基于生成式预训练提升语言理解

Alec Radford Karthik Narasimhan Tim Salimans Ilya Sutskever
OpenAI OpenAI OpenAI OpenAI
alec@openai.com karthikn@openai.com tim@openai.com ilyasu@openai.com

Natural language understanding comprises a wide range of diverse tasks such as textual entailment, question answering, semantic similarity assessment, and document classification. Although large unlabeled text corpora are abundant, labeled data for learning these specific tasks is scarce, making it challenging for discriminatively trained models to perform adequately. We demonstrate that large gains on these tasks can be realized by generative pre-training of a language model on a diverse corpus of unlabeled text, followed by discriminative fine-tuning on each specific task.

自然语言理解包含很多任务，比如文本蕴含、问答、语义相似性评估、文档分类。尽管大型未标记预料库很丰富，然而学习这些任务的标记数据却很少，这样的话，判别模型就难以表现完美。我们发现基于大量未标记语料进行生成式训练语言模型训练，然后基于与训练模型进行微调，在这些任务上取得了巨大的进步。

In contrast to previous approaches, we make use of task-aware input transformations during fine-tuning to achieve effective transfer while requiring minimal changes to the model architecture. We demonstrate the effectiveness of our approach on a wide range of benchmarks for natural language understanding.

与其它的方法相比，在微调时，我们使用了任务感知的输入变换，来达到有效的变换，同时对网络架构不需要做大的调整。我们在广泛的自然语言理解的任务上，证明了我们的方法是有效的。

Our general task-agnostic model outperforms discriminatively trained models that use architectures specifically crafted for each task, significantly improving upon the state of the art in 9 out of the 12 tasks studied. For instance, we achieve absolute improvements of 8.9% on commonsense reasoning (Stories Cloze Test), 5.7% on question answering (RACE), and 1.5% on textual entailment (MultiNLI).

我们的通用模型超越了特定任务特定网络架构的模型，在12项任务中，有9项任务有了显著改进。例如，我们在常识推理（故事完形填空测试）中有8.9%的绝对改进，在问答任务（RACE）中有5.7%的改进，在文本蕴含（MultiNLI）中有1.5%的改进。

1. 引言

The ability to learn effectively from raw text is crucial to alleviating the dependence on supervised learning in natural language processing (NLP). Most deep learning methods require substantial amounts of manually labeled data, which restricts their applicability in many domains that suffer from a dearth of annotated resources [61].

在NLP中，从原始文本中有效得学习，对于减少对监督学习的依赖至关重要。大多数深度学习方法需要大量的人工标注数据，这限制了它们在缺乏标注资源的领域应用。

In these situations, models that can leverage linguistic information from unlabeled data provide a valuable alternative to gathering more annotation, which can be time-consuming and expensive. Further, even in cases where considerable supervision is available, learning good representations in an unsupervised fashion can provide a significant performance boost. The most compelling evidence for this so far has been the extensive use of pretrained word embeddings [10, 39, 42] to improve performance on a range of NLP tasks [8, 11, 26, 45].

在这些情况下，可以从未标注的数据中收集语言信息的模型，是一个有价值的替代方案，因为收集标注信息既耗时，又昂贵。此外，即使在有相当多的监督数据的情况下，以无监督的方式学习也会有显著的效果提升。目前，最令人信服的证据是预训练词嵌入，它可以在许多NLP任务中提升效果。

Leveraging more than word-level information from unlabeled text, however, is challenging for two main reasons. First, it is unclear what type of optimization objectives are most effective at learning text representations that are useful for transfer. Recent research has looked at various objectives such as language modeling [44], machine translation [38], and discourse coherence [22], with each method outperforming the others on different tasks.1 Second, there is no consensus on the most effective way to transfer these learned representations to the target task. Existing techniques involve a combination of making task-specific changes to the model architecture [43, 44], using intricate learning schemes [21] and adding auxiliary learning objectives [50]. These uncertainties have made it difficult to develop effective semi-supervised learning approaches for language processing.

然而，从未标注数据中收集词级别以上的信息是有挑战的，这有两个原因。第一个，不清楚到底是哪种目标函数对文本表示（用于迁移学习）是有效的。最近的研究有：语言模型、机器翻译、语言连贯性，每种方法都在不同的任务上优于其它的方法。第二个，如何将学习到的表示迁移到目标任务上，还没有达成共识。现有的技术是同时使用复杂技术针对特定任务更改网络架构和添加辅助学习任务。这些不确定性，使得开发有效得半监督学习模型变得困难。

In this paper, we explore a semi-supervised approach for language understanding tasks using a combination of unsupervised pre-training and supervised fine-tuning. Our goal is to learn a universal representation that transfers with little adaptation to a wide range of tasks. We assume access to a large corpus of unlabeled text and several datasets with manually annotated training examples (target tasks). Our setup does not require these target tasks to be in the same domain as the unlabeled corpus. We employ a two-stage training procedure. First, we use a language modeling objective on the unlabeled data to learn the initial parameters of a neural network model. Subsequently, we adapt these parameters to a target task using the corresponding supervised objective.

在这篇文章，我们针对语言理解探索了一个半监督学习方法，它是无监督的与训练和有监督的微调组合而成。我们的目标是学习一个通用的表示，基于此，不用有什么改动就可以应用到各种任务中。我

们使用了大量的未标记语料，和一些用于目标任务的标记数据。我们也不要求目标任务和未标记预料在同一领域。我们使用了两阶段训练过程。首先，我们用语言模型目标函数在未标记数据上学习了一个神经网络模型参数。随后，我们使用相应的监督目标函数将这些参数应用到目标任务上。

For our model architecture, we use the Transformer [62], which has been shown to perform strongly onvarious tasks such as machine translation [62], document generation [34], and syntactic parsing [29].This model choice provides us with a more structured memory for handling long-term dependencies in text, compared to alternatives like recurrent networks, resulting in robust transfer performance across diverse tasks. During transfer, we utilize task-specific input adaptations derived from traversal-style approaches [52], which process structured text input as a single contiguous sequence of tokens. As we demonstrate in our experiments, these adaptations enable us to fine-tune effectively with minimal changes to the architecture of the pre-trained model.

