**Improving Language Understanding by Generative Pre-Training**

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever

**Abstract**

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. 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. 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).

자연어 이해는 텍스트 수반, 질문 답변, 의미 유사성 평가 및 문서 분류와 같은 다양한 작업으로 구성됩니다. 레이블이 지정되지 않은 큰 텍스트 말뭉치는 풍부하지만 이러한 특정 작업을 학습하기 위한 레이블이 지정된 데이터는 드물기 때문에 판별식으로 훈련된 모델이 적절하게 수행하기 어렵습니다. 우리는 레이블이 지정되지 않은 텍스트의 다양한 코퍼스에 대한 언어 모델의 생성적 사전 훈련과 각 특정 작업에 대한 차별적 미세 조정을 통해 이러한 작업에 대한 큰 이득을 실현할 수 있음을 보여줍니다. 이전 접근 방식과 달리 모델 아키텍처에 대한 변경을 최소화하면서 효과적인 전송을 달성하기 위해 미세 조정 중에 작업 인식 입력 변환을 사용합니다. 우리는 자연어 이해를 위한 광범위한 벤치마크에 대한 접근 방식의 효율성을 보여줍니다. 우리의 일반 작업 불가지론 모델은 각 작업에 대해 특별히 제작된 아키텍처를 사용하는 차별적으로 훈련된 모델보다 성능이 뛰어나 연구된 12개 작업 중 9개 작업에서 최신 기술을 크게 향상시킵니다. 예를 들어 상식 추론(Stories Cloze Test)에서 8.9%, 질문 답변(RACE)에서 5.7%, 텍스트 포함(MultiNLI)에서 1.5%의 절대적인 향상을 달성했습니다.

**1 Introduction**

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]. 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].

원시 텍스트에서 효과적으로 학습하는 능력은 NLP(자연어 처리)에서 지도 학습에 대한 의존도를 줄이는 데 중요합니다. 대부분의 딥 러닝 방법은 수동으로 레이블이 지정된 상당한 양의 데이터를 필요로 하므로 주석이 달린 리소스가 부족한 많은 영역에서 적용 가능성이 제한됩니다[61]. 이러한 상황에서 레이블이 지정되지 않은 데이터의 언어 정보를 활용할 수 있는 모델은 시간과 비용이 많이 소요될 수 있는 더 많은 주석을 수집하는 것에 대한 귀중한 대안을 제공합니다. 또한 상당한 감독이 가능한 경우에도 감독되지 않은 방식으로 좋은 표현을 학습하면 상당한 성능 향상을 제공할 수 있습니다. 지금까지 이것에 대한 가장 강력한 증거는 다양한 NLP 작업[8, 11, 26, 45]에서 성능을 향상시키기 위해 사전 훈련된 단어 임베딩[10, 39, 42]을 광범위하게 사용했다는 것입니다.

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.

그러나 레이블이 지정되지 않은 텍스트에서 단어 수준 이상의 정보를 활용하는 것은 두 가지 주요 이유로 어렵습니다. 첫째, 어떤 유형의 최적화 목표가 전송에 유용한 텍스트 표현을 학습하는 데 가장 효과적인지 명확하지 않습니다. 최근 연구에서는 언어 모델링[44], 기계 번역[38], 담화 일관성[22]과 같은 다양한 목표를 살펴보았으며, 각 방법은 다른 작업에서 다른 방법보다 성능이 뛰어납니다.1 둘째, 가장 효과적인 방법에 대한 합의가 없습니다. 이러한 학습된 표현을 대상 작업으로 전송합니다. 기존 기술은 모델 아키텍처에 대한 작업별 변경[43, 44], 복잡한 학습 계획 사용[21] 및 보조 학습 목표 추가[50]의 조합을 포함합니다. 이러한 불확실성으로 인해 언어 처리를 위한 효과적인 반 지도 학습 접근 방식을 개발하기가 어렵습니다.

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.

이 논문에서 우리는 감독되지 않은 사전 훈련과 감독된 미세 조정의 조합을 사용하여 언어 이해 작업에 대한 반 감독 접근 방식을 탐구합니다. 우리의 목표는 광범위한 작업에 거의 적응하지 않고 전환하는 보편적인 표현을 배우는 것입니다. 레이블이 지정되지 않은 텍스트의 대규모 코퍼스와 수동으로 주석이 달린 훈련 예제(대상 작업)가 있는 여러 데이터 세트에 대한 액세스를 가정합니다. 우리의 설정에서는 이러한 대상 작업이 레이블이 지정되지 않은 말뭉치와 동일한 도메인에 있을 필요가 없습니다. 우리는 2단계 교육 절차를 사용합니다. 먼저 레이블이 지정되지 않은 데이터에 언어 모델링 목표를 사용하여 신경망 모델의 초기 매개변수를 학습합니다. 그 후, 해당 감독 목표를 사용하여 이러한 매개변수를 대상 작업에 적용합니다.

