，目前的很多方法都是采用无监督学习来学习句子编码，比如 skip-thought 或者 FastSent。在这里，我们调查是否可以采用监督学习来学习句子表达，该灵感来源于计算机视觉领域的已经经验，计算机视觉领域很多任务使用在 imagenet 上预先训练的模型。我们对比了在多个有监督任务上的句子嵌入训练，发现自然语言推理任务生成的句子嵌入向量迁移到其他任务上有最好的结果。经过 NLI task 训练过的模型之所以能表现出优越的性能，是因为能够更深入的了解句子间的语义关系。

不同于计算机视觉领域，卷积神经网络占据主导地位，我们可以采用多种神经网络来编码句子。因此，我们调查句子编码结构对表达迁移性的影响，并对比 cnn/rnn/更简单的词组合方案。实验表明，bi-lstm 加上 max-pooling 这种网络结构是最好的。我们在更多迁移任务上测试句子

表达性能来捕捉更普遍和有用的信息。

三、 相关工作

介绍迁移学习，句子表达，数据等方面的工作

四、 方法

这项工作结合了两个研究方向，我们在以下描述。首先,我们解释一下如何使用NLI任务来训练使用SNLI任务的通用语句编码模型。我们找到目前常用的句子编码，并对这些模型结构进行了详细的调查。

3.1 自然语言推理任务

Text	Judgments	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction C C C C C	The man is sleeping
An older and younger man smiling.	neutral N N E N N	Two men are smiling and laughing at the cats playing on the floor.
A black race car starts up in front of a crowd of people.	contradiction C C C C C	A man is driving down a lonely road.
A soccer game with multiple males playing.	entailment E E E E E	Some men are playing a sport.
A smiling costumed woman is holding an umbrella.	neutral N N E C N	A happy woman in a fairy costume holds an umbrella.

SNLI dataset 是由 570,000 人工标注的英语句子对组成，每个句子对都有对应的标签。标签一共有三种，分别是 entailment，contradiction 和 neutral(蕴涵，矛盾和中性)。它捕捉到自然语言推理以及针对句子语义理解构造最大的高品质标签资源，自然语言推理这之前也被认为是识别文本认证（RTE)。。。。。