我们的模型架构选择的是Transformer，Transformer已经证实了在机器翻译、文档生成和语法解析方面有很强的表现。这个模型更适合处理长文本依赖，与RNN网络相比，在很多任务上表现出了强大的迁移性能。在迁移过程中，我们将结构化文本输入当作一个连续的token串来处理。正如我们在实验中所证明，这样处理可以使我们微调更高效，同时对预训练模型改动最少。

We evaluate our approach on four types of language understanding tasks - natural language inference,question answering, semantic similarity, and text classification. Our general task-agnostic model outperforms discriminatively trained models that employ architectures specifically crafted for each task, significantly improving upon the state of the art in 9 out of the 12 tasks studied. For instance,we achieve absolute improvements of 8.9% on common sense reasoning (Stories Cloze Test) [40],5.7% on question answering (RACE) [30], 1.5% on textual entailment (MultiNLI) [66] and 5.5% on the recently introduced GLUE multi-task benchmark [64]. We also analyzed zero-shot behaviors of the pre-trained model on four different settings and demonstrate that it acquires useful linguistic knowledge for downstream tasks.

我们在四种语言理解模型上进行了评估：语言推理、问答、语义相似度和文本分类。我们的通用的与任务无关的模型表现超越了为每个任务单独设计结构的判别式模型，在12个研究任务中有9个达到了最先进的水平。例如，我们在普通的语义推理（故事完形填空）上提升了8.9%的绝对水平，在问答（RACE）上提升了5.7%的绝对水平，在文本蕴含（MultiNLI）上提升了1.5%的绝对水平，在最近新提出的GLUE基准上提升了5.5%的绝对水平。我们还分析了预训练模型在四各不同设置下的零样本行为，并证明它获得了对下游任务有用的语言知识。

2. 相关工作

Semi-supervised learning for NLP Our work broadly falls under the category of semi-supervised learning for natural language. This paradigm has attracted significant interest, with applications to tasks like sequence labeling [24, 33, 57] or text classification [41, 70]. The earliest approaches used unlabeled data to compute word-level or phrase-level statistics, which were then used as features in a supervised model [33]. Over the last few years, researchers have demonstrated the benefits of using word embeddings [11, 39, 42], which are trained on unlabeled corpora, to improve performance on avariety of tasks [8, 11, 26, 45]. These approaches, however, mainly transfer word-level information,whereas we aim to capture higher-level semantics.

NLP的半监督学习

我们的工作大致属于自然语言的半监督学习范畴。这种学习方式引起了我们极大的兴趣，它可以应用序列标注、文本分类等任务。最早的方法先计算未标注数据的词级别和短语级的统计量，然后将它们作为特征送入监督学习模型。在过去的几年，研究者们发现使用词嵌入的好处，它们是在未标注的语料上进行训练，然后可以提升一系列任务的性能。然而，这些方法主要是迁移了词级的信息，我们的主要目标是获取更高级别的语义。

Recent approaches have investigated learning and utilizing more than word-level semantics from
unlabeled data. Phrase-level or sentence-level embeddings, which can be trained using an unlabeled
corpus, have been used to encode text into suitable vector representations for various target tasks [28,
32, 1, 36, 22, 12, 56, 31].

最近有人研究了从未标记数据中学习和使用超越词级语义的方法。从未标记语料中训练短语级和句子级的词嵌入，记经用来对文本编码成合适的向量表示，并用于各种目标任务中。

Unsupervised pre-training Unsupervised pre-training is a special case of semi-supervised learning where the goal is to find a good initialization point instead of modifying the supervised learning objective. Early works explored the use of the technique in image classification [20, 49, 63] and regression tasks [3]. Subsequent research [15] demonstrated that pre-training acts as a regularization scheme, enabling better generalization in deep neural networks. In recent work, the method has been used to help train deep neural networks on various tasks like image classification [69], speech recognition [68], entity disambiguation [17] and machine translation [48].

无监督预训练

无监督预训练是半监督学习的一种特殊情况，其目标是找到一个良好的初始化点，而不是修改监督学习目标。早期的工作探索了该技术在图像分类和回归任务中的应用。后期的研究表明预训练可以作为一种正则化的方案，使深度神经网络有更好的泛化性。最近，该方法已被用于帮助训练深度神经网络，比如图像分类、语音识别、实体消歧和机器翻译。

The closest line of work to ours involves pre-training a neural network using a language modeling
objective and then fine-tuning it on a target task with supervision. Dai et al. [13] and Howard and

Ruder [21] follow this method to improve text classification. However, although the pre-training phase helps capture some linguistic information, their usage of LSTM models restricts their prediction ability to a short range. In contrast, our choice of transformer networks allows us to capture longer range linguistic structure, as demonstrated in our experiments. Further, we also demonstrate the effectiveness of our model on a wider range of tasks including natural language inference, paraphrase detection and story completion. Other approaches [43, 44, 38] use hidden representations from a pre-trained language or machine translation model as auxiliary features while training a supervised model on the target task. This involves a substantial amount of new parameters for each separate target task, whereas we require minimal changes to our model architecture during transfer.

与我们工作最接近的研究是：先使用语言模型目标函数预训练一个神经网络，然后用监督的方法在目标任务上微调。Dai et al、Howard和Ruder使用这个方法提升了文本分类效果，尽管预训练的短语帮助获取了语言信息，然而，他们使用的LSTM模型限制了输入长度。相反， 我们的transformer网络允许输入的更长，这一点已在我们实验证实。此外，我们也证明了我们的模型在一系列任务上的有效性，这些任务包括语言推理、释义检测和故事创作。共它的一些方法有从语言模型或机器翻译模型中使用隐藏层表示来作为辅助特征，然后在目标任务上进行监督学习，这个对于每个单独的目标任务是需要大量的新参数。而我们在做迁移的时候只需要很小的改变。

Auxiliary training objectives Adding auxiliary unsupervised training objectives is an alternative form of semi-supervised learning. Early work by Collobert and Weston [10] used a wide variety of auxiliary NLP tasks such as POS tagging, chunking, named entity recognition, and language modeling to improve semantic role labeling. More recently, Rei [50] added an auxiliary language modeling objective to their target task objective and demonstrated performance gains on sequence labeling tasks. Our experiments also use an auxiliary objective, but as we show, unsupervised pre-training already learns several linguistic aspects relevant to target tasks.