For our model architecture, we use the Transformer [62], which has been shown to perform strongly on various 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.

우리의 모델 아키텍처에는 기계 번역[62], 문서 생성[34], 구문 분석[29]과 같은 다양한 작업을 강력하게 수행하는 것으로 입증된 Transformer[62]를 사용합니다. 이 모델 선택은 순환 네트워크와 같은 대안과 비교하여 텍스트의 장기 종속성을 처리하기 위한 보다 구조화된 메모리를 제공하여 다양한 작업에서 강력한 전송 성능을 제공합니다. 전송하는 동안 구조화된 텍스트 입력을 단일 연속 토큰 시퀀스로 처리하는 순회 스타일 접근 방식에서 파생된 작업별 입력 적응을 활용합니다. 실험에서 보여주듯이 이러한 적응을 통해 사전 훈련된 모델의 아키텍처를 최소한으로 변경하여 효과적으로 미세 조정할 수 있습니다.

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 commonsense 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.

우리는 네 가지 유형의 언어 이해 작업(자연어 추론, 질문 답변, 의미론적 유사성, 텍스트 분류)에 대한 접근 방식을 평가합니다. 우리의 일반 작업 불가지론 모델은 각 작업에 대해 특별히 제작된 아키텍처를 사용하는 차별적으로 훈련된 모델보다 성능이 뛰어나 연구된 12개 작업 중 9개 작업에서 최신 기술을 크게 향상시킵니다. 예를 들어 상식 추론(Stories Cloze Test)에서 8.9%, 질문 답변(RACE)[30]에서 5.7%, 텍스트 수반(MultiNLI)에서 1.5%, [66]에서 5.5%의 절대적인 향상을 달성했습니다. 최근 GLUE 다중 작업 벤치마크를 도입했습니다[64]. 또한 4가지 다른 설정에서 사전 훈련된 모델의 제로샷 동작을 분석하고 다운스트림 작업에 유용한 언어 지식을 획득함을 보여줍니다.

**2 Related Work**

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 a variety of tasks [8, 11, 26, 45]. These approaches, however, mainly transfer word-level information, whereas we aim to capture higher-level semantics.

NLP를 위한 준지도 학습 우리의 작업은 광범위하게 자연어에 대한 준지도 학습 범주에 속합니다. 이 패러다임은 시퀀스 라벨링 또는 텍스트 분류와 같은 작업에 적용하여 상당한 관심을 끌었습니다. 초기 접근 방식은 레이블이 지정되지 않은 데이터를 사용하여 단어 수준 또는 구문 수준 통계를 계산한 다음 지도 모델의 기능으로 사용했습니다. 지난 몇 년 동안 연구자들은 레이블이 없는 말뭉치에 대해 훈련된 단어 임베딩을 사용하여 다양한 작업의 성능을 향상시키는 이점을 입증했습니다. 그러나 이러한 접근 방식은 주로 단어 수준 정보를 전송하는 반면 우리는 더 높은 수준의 의미를 포착하는 것을 목표로 합니다.

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].

최근 접근 방식은 레이블이 지정되지 않은 데이터에서 단어 수준 의미론 이상의 학습 및 활용을 조사했습니다. 레이블이 지정되지 않은 말뭉치를 사용하여 학습할 수 있는 구문 수준 또는 문장 수준 임베딩은 텍스트를 다양한 대상 작업에 적합한 벡터 표현으로 인코딩하는 데 사용되었습니다.

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].

감독되지 않은 사전 훈련 감독되지 않은 사전 훈련은 감독 학습 목표를 수정하는 대신 좋은 초기화 지점을 찾는 것이 목표인 반 감독 학습의 특별한 경우입니다. 초기 작업은 이미지 분류 및 회귀 작업에서 이 기술의 사용을 탐구했습니다. 후속 연구는 사전 훈련이 정규화 방식으로 작용하여 심층 신경망에서 더 나은 일반화를 가능하게 함을 보여주었습니다. 최근 작업에서 이 방법은 이미지 분류, 음성 인식, 개체 명확화 및 기계 번역과 같은 다양한 작업에서 심층 신경망을 훈련시키는 데 사용되었습니다.

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 longerrange 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 2 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.