作者给出了在 SNLI 上训练 encoder 的两种方法：

i. encoder 对输入句子单独编码得到 representation，句子之

间没有交互；

ii. encoder对输入句子对联合编码(可以用到cross-features或者注意力机制)。

在文中，作者采取了第一种做法。

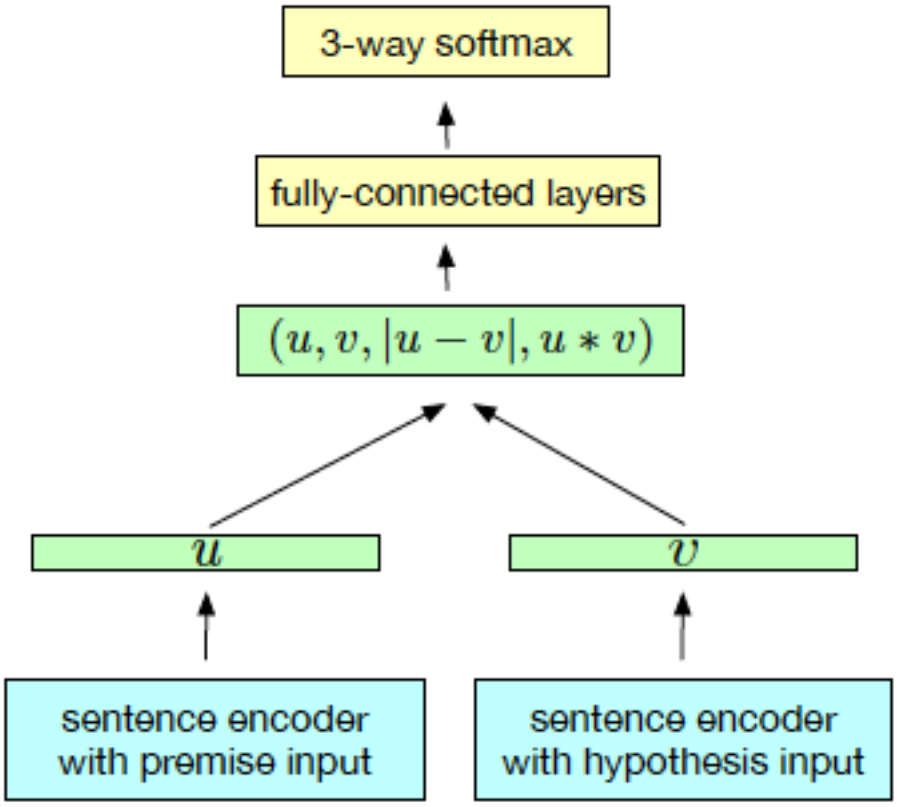

Figure 1: **Generic NLI training scheme.**

在上图中，展示了通用自然语言推理训练方案，SNLI 数据集中的 premise 和 hypothesis 经过 sentence encoder 后对应的向量表示分别为 u 和 v，然后作者用 3 种方法来抽取 u 和 v 的关系：

i. concatenation: 将 u 和 v 的表示首尾相连得到(u, v)

ii. element-wise product: 将 u 和 v 对应维度上的值相乘得到新的表示 u*v

iii. absolute element-wise difference: 将 u 和 v 对应维度上的值相减得到新的表示 |u-v|

最后将得到的表示送入一个 3 分类的分类器，分类器由多个全连接层和一个 softmax 层组成，最终得到输入 premise 和 hypothesis 的标签的概率分布。

3.2 句子编码结构

目前，有多种多样的神经网络能将句子编码成固定大小的向量表示,并且也没有明确的研究支出哪一种编码方法最好。因此，作者选择了 7 种不同的 architectures：

1. standard recurrent encoders with LSTM

2. standard recurrent encoders with GRU

上述两种是基础的 recurrent encoder，在句子建模中通常将网络中的最后一个隐藏状态作为 sentence representation;

3. conncatenation of last hidden states of forward and backward GRU

这种方法是将单向的网络变成了双向的网络，然后用将前向和后向的最后一个状态进行连接，得到句子向量;

4. Bi-directional LSTMs (BiLSTM) with mean pooling

5. Bi-directional LSTMs (BiLSTM) with max pooling

这两种方法使用了双向LSTM结合一个pooling层的方法来获取句子表示，具体公式如下：

$$
\begin{aligned}
\overrightarrow{h_t} &= \overrightarrow{\mathrm{LSTM}}_t(w_1, \ldots, w_T) \\
\overleftarrow{h_t} &= \overleftarrow{\mathrm{LSTM}}_t(w_1, \ldots, w_T) \\
h_t &= [\overrightarrow{h_t}, \overleftarrow{h_t}]
\end{aligned}
$$

按照上述公式得到每个时刻 t 的隐藏状态 ht 后，经过一个 max/mean pooling 得到最终的句子表示。其中，max/mean pooling 的意思是将每个时刻 t 下 ht 对应维度上的值进行比较，max 表示去对应维度上最大的值，mean 表示将该维度上所有值进行加和平均。网络模型如下：