辅助训练目标

添加辅助无监督训练目标也是半监督学习的一种选择。早期由Collobert和Weston进行的研究，使用了很多辅助的NLP任务：词性标注、分块、实体识别和语言模型来提升语义角色标注。最近，Rei添加了辅助语言模型目标函数到目标任务的目标函上，而且证明了在序列标注任务上有了很好的提升。我们的实验也使用了一个辅助目标函数，但是正如我们展示的，无监督的预训练已经学习到了与目标任务相关的几个语言部分。

3 算法框架

Our training procedure consists of two stages. The first stage is learning a high-capacity language
model on a large corpus of text. This is followed by a fine-tuning stage, where we adapt the model to
a discriminative task with labeled data.

我们的训练过程包括两步。第一步是在大量的语料上学习一个高性能的语言模型，第二步是微调，也就是将第一步学习到的模型应用到带有标注数据的判别任务上。

3.1 无监督预训练

Given an unsupervised corpus of tokens U = {u1, . . . , un}, we use a standard language modeling
objective to maximize the following likelihood:

where k is the size of the context window, and the conditional probability P is modeled using a neural network with parameters Θ. These parameters are trained using stochastic gradient descent [51].In our experiments, we use a multi-layer Transformer decoder [34] for the language model, which is a variant of the transformer [62]. This model applies a multi-headed self-attention operation over the input context tokens followed by position-wise feedforward layers to produce an output distribution over target tokens:

where U = (u−k, . . . , u−1) is the context vector of tokens, n is the number of layers, We is the token embedding matrix, and Wp is the position embedding matrix.

给定一个无监督的token串U={u1, ..., un}，我们使用一个标准的语言模型目标函数来最大化下面的似然函数：

这里k是内容窗口，条件概率P是带参数θ的神经网络模型计算得到。这些参数使用随机梯度下降算法训练而得到。在我们的实验中，我们使用了一个多层的transformer解码器来表示这个语言模型，这是transformer的一个变体。这个模型在输入文本串上使用了一个多头自注意力操作，然后用了一个基于位置的前馈神经网络层，最后产生一个与目标tokens对应的输出分布。

这里U = (u−k, . . . , u−1) 是输入token向量，n是层数，We是token嵌入矩阵，Wp是位置嵌入矩阵。

3.2 有监督微调

After training the model with the objective in Eq. 1, we adapt the parameters to the supervised targettask. We assume a labeled dataset C, where each instance consists of a sequence of input tokens,x1, . . . , xm, along with a label y. The inputs are passed through our pre-trained model to obtain the final transformer block' s activation hml, which is then fed into an added linear output layer withparameters Wy to predict y:

This gives us the following objective to maximize:

We additionally found that including language modeling as an auxiliary objective to the fine-tuning helped learning by (a) improving generalization of the supervised model, and (b) accelerating convergence. This is in line with prior work [50, 43], who also observed improved performance with such an auxiliary objective. Specifically, we optimize the following objective (with weight λ):

Figure 1: (left) Transformer architecture and training objectives used in this work. (right) Input transformations for fine-tuning on different tasks. We convert all structured inputs into token sequences to be processed by our pre-trained model, followed by a linear+softmax layer.
通过目标函数L1训练出一个模型后，我们将参数用于监督目标任务上。我们假设带标签的数据为C，每个实例包一串输入的tokens，x1,...,xm，以及输入标签y。输入通过训练模型最后得到一个transformer的输出hml，然后将该输出送入线性输出层，乘以Wy得到y的预测值。

基于此我们得到要最大化的目标函数：

我们还发现将语言模型作为辅助目标函数加到微调任务上，可以提升监督模型的泛化能力，而且可以加速收敛。这与之前的有个工作是相似的，他们发现使用这样的辅助目标函数可以提升表现。因此，我们将目标函数定义为以下形式：

图 1: (左) Transformer架构和本工作中作使用得训练目标。 (右) 在不同任务上进行微调时得输入变换。我们将结构化的输入转换成token序列，然后送入预训练模型，然后在下游接一个线性层和softmax层。

3.3 任务输入转换

For some tasks, like text classification, we can directly fine-tune our model as described above.Certain other tasks, like question answering or textual entailment, have structured inputs such as ordered sentence pairs, or triplets of document, question, and answers. Since our pre-trained model was trained on contiguous sequences of text, we require some modifications to apply it to these tasks.Previous work proposed learning task specific architectures on top of transferred representations [44].Such an approach re-introduces a significant amount of task-specific customization and does not use transfer learning for these additional architectural components. Instead, we use a traversal-style approach [52], where we convert structured inputs into an ordered sequence that our pre-trained model can process. These input transformations allow us to avoid making extensive changes to the architecture across tasks. We provide a brief description of these input transformations below and Figure 1 provides a visual illustration. All transformations include adding randomly initialized start and end tokens (hsi, hei).

对于一些任务，比如文本分类，正如以上所说，我们可以直接微调。一些其它任务，比如问答或者文本蕴含，是有结构化的输入的，比如排好序的句子对，文档、问题、答案三元组。因为我们的与训练模型是基于一些连续的文本串训练的，我们需要一些改变来适应这些任务。先前的工作提出了基于迁移表征的学习任务特定架构[44]。这种方法重新引入了大量特定于任务的自定义，并且没有对这些额外的体系结构组件使用迁移学习。相反，我们使用遍历式方法，将结构化的输入转换成有序的序列，以便于与训练模型可以处理。这种输入转换使我们免去了基于特定任务的昂贵变换。下面我们列出了这些输入变换的简要说明，图1也将此说明可视化了。所有的转换都包括随机初始化的开始和结束符。

Textual entailment For entailment tasks, we concatenate the premise p and hypothesis h token sequences, with a delimiter token ($) in between.

文本蕴含 对于文本蕴含任务，我们我们用符号（$）将前提p和建设h的token序列连接起来。

Similarity For similarity tasks, there is no inherent ordering of the two sentences being compared.To reflect this, we modify the input sequence to contain both possible sentence orderings (with a delimiter in between) and process each independently to produce two sequence representations hml which are added element-wise before being fed into the linear output layer.