우리와 가장 가까운 작업 라인은 언어 모델링 목표를 사용하여 신경망을 사전 훈련한 다음 감독하에 대상 작업에서 미세 조정하는 것입니다. Dai et al. [13], Howard and Ruder [21]는 텍스트 분류를 개선하기 위해 이 방법을 따릅니다. 그러나 사전 훈련 단계가 일부 언어 정보를 캡처하는 데 도움이 되지만 LSTM 모델을 사용하면 예측 능력이 짧은 범위로 제한됩니다. 대조적으로, 변압기 네트워크를 선택하면 실험에서 입증된 것처럼 더 긴 범위의 언어 구조를 캡처할 수 있습니다. 또한 자연어 추론, 의역 감지 및 스토리 완성을 포함한 광범위한 작업에 대한 모델의 효율성도 보여줍니다. 다른 접근법[43, 44, 38]은 대상 작업에 대해 지도 모델을 훈련하는 동안 2개의 사전 훈련된 언어 또는 기계 번역 모델의 숨겨진 표현을 보조 기능으로 사용합니다. 여기에는 각각의 개별 대상 작업에 대한 상당한 양의 새 매개변수가 포함되지만 전송하는 동안 모델 아키텍처에 대한 최소한의 변경이 필요합니다.

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.

보조 교육 목표 보조 비지도 교육 목표를 추가하는 것은 반 지도 학습의 대안적인 형태입니다. Collobert와 Weston[10]의 초기 작업은 의미론적 역할 레이블링을 개선하기 위해 POS 태깅, 청킹, 명명된 엔터티 인식 및 언어 모델링과 같은 다양한 보조 NLP 작업을 사용했습니다. 보다 최근에 Rei는 목표 작업 목표에 보조 언어 모델링 목표를 추가하고 시퀀스 라벨링 작업에서 성능 향상을 시연했습니다. 우리의 실험은 또한 보조 목표를 사용하지만, 우리가 보여주듯이 감독되지 않은 사전 훈련은 이미 대상 작업과 관련된 여러 언어적 측면을 학습합니다.

**3 Framework**

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.

우리의 교육 절차는 두 단계로 구성됩니다. 첫 번째 단계는 대량의 텍스트에서 고용량 언어 모델을 학습하는 것입니다. 그 다음에는 레이블이 지정된 데이터를 사용하여 식별 작업에 모델을 적용하는 미세 조정 단계가 이어집니다.

**3.1 Unsupervised pre-training**

Given an unsupervised corpus of tokens we use a standard language modeling objective to maximize the following likelihood:

감독되지 않은 토큰 모음 이 주어지면 표준 언어 모델링 목표를 사용하여 다음 가능성을 최대화합니다.

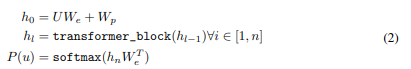


where k is the size of the context window, and the conditional probability P is modeled using a neural network with parameters Θ. These parameters are trained using stochastic gradient descent [51].

여기서 k는 컨텍스트 창의 크기이고 조건부 확률 P는 매개변수 가 있는 신경망을 사용하여 모델링됩니다. 이러한 매개변수는 확률적 경사하강법을 사용하여 훈련됩니다[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:

우리의 실험에서 우리는 변환기[62]의 변형인 언어 모델에 대해 다층 변환기 디코더[34]를 사용합니다. 이 모델은 대상 토큰에 대한 출력 분포를 생성하기 위해 위치별 피드포워드 레이어가 뒤따르는 입력 컨텍스트 토큰에 대해 다중 헤드 셀프 어텐션(multi-headed self-attention) 작업을 적용합니다.



where is the context vector of tokens, n is the number of layers, is the token embedding matrix, and is the position embedding matrix.

여기서 U=(u\_k,…,u\_1)은 토큰의 컨텍스트 벡터, n은 레이어 수, W\_e는 토큰 임베딩 행렬, W\_p는 위치 임베딩 행렬입니다.

**3.2 Supervised fine-tuning**

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

Eq.의 목적으로 모델을 훈련시킨 후 1, 매개변수를 감독 대상 작업에 적용합니다. 레이블이 지정된 데이터 세트 C를 가정합니다. 여기서 각 인스턴스는 레이블 y와 함께 일련의 입력 토큰 으로 구성됩니다. 입력은 사전 훈련된 모델을 통해 전달되어 최종 변환기 블록의 활성화 을 얻은 다음 매개변수 를 사용하여 추가된 선형 출력 레이어에 공급되어 y를 예측합니다.



This gives us the following objective to maximize:

이를 통해 다음과 같은 목표를 극대화할 수 있습니다.



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):

우리는 또한 미세 조정에 대한 보조 목표로 언어 모델링을 포함하는 것이 (a) 지도 모델의 일반화를 개선하고 (b) 수렴을 가속화함으로써 학습에 도움이 된다는 것을 발견했습니다. 이는 이러한 보조 목적으로 향상된 성능을 관찰한 이전 작업[50, 43]과 일치합니다. 특히 다음 목표를 최적화합니다(가중치 λ 사용).