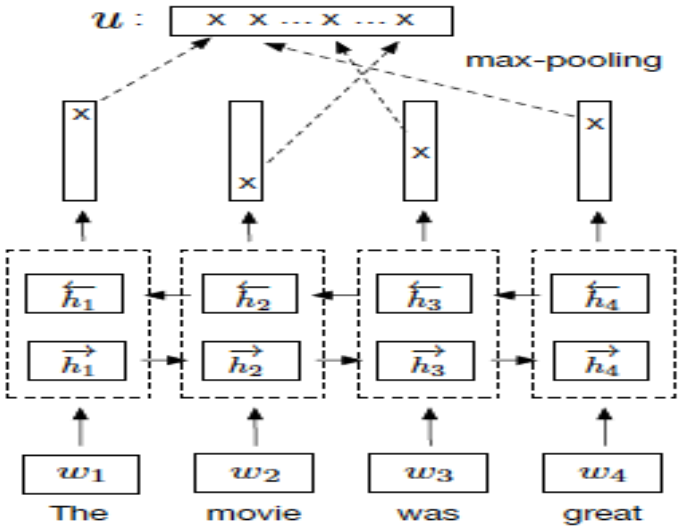

Figure 2: **Bi-LSTM max-pooling network.**

6. self-attentive network

这个网络在双向 LSTM 的基础上加入了 attention 机制,具体网络结构如下：

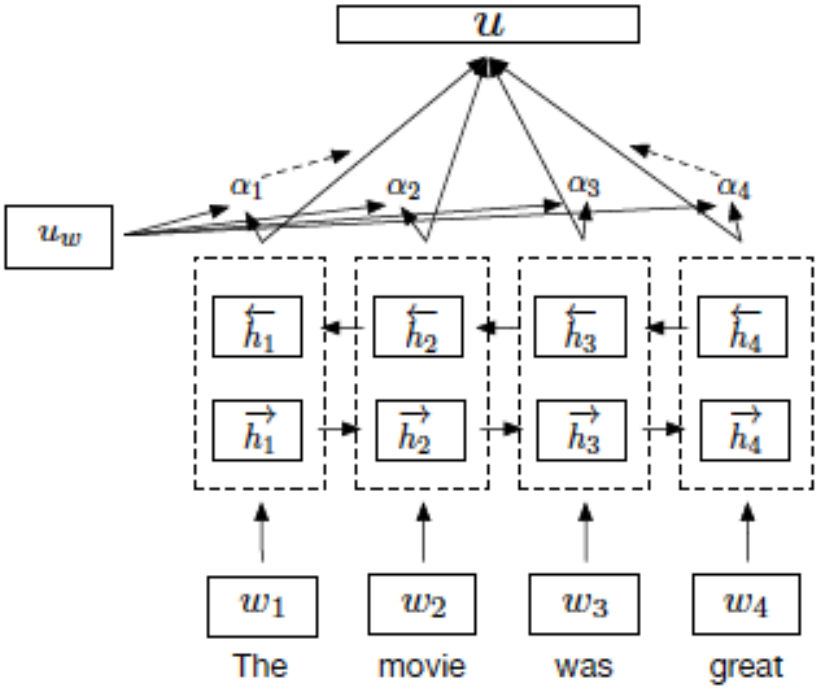

Figure 3: Inner Attention network architecture.

BiLSTM 的输出首先经过一次线性变换和一次非线性 tanh 变换，变成想要的形状，然后计算每个隐藏状态对应的注意力权重，最后加权求和得到最终表示。

$$
\begin{aligned}
\bar{h}_i &= \tanh(W h_i + b_w) \\
\alpha_i &= \frac{e^{\bar{h}_i^T u_w}}{\sum_i e^{\bar{h}_i^T u_w}} \\
u &= \sum_t \alpha_i h_i
\end{aligned}
$$