文本相似 对于文本相似任务，对于要比较的两个句子没有固定的顺序。为了反映这一点，我们改变了输入序列以包含两种可能可顺序，然后分别对两个序列表示生成表示，然后按元素相加，再送入到线性输出层。

Question Answering and Commonsense Reasoning For these tasks, we are given a context document z, a question q, and a set of possible answers {ak}. We concatenate the document context and question with each possible answer, adding a delimiter token in between to get [z; q; $; ak]. Each of these sequences are processed independently with our model and then normalized via a softmax layer to produce an output distribution over possible answers.

问答和推理 对于这些任务，我们面对的是一个文档z，一个问题q，和一系列可能的答案{ak}。我们将文档、问题和每个答案用分隔符连在一起，用【z; q; $; ak】表示。每个序列用模型独立进行表示，然后通过softmax层进行标准化，得到输出的分布，对应于可能的答案。

4 实验

4.1 设置

Unsupervised pre-training We use the BooksCorpus dataset [71] for training the language model.It contains over 7,000 unique unpublished books from a variety of genres including Adventure,Fantasy, and Romance. Crucially, it contains long stretches of contiguous text, which allows the generative model to learn to condition on long-range information. An alternative dataset, the 1B Word Benchmark, which is used by a similar approach, ELMo [44], is approximately the same size but is shuffled at a sentence level - destroying long-range structure. Our language model achieves a very low token level perplexity of 18.4 on this corpus.

无监督预训练 我们使用BookCorpus数据集来训练语言模型。该数据集包含7000本未出版的书籍，有探险、幻想和浪漫。关键在于，它包含长文本，这使得生成模型可以基于长段的信息进行学习。还有一个数据集是“1B Word Benchmark”，Elmo也使用了这个数据集，二者的大小相似，但是Elmo在句子级上进行了洗牌，这破坏了长文本结构。我们的语言模型在这个语料上实现了很低的token困惑度，仅为18.4。

Model specifications Our model largely follows the original transformer work [62]. We trained a 12-layer decoder-only transformer with masked self-attention heads (768 dimensional states and 12 attention heads). For the position-wise feed-forward networks, we used 3072 dimensional inner states.We used the Adam optimization scheme [27] with a max learning rate of 2.5e-4. The learning rate was increased linearly from zero over the first 2000 updates and annealed to 0 using a cosine schedule.We train for 100 epochs on minibatches of 64 randomly sampled, contiguous sequences of 512 tokens.Since layernorm [2] is used extensively throughout the model, a simple weight initialization of N(0, 0.02) was sufficient. We used a bytepair encoding (BPE) vocabulary with 40,000 merges [53]and residual, embedding, and attention dropouts with a rate of 0.1 for regularization. We also employed a modified version of L2 regularization proposed in [37], with w = 0.01 on all non bias orgain weights. For the activation function, we used the Gaussian Error Linear Unit (GELU) [18]. We used learned position embeddings instead of the sinusoidal version proposed in the original work.We use the ftfy library2 to clean the raw text in BooksCorpus, standardize some punctuation and whitespace, and use the spaCy tokenizer.3

模型细节

我们的模型在很大程度上遵循了原始transformer的工作。我们训练了一个12层的带掩码的多头自注意力机制的Transformer Decoder（768维、12个头）。对于位置编码的前馈神经网络，我们使用了3072维内部状态。我们使用了Adam优化器，其中的最大学习率为2.5e-4。学习率在前2000次更新中从0开始线性增加，然后使用余弦调度退火到0。我们在64个随机抽样的连续序列的小批量上进行了100轮训练，每个序列包含512个tokens。由于层标准化在模型中广泛使用，因此使用一个正态分布N(0, 0.02)生成的简单权重来作为初始化权重就足够了。我们使用了40,000个合并的字节对编码（BPE）词汇表[53]，并使用了残差、嵌入和注意力丢失，其正则化率为0.1。我们还使用了【37】所提到的改进版本的L2正则，作用在了初始的非偏置项的权重w=0.01上。对于激活函数，我们用的是高斯误差线性单元（GELU）。我们使用了学习的位置编码，而不是原始工作中提到的正弦编码。我们使用ftfy库来清理BookCorpus的原始文本，标准化一些标点符号和空格，还使用了分词器spaCy。

Fine-tuning details Unless specified, we reuse the hyperparameter settings from unsupervised pre-training. We add dropout to the classifier with a rate of 0.1. For most tasks, we use a learning rate of 6.25e-5 and a batchsize of 32. Our model finetunes quickly and 3 epochs of training was sufficient for most cases. We use a linear learning rate decay schedule with warmup over 0.2% of training. λ was set to 0.5.

微调细节

除非有特殊情况，我们一般是复用监督预训练的超参。我们将丢失率为0.1的dropout添加到分类器上。对于大多数任务，我们使用学习率为6.25e-5，batchsize为32。我们的模型微调的非常快，3轮训练对于大多数案例来说足够了。我们使用线性学习率衰减计划，在0.2%的训练中进行预热。 λ设置为0.5。

4.2 监督微调

We perform experiments on a variety of supervised tasks including natural language inference,question answering, semantic similarity, and text classification. Some of these tasks are available as part of the recently released GLUE multi-task benchmark [64], which we make use of. Figure 1 provides an overview of all the tasks and datasets.

我们在各种监督任务上进行实验，包括自然语言推理、问答、语义相似和文本分类。其中一些任务是最近发布的GLUE多任务基准测试中的。图1展示了所有的任务和数据集。

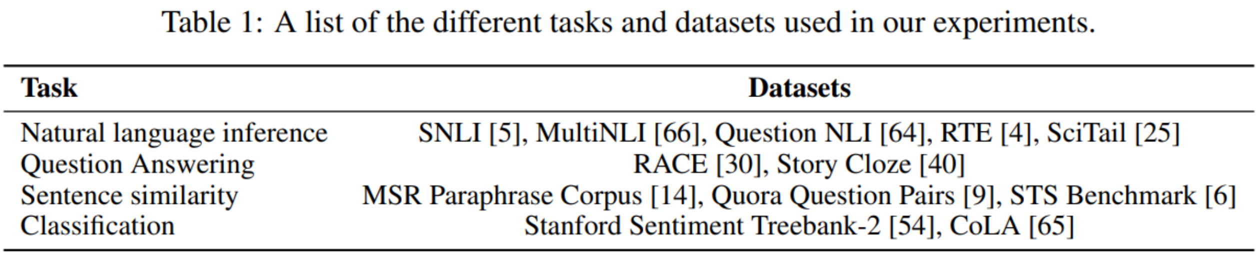

Table 1: A list of the different tasks and datasets used in our experiments.