Overall, the only extra parameters we require during fine-tuning are Wy, and embeddings for delimiter tokens (described below in Section 3.3).

전반적으로 미세 조정 중에 필요한 유일한 추가 매개변수는 Wy와 구분자 토큰에 대한 임베딩입니다(아래 섹션 3.3에서 설명).

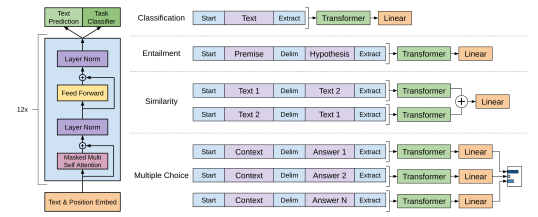


Figure 1: (left) Transformer architecture and training objectives used in this work. (right) Input transformations for fine-tuning on different tasks. We convert all structured inputs into token sequences to be processed by our pre-trained model, followed by a linear+softmax layer.

그림 1: (왼쪽) 이 작업에 사용된 Transformer 아키텍처 및 교육 목표. (오른쪽) 다양한 작업에 대한 미세 조정을 위한 입력 변환. 모든 구조화된 입력을 사전 훈련된 모델에 의해 처리될 토큰 시퀀스로 변환한 다음 linear+softmax 레이어가 뒤따릅니다.

**3.3 Task-specific input transformations**

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 (<s>,<e>).

텍스트 분류와 같은 일부 작업의 경우 위에서 설명한 대로 모델을 직접 미세 조정할 수 있습니다. 질문 응답 또는 텍스트 수반과 같은 특정 다른 작업에는 정렬된 문장 쌍 또는 문서, 질문 및 답변의 삼중항과 같은 구조화된 입력이 있습니다. 사전 훈련된 모델은 연속된 텍스트 시퀀스에 대해 훈련되었기 때문에 이러한 작업에 적용하려면 약간의 수정이 필요합니다. 이전 작업에서는 전송된 표현 위에 학습 작업별 아키텍처를 제안했습니다[44]. 이러한 접근 방식은 상당한 양의 작업별 사용자 지정을 다시 도입하고 이러한 추가 아키텍처 구성 요소에 대해 전이 학습을 사용하지 않습니다. 대신 우리는 구조화된 입력을 사전 훈련된 모델이 처리할 수 있는 순서화된 시퀀스로 변환하는 순회 스타일 접근 방식[52]을 사용합니다. 이러한 입력 변환을 통해 작업 전반에 걸쳐 아키텍처에 대한 광범위한 변경을 피할 수 있습니다. 아래에서 이러한 입력 변환에 대한 간략한 설명을 제공하고 그림 1은 시각적 설명을 제공합니다. 모든 변환에는 무작위로 초기화된 시작 및 종료 토큰(<s>,<e>) 추가가 포함됩니다.

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

텍스트 포함 포함 작업의 경우 구분 기호 토큰($)을 사용하여 전제 p와 가설 h 토큰 시퀀스를 연결합니다.

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 which are added element-wise before being fed into the linear output layer.

유사성 유사성 작업의 경우 비교되는 두 문장의 고유한 순서가 없습니다. 이를 반영하기 위해 입력 시퀀스를 수정하여 가능한 문장 순서(사이에 구분 기호 포함)를 모두 포함하고 선형 출력 레이어에 공급되기 전에 요소별로 추가되는 두 시퀀스 표현 을 생성하기 위해 각각을 독립적으로 처리합니다.

Question Answering and Commonsense Reasoning For these tasks, we are given a context document z, a question q, and a set of possible answers {}. We concatenate the document context and question with each possible answer, adding a delimiter token in between to get . 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.

질문 답변 및 상식적인 추론 이러한 작업을 위해 컨텍스트 문서 z, 질문 q 및 가능한 답변 세트 {}가 제공됩니다. 문서 컨텍스트와 질문을 가능한 각 답변과 연결하고 그 사이에 구분자 토큰을 추가하여 를 얻습니다. 이러한 각 시퀀스는 우리 모델과 독립적으로 처리된 다음 가능한 답변에 대한 출력 분포를 생성하기 위해 softmax 레이어를 통해 정규화됩니다.

**4 Experiments**

**4.1 Setup**

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.

비지도 사전 훈련 언어 모델 훈련을 위해 BooksCorpus 데이터 세트[71]를 사용합니다. 어드벤처, 판타지, 로맨스 등 다양한 장르의 7,000권이 넘는 미출간 도서가 포함되어 있습니다. 결정적으로, 여기에는 긴 연속 텍스트가 포함되어 있어 생성 모델이 장거리 정보를 조건으로 학습할 수 있습니다. 유사한 접근 방식인 ELMo[44]에서 사용하는 대체 데이터 세트인 1B Word Benchmark는 크기가 거의 같지만 문장 수준에서 뒤섞여서 장거리 구조를 파괴합니다. 우리의 언어 모델은 이 말뭉치에서 18.4의 매우 낮은 토큰 수준 혼란을 달성합니다.