7. hierarchical convolutional networks

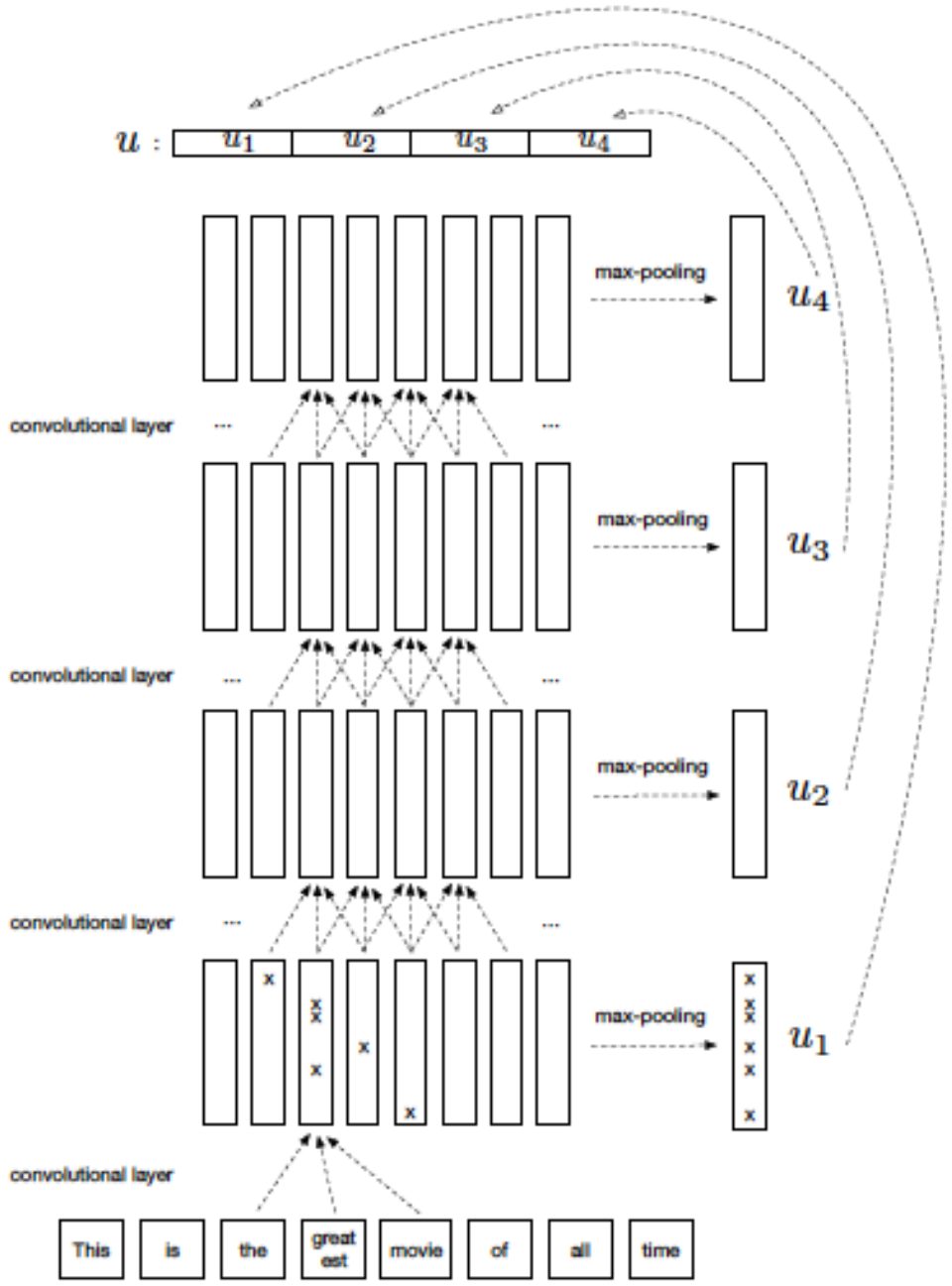

Figure 4: **Hierarchical ConvNet architecture.**

通过多层卷积神经网络可以对输入句子进行不同层级的抽象，在每一层作者通过 max-pooling 得到一个表示 ui，最终将这些

表示连接得到 u=[u1,u2,...,un]

3.3 训练细节

数据集：Standford Natural Language Inference dataset

参数更新方法：SGD 随机梯度下降

学习率：0.1

学习率衰减：0.99，即每个 epoch 的学习率是上一个 epoch 的 0.99

mini-batches 大小：64

训练停止条件：学习率小于 10^-5

分类器：多层感知机，隐藏层节点数为 512

词向量：300D Glove vectors trained on Common Crawl 840B

五、 对句子表示的评估

我们的目标是获取通用语句嵌入，捕获对广泛的任务有用的通用信息。为了评估这些表示的质量，我们将它们用作 12 个转移任务中的特征。我们在本节介绍我们的句子嵌入评估程序。我们构建了一个句子评估工具，对本文提到的所有任务进行自动评估。该工具使用 Adam 训练逻辑回归分类器，批量大小为 64。

主要涉及了如下任务：

Binary and multi-class classification

我们使用了一系列句子二分类任务。我们生成句子向量并在之上训练逻辑回归模型。线性分类器比 MLP 要求更少的参数，因此适合于小数据集，也特别适合转移学习。我们在验证集上利用网格搜索调整带 12 惩罚的逻辑回归模型。

Entailment and semantic relatedness

同上

STS14 - Semantic Textual Similarity

Paraphrase detection

Caption-Image retrieval

六、 实验结果

在这个部分，在迁移任务的准确率上我们提到了“微观”和“宏观”平均值：“宏观”对应测试集整体准确率，“微观”指的是测试集准确率的加权和，权重是测试集样本数。

5.1 结构影响

1、模型

2、嵌入大小

5.2 任务迁移

1、与 skipthought 的对比

2、NLI 作为监督训练集

3、对 SICK 任务进行域适应

4、图像字幕检索结果

5、MultiGenre NL

七、结论

这篇文章在 12 个不同的任务上对从有标签数据集（SNLI）训练得到的句子向量进行研究。结果表明,在自然语言推理(NLI)任务中训练得到的模型的性能要优于一些从其他有监督任务或者无监督条件下学习得到的模型。并且通过对不同编码器的比较，作者发现 BiLSTM with max pooling 是其中最优秀的生成句子表示的方法，胜过了 SkipThought vectors。