Task	Datasets
Natural language inference	SNLI [5], MultiNLI [66], Question NLI [64], RTE [4], SciTail [25]
Question Answering	RACE [30], Story Cloze [40]
Sentence similarity	MSR Paraphrase Corpus [14], Quora Question Pairs [9], STS Benchmark [6]
Classification	Stanford Sentiment Treebank-2 [54], CoLA [65]

Natural Language Inference The task of natural language inference (NLI), also known as recognizing textual entailment, involves reading a pair of sentences and judging the relationship between them from one of entailment, contradiction or neutral. Although there has been a lot of recent interest [58, 35, 44], the task remains challenging due to the presence of a wide variety of phenomena like lexical entailment, coreference, and lexical and syntactic ambiguity. We evaluate on five datasets with diverse sources, including image captions (SNLI), transcribed speech, popular fiction, and government reports (MNLI), Wikipedia articles (QNLI), science exams (SciTail) or news articles (RTE).

自然语言推理

自然语言推理任务（NLI）,或者称之为文本蕴含，是阅读一对句子然后判断他们之间的关系，关系有蕴含、矛盾和中性。尽管近年来引起了很多关注[58，35，44]，但由于存在各种现象，如词汇蕴涵、指代和词汇和句法歧义，该任务仍然具有挑战性。我们在五个数据集上进行了评估，包括图像标题（SNLI）、转录语音、流行小说、政府报告（MNLI）、维基百科文章（QNLI）、科学考试（SciTail）和新闻文章（RTE）。

Table 2 details various results on the different NLI tasks for our model and previous state-of-the-art approaches. Our method significantly outperforms the baselines on four of the five datasets, achieving absolute improvements of upto 1.5% on MNLI, 5% on SciTail, 5.8% on

QNLI and 0.6% on SNLI over the previous best results. This demonstrates our model’s ability to better reason over multiple sentences, and handle aspects of linguistic ambiguity. On RTE, one of the smaller datasets we evaluate on (2490 examples), we achieve an accuracy of 56%, which is below the 61.7% reported by a multi-task biLSTM model. Given the strong performance of our approach on larger NLI datasets, it is likely our model will benefit from multi-task training as well but we have not explored this currently.

表2详细说明了我们的模型和之前最新先进模型在不同NLI任务上的结果。我们的方法在五个数据集上都击败了之前的baseline，相较于之前最好的结果在MNLI上提升了1.5%绝对百分点、在SciTail上提升了5%绝对百分点、在QNLI上提升了5.8%绝对百分点、在SNLI上提升了0.6%绝对百分点。这证明了，我们的模型有能力更好地推理多句子，更好地处理语言模型糊性。在RTE这个较小的数据集上（2490个样本）我们进行了评估，我们获得了56%的准确率，低于多任务biLSTM模型所报告的61.7%。鉴于我们模型在大型NLI数据集上的强劲表现，我们模型很有可能出会在多任务训练上获益，但我们目前尚未这么做。

表2：NLI上的实验结果，对比了我们的模型和最先进的方法。5x代表了5个模型的集成。所有数据集使用准确率作为评估标准。

Question answering and commonsense reasoning Another task that requires aspects of single and multi-sentence reasoning is question answering. We use the recently released RACE dataset [30],consisting of English passages with associated questions from middle and high school exams. This corpus has been shown to contain more reasoning type questions than other datasets like CNN [19] orSQuaD [47], providing the perfect evaluation for our model which is trained to handle long-range contexts. In addition, we evaluate on the Story Cloze Test [40], which involves selecting the correctending to multi-sentence stories from two options. On these tasks, our model again outperforms the previous best results by significant margins - up to 8.9% on Story Cloze, and 5.7% overall on RACE.This demonstrates the ability of our model to handle long-range contexts effectively.

问答和常识推理

另一个需要单句或多句推理的任务是问答。我们使用了最近发布的RACE数据集，这是来自中学和高中的考试题目，包含了英文段落和相关问题。这个语料已被证明比其实数据集CNN和SQuaD包含更多的推理类型的问题，为我们的模型提供了完美的评估，我们的模型经过训练可以处理长文本。此外，我们还在Story Cloze Test上进行了评估，这涉及从两个选项中选择多句故事的正确结尾。在这些任务上，我们的模型再一次大幅超越了以往的最好结果，在Story Cloze上超越了8.9%，在RACE上超越了5.7%。这证实了我们模型在高效处理长文本上的能力。

表3：问答和常识推理上的结果，对比我们模型和当前最先进的方法。9x代表9个模型的集成。

Semantic Similarity Semantic similarity (or paraphrase detection) tasks involve predicting whether two sentences are semantically equivalent or not. The challenges lie in recognizing rephrasing of concepts, understanding negation, and handling syntactic ambiguity. We use three datasets for this task – the Microsoft Paraphrase corpus (MRPC) [14] (collected from news sources), the QuoraQuestion Pairs (QQP) dataset [9], and the Semantic Textual Similarity benchmark (STS-B) [6].We obtain state-of-the-art results on two of the three semantic similarity tasks (Table 4) with a 1 point absolute gain on STS-B. The performance delta on QQP is significant, with a 4.2% absolute improvement over Single-task BiLSTM + ELMo + Attn.