텍스트이(가) 표시된 사진

자동 생성된 설명

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

표 1: 실험에 사용된 다양한 작업 및 데이터 세트의 목록입니다.

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 or gain 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

모델 사양 우리 모델은 원래 변압기 작업 [62]을 크게 따릅니다. 우리는 마스크된 자기 주의 헤드(768차원 상태 및 12주의 헤드)가 있는 12계층 디코더 전용 변환기를 훈련했습니다. 위치별 피드포워드 네트워크의 경우 3072 차원 내부 상태를 사용했습니다. 우리는 최대 학습률이 2.5e-4인 Adam 최적화 방식[27]을 사용했습니다. 학습률은 처음 2000개의 업데이트 동안 0에서 선형으로 증가했으며 코사인 일정을 사용하여 0으로 어닐링되었습니다. 무작위로 샘플링된 512개 토큰의 연속 시퀀스 64개로 구성된 미니배치에서 100개의 에포크를 훈련합니다. layernorm [2]는 모델 전체에 걸쳐 광범위하게 사용되기 때문에 N(0, 0.02)의 간단한 가중치 초기화로 충분했습니다. 우리는 정규화를 위해 40,000개의 병합[53]과 0.1의 비율로 잔차, 임베딩 및 주의 드롭아웃이 있는 바이트 쌍 인코딩(BPE) 어휘를 사용했습니다. 우리는 또한 모든 비편향 또는 이득 가중치에 대해 w = 0.01인 [37]에서 제안된 L2 정규화의 수정된 버전을 사용했습니다. 활성화 함수의 경우 GELU(Gaussian Error Linear Unit)[18]를 사용했습니다. 우리는 원래 작업에서 제안된 사인파 버전 대신 학습된 위치 임베딩을 사용했습니다. 우리는 ftfy 라이브러리2를 사용하여 BooksCorpus에서 원시 텍스트를 정리하고, 일부 구두점과 공백을 표준화하고, spaCy 토크나이저를 사용합니다.3

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.

세부 사항 미세 조정 지정되지 않는 한 감독되지 않은 사전 훈련의 하이퍼 매개 변수 설정을 재사용합니다. 0.1의 비율로 분류기에 드롭아웃을 추가합니다. 대부분의 작업에 대해 6.25e-5의 학습률과 32의 배치 크기를 사용합니다. 우리의 모델은 빠르게 미세 조정되며 대부분의 경우 3회의 훈련으로 충분합니다. 훈련의 0.2% 이상의 워밍업과 함께 선형 학습률 감소 일정을 사용합니다. λ는 0.5로 설정되었습니다.

**4.2 Supervised fine-tuning**

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.

우리는 자연어 추론, 질문 답변, 의미론적 유사성, 텍스트 분류를 포함한 다양한 지도 작업에 대한 실험을 수행합니다. 이러한 작업 중 일부는 최근 출시된 GLUE 다중 작업 벤치마크[64]의 일부로 사용할 수 있으며 이를 사용합니다. 그림 1은 모든 작업과 데이터 세트의 개요를 제공합니다.

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).

자연어 추론 자연어 추론(NLI) 작업은 텍스트 함축 인식으로도 알려져 있으며 한 쌍의 문장을 읽고 함축, 모순 또는 중립 중 하나로 두 문장 간의 관계를 판단하는 작업을 포함합니다. 최근 많은 관심이 있었지만[58, 35, 44], 어휘 수반, 상호 참조, 어휘 및 구문 모호성과 같은 다양한 현상의 존재로 인해 작업이 여전히 도전적입니다. 우리는 이미지 캡션(SNLI), 필사된 연설, 대중 소설 및 정부 보고서(MNLI), Wikipedia 기사(QNLI), 과학 시험(SciTail) 또는 뉴스 기사(RTE)를 포함하여 다양한 소스가 있는 5개의 데이터 세트를 평가합니다.

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.

표 2는 우리 모델과 이전의 최첨단 접근 방식에 대한 다양한 NLI 작업에 대한 다양한 결과를 자세히 설명합니다. 우리의 방법은 5개 데이터 세트 중 4개 데이터 세트의 기준선을 크게 능가하여 이전 최고의 결과에 비해 MNLI에서 최대 1.5%, SciTail에서 5%, QNLI에서 5.8%, SNLI에서 0.6%의 절대 개선을 달성했습니다. 이것은 우리 모델이 여러 문장에 대해 더 잘 추론하고 언어적 모호성의 측면을 처리하는 능력을 보여줍니다. 우리가 평가하는 더 작은 데이터 세트 중 하나인 RTE(2490개의 예)에서 우리는 56%의 정확도를 달성했으며, 이는 아래 표 2에 나와 있습니다. . 우리의 방법은 5개 데이터 세트 중 4개 데이터 세트의 기준선을 크게 능가하여 이전 최고의 결과에 비해 MNLI에서 최대 1.5%, SciTail에서 5%, QNLI에서 5.8%, SNLI에서 0.6%의 절대 개선을 달성했습니다. 이것은 우리 모델이 여러 문장에 대해 더 잘 추론하고 언어적 모호성의 측면을 처리하는 능력을 보여줍니다. 우리가 평가하는 더 작은 데이터 세트 중 하나인 RTE(2490개의 예)에서 56%의 정확도를 달성했으며 이는 다중 작업 biLSTM 모델에서 보고된 61.7%보다 낮습니다. 더 큰 NLI 데이터 세트에 대한 접근 방식의 강력한 성능을 감안할 때 우리 모델은 다중 작업 교육의 이점도 얻을 수 있지만 현재 이를 탐색하지 않았습니다. 다중 작업 biLSTM 모델에서 보고된 61.7%입니다. 더 큰 NLI 데이터 세트에 대한 접근 방식의 강력한 성능을 감안할 때 우리 모델은 다중 작업 교육의 이점도 얻을 수 있지만 현재 이를 탐색하지 않았습니다.

테이블이(가) 표시된 사진

자동 생성된 설명

Table 2: Experimental results on natural language inference tasks, comparing our model with current state-of-the-art methods. 5x indicates an ensemble of 5 models. All datasets use accuracy as the evaluation metric.

표 2: 자연어 추론 작업에 대한 실험 결과, 우리 모델을 현재 최첨단 방법과 비교합니다. 5x는 5개 모델의 앙상블을 나타냅니다. 모든 데이터 세트는 정확도를 평가 지표로 사용합니다.

테이블이(가) 표시된 사진

자동 생성된 설명

Table 3: Results on question answering and commonsense reasoning, comparing our model with current state-of-the-art methods.. 9x means an ensemble of 9 models.

표 3: 우리 모델을 현재 최첨단 방법과 비교한 질문 답변 및 상식 추론 결과. 9x는 9개 모델의 앙상블을 의미합니다.

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 that other datasets like CNN [19] or SQuaD [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 correct ending 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.

질문 답변 및 상식적인 추론 단일 및 다중 문장 추론의 측면을 요구하는 또 다른 작업은 질문 답변입니다. 우리는 최근에 발표된 RACE 데이터 세트[30]를 사용하며, 중학교 및 고등학교 시험의 관련 질문과 함께 영어 구절로 구성됩니다. 이 말뭉치는 CNN[19] 또는 SQuaD[47]와 같은 다른 데이터 세트보다 더 많은 추론 유형 질문을 포함하는 것으로 나타났으며, 장거리 컨텍스트를 처리하도록 훈련된 우리 모델에 대한 완벽한 평가를 제공합니다. 또한 두 가지 옵션에서 여러 문장으로 된 이야기의 올바른 결말을 선택하는 Story Cloze Test[40]를 평가합니다. 이러한 작업에서 우리 모델은 Story Cloze에서 최대 8.9%, RACE에서 전체적으로 5.7%의 상당한 마진으로 이전 최고의 결과를 다시 능가합니다. 이것은 우리 모델이 장거리 컨텍스트를 효과적으로 처리하는 능력을 보여줍니다.

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 Quora Question 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.

의미론적 유사성 의미론적 유사성(또는 의역 검색) 작업에는 두 문장이 의미적으로 동일한지 여부를 예측하는 작업이 포함됩니다. 문제는 개념의 변형을 인식하고, 부정을 이해하고, 구문적 모호성을 처리하는 데 있습니다. 이 작업을 위해 Microsoft Paraphrase corpus(MRPC)[14](뉴스 소스에서 수집), Quora Question Pairs(QQP) 데이터 세트[9], Semantic Textualsimilarity 벤치마크(STS-B)[6]의 세 가지 데이터 세트를 사용합니다. ]. 우리는 STS-B에서 1포인트 절대 이득으로 3개의 의미론적 유사성 작업 중 2개(표 4)에 대한 최신 결과를 얻습니다. QQP의 성능 델타는 단일 작업 BiLSTM + ELMo + Attn에 비해 절대적으로 4.2% 향상되어 중요합니다.

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.