语义相似

语义相似（或段落检测）任务就是预测两个句子是否在语义上相等。挑战在于识别概念改述、理解否定和处理句法歧义。我们使用了三个数据集来完成这项任务——Microsoft Paraphrase corpus (MRPC) （从新闻来源收集），Quora Question Pairs (QQP) 数据集和Semantic Textual Similarity benchmark (STS-B)。这三个语义相似任务，我们有两个得到了优异的结果，在STS-B上提升了1个绝对百分点。QQP上的表现有显著提升，比单任务BiLSTM + ELMo + Attn提升了4.2%个绝对百分点。

Classification Finally, we also evaluate on two different text classification tasks. The Corpus of Linguistic Acceptability (CoLA) [65] contains expert judgements on whether a sentence is grammatical or not, and tests the innate linguistic bias of trained models. The Stanford Sentiment Treebank (SST-2) [54], on the other hand, is a standard binary classification task. Our model obtains an score of 45.4 on CoLA, which is an especially big jump over the previous best result of 35.0,showcasing the innate linguistic bias learned by our model. The model also achieves 91.3% accuracy on SST-2, which is competitive with the state-of-the-art results. We also achieve an overall score of 72.8 on the GLUE benchmark, which is significantly better than the previous best of 68.9.

分类问题

最后，我们还在两个分类任务上做了评估。语言可接受性的语料（CoLA）包含了专家判断一个句子是否符合语法，并测试了训练模型的固有语言偏见。另一方面，斯坦幅情感树（SST-2）是一个标准的二分类任务。我们的模型在CoLA上得到了45.4分，比之前最好的结果35.0有了巨大的进步，展示了我们模型学习到的固有语言偏见。这个模型还在SST-2上得到了91.3%的准确率，可以比肩当前最好的结果。我们也在GLUE基线测试得到了72.8分的总分，这比之前的68.9分提升了多少。

表4：语义相似和分类结果，对比了我们模型和当前最好的方法。此表的所有评估是用GLUE基线来完成的。（mc=Mathews correlation, acc=Accuracy,pc=Pearson correlation）

Overall, our approach achieves new state-of-the-art results in 9 out of the 12 datasets we evaluateon, outperforming ensembles in many cases. Our results also indicate that our approach works well across datasets of different sizes, from smaller datasets such as STS-B (≈5.7k training examples) –to the largest one – SNLI (≈550k training examples).

总之，在12个数据集中，我们的方法有9个获得了最新的优异结果，在很多场景击败了集成模型。我们的结果还证明了我们的方法在各种规模的数据集上表现良好，从小的数据集STS-B（≈5.7k训练样本）到最大的SNLI（≈550k训练样本）。

5 分析

Impact of number of layers transferred We observed the impact of transferring a variable number of layers from unsupervised pre-training to the supervised target task. Figure 2(left) illustrates the performance of our approach on MultiNLI and RACE as a function of the number of layers transferred.We observe the standard result that transferring embeddings improves performance and that each transformer layer provides further benefits up to 9% for full transfer on MultiNLI. This indicates that each layer in the pre-trained model contains useful functionality for solving target tasks.

层数的影响

我们观察了可变层数对无监督与训练到有监督目标任务的影响。图2（左）展示了将层数作为变量时，我们的方法在MultiNLI和RACE上的表现。我们观察到：将嵌入进行迁移提高了性能，而且每增加一个transformer层会多增加最多9%的性能提升。这表明了预训练模型的每一层都包含解决目标任务的有用信息。

Zero-shot Behaviors We’d like to better understand why language model pre-training of transform-ers is effective. A hypothesis is that the underlying generative model learns to perform many of the tasks we evaluate on in order to improve its language modeling capability and that the more structured attentional memory of the transformer assists in transfer compared to LSTMs. We designed a series of heuristic solutions that use the underlying generative model to perform tasks without supervised finetuning. We visualize the effectiveness of these heuristic solutions over the course of generative pre-training in Fig 2(right). We observe the performance of these heuristics is stable and steadily increases over training suggesting that generative pretraining supports the learning of a wide variety of task relevant functionality. We also observe the LSTM exhibits higher variance in its zero-shot performance suggesting that the inductive bias of the Transformer architecture assists in transfer.

零样本学习行为

我们想更好地理解为什么预训练的transfomrer是有效的。一个假设是底层的生成模型学习了很多我们要评估的任务知识，以此提升了语言模型能力。另外，相比于LSTM系列，更多的transformer注意力记忆有助于迁移。我们设计了一系列启发式的解决方案，在没有监督微调的情况下，使用底层的生成模型去评估任务。在图2（右）上，我们可展示了这些启发式解决方案在生成与训练过程中的有效性。我们观察到这些启发式方案的表现稳定，而且表现随着训练次数的增加而稳定提升，这表明生成式预训练支持一系列任务相关的功能学习。我们还观察到LSTM在零样本学习上展示了更高的方差，这表明Transformer架构的感应偏置有助于迁移。

图2：（左）预训练模型的层数对下游任务RACE和MultiNLI的影响。（右）图展示了零样本学习在不同任务的表现随着与训练次数的提升而提升。每个任务的表现平均来说介于随机猜测和当前最好的模型表现之间。

For CoLA (linguistic acceptability), examples are scored as the average token log-probability the generative model assigns and predictions are made by thresholding. For SST-2 (sentiment analysis),we append the token very to each example and restrict the language model’s output distribution to only the words positive and negative and guess the token it assigns higher probability to as the prediction.For RACE (question answering), we pick the answer the generative model assigns the highest averagetoken log-probability when conditioned on the document and question. For DPRD [46] (winogradschemas), we replace the definite pronoun with the two possible referrents and predict the resolutionthat the generative model assigns higher average token log-probability to the rest of the sequenceafter the substitution.（暂不好翻译）

Ablation studies We perform three different ablation studies (Table 5). First, we examine the performance of our method without the auxiliary LM objective during fine-tuning. We observe that the auxiliary objective helps on the NLI tasks and QQP. Overall, the trend suggests that larger datasets benefit from the auxiliary objective but smaller datasets do not. Second, we analyze the effect of theTransformer by comparing it with a single layer 2048 unit LSTM using the same framework. We observe a 5.6 average score drop when using the LSTM instead of the Transformer. The LSTM only outperforms the Transformer on one dataset – MRPC. Finally, we also compare with our transformer architecture directly trained on supervised target tasks, without pre-training. We observe that the lack of pre-training hurts performance across all the tasks, resulting in a 14.8% decrease compared to our full model。

消融实验

我们尝试了三个不同的消融实验（表5）。首先，我们在微调时不用辅助的LM目标函数的情况下检查模型表现。我们发现辅助目标有助于NLI和QQP任务。整体来说，这个变化趋势表明越大的数据集越能从辅助目标上得到好处，而越小的数据集却不能。其次，我们用相同的网络架构对比了单层2048单元的LSTM和Transformer。我们观察到用LSTM来代替Transformer的话平均分要下降5.6分。LSTM仅仅在MRPC上击败了Transformer。最后，我们对比了直接在目标监督任务上训练，而不使用预训练。我们发现不用预训练的会影响所有任务的性能，跟我们的全模型相比下降了14.8%。

表5：不同任务上的几种消融实验分析。平均分是没有权重的。（mc= Mathews correlation, acc=Accuracy, pc=Pearson correlation）

6 结论

We introduced a framework for achieving strong natural language understanding with a single task-agnostic model through generative pre-training and discriminative fine-tuning. By pre-training on a diverse corpus with long stretches of contiguous text our model acquires significant world knowledge and ability to process long-range dependencies which are then successfully transferred to solving discriminative tasks such as question answering, semantic similarity assessment, entailment determination, and text classification, improving the state of the art on 9 of the 12 datasets we study. Using unsupervised (pre-)training to boost performance on discriminative tasks has long been an important goal of Machine Learning research. Our work suggests that achieving significant performance gains is indeed possible, and offers hints as to what models (Transformers) and data sets(text with long range dependencies) work best with this approach. We hope that this will help enable new research into unsupervised learning, for both natural language understanding and other domains,further improving our understanding of how and when unsupervised learning works.

我们介绍了一个框架，使用任务无关的预训练和微调模型来实现强大的自然语言理解。通过在不同的长文本语料上进行预训练，我们模型获得了重要的世界级的知识和能力去处理长文本依赖，然后成功迁移到其它任务，如问答、语义相似判别、蕴含识别和文本分类，在12个数据集上有9个效果有重大提升。使用无监督预训练去提升判别任务的效果长期以来是机器学习的重要目标。我们的工作表明取得重要的效果提升的确可行，并且提供了哪种模型（Transformer）和数据（具有长距离依赖性的文本）更适合这种方法。我们希望这将会助力无监督学习方面新的研究，不管是自然语言理解还是其它领域，进一步提升我们对无监督工作的理解。

References

[1] S. Arora, Y. Liang, and T. Ma. A simple but tough-to-beat baseline for sentence embeddings. 2016.

[2] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. arXiv preprint arXiv:1607.06450, 2016.

[3] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. Greedy layer-wise training of deep networks. In Advances in neural information processing systems, pages 153–160, 2007.

[4] L. Bentivogli, P. Clark, I. Dagan, and D. Giampiccolo. The fifth pascal recognizing textual entailment challenge. In TAC, 2009.

[5] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning. A large annotated corpus for learning natural language inference. EMNLP, 2015.

[6] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. arXiv preprint arXiv:1708.00055, 2017.

[7] S. Chaturvedi, H. Peng, and D. Roth. Story comprehension for predicting what happens next. In Proceedings

of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1603–1614, 2017.

[8] D. Chen and C. Manning. A fast and accurate dependency parser using neural networks. In Proceedings
of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 740–750,
2014.

[9] Z. Chen, H. Zhang, X. Zhang, and L. Zhao. Quora question pairs. https://data.quora.com/First-QuoraDataset-Release-Question-Pairs, 2018.

[10] R. Collobert and J. Weston. A unified architecture for natural language processing: Deep neural networks
with multitask learning. In Proceedings of the 25th international conference on Machine learning, pages
160–167. ACM, 2008.

[11] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing
(almost) from scratch. Journal of Machine Learning Research, 12(Aug):2493–2537, 2011.

[12] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes. Supervised learning of universal sentence
representations from natural language inference data. EMNLP, 2017.

[13] A. M. Dai and Q. V. Le. Semi-supervised sequence learning. In Advances in Neural Information Processing
Systems, pages 3079–3087, 2015.

[14] W. B. Dolan and C. Brockett. Automatically constructing a corpus of sentential paraphrases. In Proceedings
of the Third International Workshop on Paraphrasing (IWP2005), 2005.

[15] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio. Why does unsupervised
pre-training help deep learning? Journal of Machine Learning Research, 11(Feb):625–660, 2010.

[16] S. Gray, A. Radford, and K. P. Diederik. Gpu kernels for block-sparse weights. 2017.

[17] Z. He, S. Liu, M. Li, M. Zhou, L. Zhang, and H. Wang. Learning entity representation for entity disambiguation. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics
(Volume 2: Short Papers), volume 2, pages 30–34, 2013.

[18] D. Hendrycks and K. Gimpel. Bridging nonlinearities and stochastic regularizers with gaussian error linear

units. arXiv preprint arXiv:1606.08415, 2016.

[19] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom. Teaching machines to read and comprehend. In Advances in Neural Information Processing Systems, pages 1693–1701, 2015.

[20] G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. Neural computation, 18(7):1527–1554, 2006.

[21] J. Howard and S. Ruder. Universal language model fine-tuning for text classification. Association for Computational Linguistics (ACL), 2018.

[22] Y. Jernite, S. R. Bowman, and D. Sontag. Discourse-based objectives for fast unsupervised sentence representation learning. arXiv preprint arXiv:1705.00557, 2017.

[23] Y. Ji and J. Eisenstein. Discriminative improvements to distributional sentence similarity. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 891–896, 2013.

[24] F. Jiao, S. Wang, C.-H. Lee, R. Greiner, and D. Schuurmans. Semi-supervised conditional random fields for improved sequence segmentation and labeling. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, pages 209–216. Association for Computational Linguistics, 2006.

[25] T. Khot, A. Sabharwal, and P. Clark. Scitail: A textual entailment dataset from science question answering. In Proceedings of AAAI, 2018.

[26] Y. Kim. Convolutional neural networks for sentence classification. EMNLP, 2014.

[27] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[28] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler. Skip-thought vectors. In Advances in neural information processing systems, pages 3294–3302, 2015.