분류 마지막으로 두 가지 다른 텍스트 분류 작업에 대해서도 평가합니다. 언어적 수용 가능성의 코퍼스(CoLA)[65]는 문장이 문법적인지 아닌지에 대한 전문가 판단을 포함하고 훈련된 모델의 타고난 언어적 편향을 테스트합니다. 반면 Stanford Sentiment Treebank(SST-2)[54]는 표준 이진 분류 작업입니다. 우리 모델은 CoLA에서 45.4점을 얻었습니다. 이는 이전의 최고 결과인 35.0점보다 특히 크게 높아서 우리 모델에서 학습한 타고난 언어적 편향을 보여줍니다. 또한 이 모델은 SST-2에서 91.3%의 정확도를 달성하여 최첨단 결과와 경쟁합니다. 우리는 또한 GLUE 벤치마크에서 전체 점수 72.8을 달성했는데, 이는 이전 최고 점수인 68.9보다 훨씬 높습니다.

테이블이(가) 표시된 사진

자동 생성된 설명

Table 4: Semantic similarity and classification results, comparing our model with current state-of-theart methods. All task evaluations in this table were done using the GLUE benchmark. (mc= Mathews correlation, acc=Accuracy, pc=Pearson correlation)

표 4: 의미론적 유사성 및 분류 결과, 우리 모델을 현재 최신 방법과 비교합니다. 이 표의 모든 작업 평가는 GLUE 벤치마크를 사용하여 수행되었습니다. (mc= 매튜스 상관관계, acc=정확도, pc=피어슨 상관관계)

Overall, our approach achieves new state-of-the-art results in 9 out of the 12 datasets we evaluate on, 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).

전반적으로, 우리의 접근 방식은 우리가 평가한 12개 데이터 세트 중 9개에서 새로운 최첨단 결과를 달성하여 많은 경우 앙상블을 능가합니다. 우리의 결과는 또한 우리의 접근 방식이 STS-B(≈5.7k 교육 예제)와 같은 더 작은 데이터 세트에서 가장 큰 데이터 세트인 SNLI(≈550k 교육 예제)에 이르기까지 다양한 크기의 데이터 세트에서 잘 작동함을 나타냅니다.

**5 Analysis**

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.

전송된 레이어 수의 영향 우리는 감독되지 않은 사전 훈련에서 감독 대상 작업으로 다양한 수의 레이어를 전송하는 영향을 관찰했습니다. 그림 2(왼쪽)는 MultiNLI 및 RACE에 대한 접근 방식의 성능을 전송된 레이어 수의 함수로 보여줍니다. 임베딩을 전송하면 성능이 향상되고 각 변환기 레이어는 MultiNLI에서 전체 전송에 대해 최대 9%의 추가 이점을 제공한다는 표준 결과를 관찰합니다. 이는 사전 훈련된 모델의 각 계층에 대상 작업을 해결하는 데 유용한 기능이 포함되어 있음을 나타냅니다.

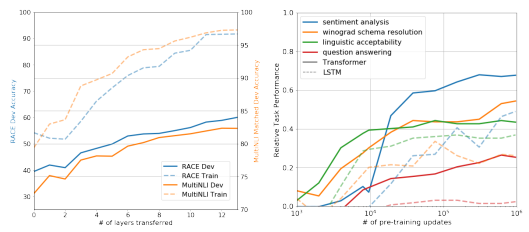


Figure 2: (left) Effect of transferring increasing number of layers from the pre-trained language model on RACE and MultiNLI. (right) Plot showing the evolution of zero-shot performance on different tasks as a function of LM pre-training updates. Performance per task is normalized between a random guess baseline and the current state-of-the-art with a single model.

그림 2: (왼쪽) RACE 및 MultiNLI에서 사전 훈련된 언어 모델에서 증가하는 레이어 수를 전송하는 효과. (오른쪽) LM 사전 훈련 업데이트의 기능으로 다양한 작업에서 제로샷 성능의 진화를 보여주는 플롯. 작업당 성능은 무작위 추측 기준선과 단일 모델을 사용한 최신 기술 사이에서 정규화됩니다.

Zero-shot Behaviors We’d like to better understand why language model pre-training of transformers 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.

Zero-shot Behaviors 우리는 왜 변환기의 언어 모델 사전 훈련이 효과적인지 더 잘 이해하고 싶습니다. 가설은 기본 생성 모델이 언어 모델링 기능을 향상시키기 위해 우리가 평가하는 많은 작업을 수행하는 방법을 배우고 변환기의 더 구조화된 주의 메모리가 LSTM과 비교하여 전달을 지원한다는 것입니다. 우리는 기본 생성 모델을 사용하여 감독 미세 조정 없이 작업을 수행하는 일련의 휴리스틱 솔루션을 설계했습니다. 생성적 사전 훈련 과정에서 이러한 휴리스틱 솔루션의 효과를 그림 2(오른쪽)에서 시각화합니다. 우리는 이러한 휴리스틱의 성능이 안정적이고 훈련을 통해 꾸준히 증가하는 것을 관찰했으며, 이는 생성적 사전 훈련이 다양한 작업 관련 기능의 학습을 지원함을 시사합니다. 우리는 또한 LSTM이 트랜스포머 아키텍처의 유도성 바이어스가 전송을 지원한다는 것을 암시하는 제로샷 성능에서 더 높은 변동을 보이는 것을 관찰했습니다.