[29] N. Kitaev and D. Klein. Constituency parsing with a self-attentive encoder. ACL, 2018.

[30] G. Lai, Q. Xie, H. Liu, Y. Yang, and E. Hovy. Race: Large-scale reading comprehension dataset from examinations. EMNLP, 2017.

[31] G. Lample, L. Denoyer, and M. Ranzato. Unsupervised machine translation using monolingual corpora only. ICLR, 2018.

[32] Q. Le and T. Mikolov. Distributed representations of sentences and documents. In International Conference on Machine Learning, pages 1188–1196, 2014.

[33] P. Liang. Semi-supervised learning for natural language. PhD thesis, Massachusetts Institute of Technology, 2005.

[34] P. J. Liu, M. Saleh, E. Pot, B. Goodrich, R. Sepassi, L. Kaiser, and N. Shazeer. Generating wikipedia by summarizing long sequences. ICLR, 2018.

[35] X. Liu, K. Duh, and J. Gao. Stochastic answer networks for natural language inference. arXiv preprint arXiv:1804.07888, 2018.

[36] L. Logeswaran and H. Lee. An efficient framework for learning sentence representations. ICLR, 2018.

[37] I. Loshchilov and F. Hutter. Fixing weight decay regularization in adam. arXiv preprint arXiv:1711.05101, 2017.

[38] B. McCann, J. Bradbury, C. Xiong, and R. Socher. Learned in translation: Contextualized word vectors. In Advances in Neural Information Processing Systems, pages 6297–6308, 2017.

[39] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, pages 3111–3119, 2013.

[40] N. Mostafazadeh, M. Roth, A. Louis, N. Chambers, and J. Allen. Lsdsem 2017 shared task: The story cloze test. In Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics, pages 46–51, 2017.

[41] K. Nigam, A. McCallum, and T. Mitchell. Semi-supervised text classification using em. Semi-Supervised Learning, pages 33–56, 2006.

[42] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 1532–1543, 2014.

[43] M. E. Peters, W. Ammar, C. Bhagavatula, and R. Power. Semi-supervised sequence tagging with bidirectional language models. ACL, 2017.

[44] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. NAACL, 2018.

[45] Y. Qi, D. S. Sachan, M. Felix, S. J. Padmanabhan, and G. Neubig. When and why are pre-trained word embeddings useful for neural machine translation? NAACL, 2018.

[46] A. Rahman and V. Ng. Resolving complex cases of definite pronouns: the winograd schema challenge. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 777–789. Association for Computational Linguistics, 2012.

[47] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. Squad: 100,000+ questions for machine comprehension of text. EMNLP, 2016.

[48] P. Ramachandran, P. J. Liu, and Q. V. Le. Unsupervised pretraining for sequence to sequence learning. arXiv preprint arXiv:1611.02683, 2016.

[49] M. Ranzato, C. Poultney, S. Chopra, and Y. LeCun. Efficient learning of sparse representations with an energy-based model. In Advances in neural information processing systems, pages 1137–1144, 2007.

[50] M. Rei. Semi-supervised multitask learning for sequence labeling. ACL, 2017.

[51] H. Robbins and S. Monro. A stochastic approximation method. The annals of mathematical statistics, pages 400–407, 1951.

[52] T. Rocktäschel, E. Grefenstette, K. M. Hermann, T. Kocisk ˇ y, and P. Blunsom. Reasoning about entailment ` with neural attention. arXiv preprint arXiv:1509.06664, 2015.

[53] R. Sennrich, B. Haddow, and A. Birch. Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909, 2015.

[54] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 conference on empirical methods in natural language processing, pages 1631–1642, 2013.

[55] S. Srinivasan, R. Arora, and M. Riedl. A simple and effective approach to the story cloze test. arXiv preprint arXiv:1803.05547, 2018.

[56] S. Subramanian, A. Trischler, Y. Bengio, and C. J. Pal. Learning general purpose distributed sentence representations via large scale multi-task learning. arXiv preprint arXiv:1804.00079, 2018.

[57] J. Suzuki and H. Isozaki. Semi-supervised sequential labeling and segmentation using giga-word scale unlabeled data. Proceedings of ACL-08: HLT, pages 665–673, 2008.

[58] Y. Tay, L. A. Tuan, and S. C. Hui. A compare-propagate architecture with alignment factorization for natural language inference. arXiv preprint arXiv:1801.00102, 2017.

[59] Y. Tay, L. A. Tuan, and S. C. Hui. Multi-range reasoning for machine comprehension. arXiv preprint arXiv:1803.09074, 2018.

[60] J. Tian, Z. Zhou, M. Lan, and Y. Wu. Ecnu at semeval-2017 task 1: Leverage kernel-based traditional nlp features and neural networks to build a universal model for multilingual and cross-lingual semantic textual similarity. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pages 191–197, 2017.

[61] Y. Tsvetkov. Opportunities and challenges in working with low-resource languages. CMU, 2017.

[62] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin.

Attention is all you need. In Advances in Neural Information Processing Systems, pages 6000–6010, 2017.

[63] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In Proceedings of the 25th international conference on Machine learning, pages 1096–1103. ACM, 2008.

[64] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461, 2018.

[65] A. Warstadt, A. Singh, and S. R. Bowman. Corpus of linguistic acceptability. http://nyu-mll.github.io/cola, 2018.

[66] A. Williams, N. Nangia, and S. R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. NAACL, 2018.

[67] Y. Xu, J. Liu, J. Gao, Y. Shen, and X. Liu. Towards human-level machine reading comprehension: Reasoning and inference with multiple strategies. arXiv preprint arXiv:1711.04964, 2017.

[68] D. Yu, L. Deng, and G. Dahl. Roles of pre-training and fine-tuning in context-dependent dbn-hmms for real-world speech recognition. In Proc. NIPS Workshop on Deep Learning and Unsupervised Feature Learning, 2010.

[69] R. Zhang, P. Isola, and A. A. Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In CVPR, volume 1, page 6, 2017.

[70] X. Zhu. Semi-supervised learning literature survey. 2005.

[71] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In Proceedings of the IEEE international conference on computer vision, pages 19–27, 2015.