테이블이(가) 표시된 사진

자동 생성된 설명

Table 5: Analysis of various model ablations on different tasks. Avg. score is a unweighted average of all the results. (mc= Mathews correlation, acc=Accuracy, pc=Pearson correlation)

표 5: 다양한 작업에 대한 다양한 모델 절제 분석. 평균 점수는 모든 결과의 가중 평균입니다. (mc= 매튜스 상관관계, acc=정확도, pc=피어슨 상관관계)

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 average token log-probability when conditioned on the document and question. For DPRD [46] (winograd schemas), we replace the definite pronoun with the two possible referrents and predict the resolution that the generative model assigns higher average token log-probability to the rest of the sequence after the substitution.

CoLA(언어적 수용 가능성)의 경우, 생성 모델이 할당하는 평균 토큰 로그 확률로 예제가 채점되고 임계값에 의해 예측이 이루어집니다. SST-2(감정 분석)의 경우 각 예제에 바로 토큰을 추가하고 언어 모델의 출력 분포를 긍정 및 부정 단어로만 제한하고 예측으로 더 높은 확률을 할당하는 토큰을 추측합니다. RACE(질문 응답)의 경우 생성 모델이 문서 및 질문을 조건으로 했을 때 가장 높은 평균 토큰 로그 확률을 할당하는 답변을 선택합니다. DPRD[46](winograd schemas)의 경우, 우리는 한정 대명사를 두 개의 가능한 참조 대상으로 바꾸고 생성 모델이 치환 후 나머지 시퀀스에 더 높은 평균 토큰 로그 확률을 할당하는 해상도를 예측합니다.

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 the Transformer 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.

절제 연구 우리는 세 가지 다른 절제 연구를 수행합니다(표 5). 먼저 미세 조정 중에 보조 LM 대물렌즈가 없는 방법의 성능을 조사합니다. 우리는 보조 목표가 NLI 작업과 QQP에 도움이 된다는 것을 관찰했습니다. 전반적으로 추세는 더 큰 데이터 세트가 보조 목표의 이점을 갖지만 더 작은 데이터 세트가 그렇지 않음을 시사합니다. 둘째, 동일한 프레임워크를 사용하는 단일 레이어 2048 유닛 LSTM과 Transformer를 비교하여 Transformer의 효과를 분석합니다. Transformer 대신 LSTM을 사용할 때 평균 점수가 5.6 감소했습니다. LSTM은 MRPC라는 하나의 데이터 세트에서만 Transformer를 능가합니다. 마지막으로 사전 훈련 없이 감독 대상 작업에 대해 직접 훈련된 트랜스포머 아키텍처와 비교합니다. 사전 훈련이 부족하면 모든 작업에서 성능이 저하되어 전체 모델에 비해 14.8% 감소하는 것으로 나타났습니다.

**6 Conclusion**

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.

생성적 사전 훈련 및 판별적 미세 조정을 통해 단일 작업 불가지론적 모델로 강력한 자연어 이해를 달성하기 위한 프레임워크를 도입했습니다. 긴 연속 텍스트가 있는 다양한 코퍼스에 대한 사전 교육을 통해 우리 모델은 중요한 세계 지식과 장거리 종속성을 처리하는 능력을 습득한 다음 질문 응답, 의미론적 유사성 평가, 수반 결정 및 텍스트 분류, 우리가 연구하는 12개 데이터 세트 중 9개에서 최신 기술을 향상시킵니다. 비지도(사전) 훈련을 사용하여 판별 작업의 성능을 높이는 것은 오랫동안 기계 학습 연구의 중요한 목표였습니다. 우리의 작업은 상당한 성능 향상을 달성하는 것이 실제로 가능하다고 제안하고 어떤 모델(변환기)과 데이터 세트(장거리 종속성이 있는 텍스트)가 이 접근 방식에서 가장 잘 작동하는지에 대한 힌트를 제공합니다. 이것이 자연어 이해와 다른 영역 모두에 대한 비지도 학습에 대한 새로운 연구를 가능하게 하고 비지도 학습이 작동하는 방식과 시기에 대한 이해를 더욱 향상시키는 데 도움이 되기를 바랍니다.