

# 어댑터 구현 프로젝트 EIDA (Estimated Intrinsic Dimension Adapter)

조성해

<https://github.com/Sunghae-Cho/EIDA>

개발기간 : 2024년 12월 - 2025년 1월

## 1 제작 동기

LoRA의 논문이 딥러닝 모델의 intrinsic dimension에 관한 선행 연구들로부터 영향을 받았다는 것을 알고 해당 연구들에 흥미가 생겼습니다. Li et al. (2018)에서는 최적화된 모델을 향해 가는 학습경로가 랜덤하게 초기화된 가중치에서 출발하는 상황을 다루며, 랜덤하게 샘플링된 저차원 부분공간 안에서 파라미터의 업데이트를 허용하였습니다. 그러나 fine-tuning을 수행하는 입장에서는 학습과정이 pre-trained model이라는 고정된 시작점을 갖게 됩니다. 그래서 모델의 현재 상태와 데이터셋에 맞는 intrinsic dimension의 추정을 수행하는 것이 가능하며, fine-tuning의 상황에 더 적합하다고 생각했습니다. Train set의 데이터를 모델에 통과시켜서 얻을 수 있는 정보를 활용하여 intrinsic dimension의 방향에 맞게 학습가능한 파라미터를 줄이는 어댑터를 구현하는 연구 프로젝트를 수행하였습니다.

## 2 프로젝트 목표

- 실험을 통해 데이터가 모델을 통과하는 각 지점에서 token representation의 분포가 저차원에 얼마나 집중되는지 측정한다.
- 분포의 저차원 구조에 맞춘 어댑터를 구현하고 fine-tuning을 수행해본다.

그리고 객관적인 목표는 아니지만, 이 프로젝트를 통해 언어 모델에서 token representation의 비등방적 분포와 편향에 관한 연구인 Ethayarajh (2019)가 다뤘던 현상을, 본 연구에서는 PCA를 수행하면서 관찰해보려는 목적이 있습니다. 더 나아가 token representation의 분포가 작은 차원에 집중되어 있다는 사실이 언어 모델의 경량화 기법에 어떻게 반영되는지 이해하는 데에 관심이 있습니다.

실험 및 구현을 다음 두 종류의 모델, 데이터셋 쌍에 대해 수행하였습니다.

- RoBERTa-base 모델, GLUE SST-2 데이터셋
- GPT2 모델, E2E NLG Challenge 데이터셋

### 3 어댑터 구조

$W$ 를 언어모델의 한 가중치라고 하겠습니다.  $W$ 에 입력되는 시퀀스 안의 각 토큰들은  $X$ 로 쓰겠습니다. 데이터 배치 하나가 모델을 다 통과하여 gradient update가 일어난 후의 가중치는  $W + \Delta W$ 로 쓰겠습니다.

$X$ 의 분포가 어떤 작은 차원 부분공간에 집중되어 있다면,  $\Delta W$ 가 입력 토큰에 어떻게 작용하는지를 이 부분공간의 원소들에 대해서만 기술해도 선형변환으로서  $\Delta W$ 의 역할을 대부분 설명하는 것이 됩니다. 이 점을 이용하여 학습과정에서 발생하는  $\Delta W$ 를 작은 크기의 좋은 근사 행렬로 표현하는 것이 어댑터 EIDA의 설계 목적입니다.

주어진 데이터셋의 문장들로 일정한 크기의 배치를 구성하여 모델에 통과시키면서,  $W$ 에 입력되는 배치에서 토큰  $X$ 의 표본을 추출합니다. 배치가 모델을 다 통과하고 나서  $\Delta W$ 가 얻어지면  $\Delta W \cdot X$ 를 계산하여 별도로 저장해둡니다.

$X$ 의 표본으로 주성분 분석을 수행하여,  $W$ 의 정의역의 원소들을  $X$ 의 분포가 집중된 부분공간에 정사영하는 함수  $A$ 를 얻습니다. 그리고  $\Delta W \cdot X$ 의 표본으로 주성분 분석을 수행하여,  $W$ 의 공역의 원소들을  $\Delta W \cdot X$ 의 분포가 집중된 부분공간에 정사영하는 함수  $C$ 를 얻습니다.

그린 다음 0으로 초기화된 학습가능한 파라미터  $B$ 를  $A$ 와  $C$  사이에 두는 형태로 어댑터를 구성합니다. 가중치  $W$ 를  $W + C^T \cdot B \cdot A$ 로 교체합니다. 여기서  $C^T$ 는 행렬  $C$ 의 transpose를 의미하며, 주성분 분석에서 얻어지는 정사영 행렬의 각 열이 서로 orthonormal하여  $C \cdot C^T = I$ 가 되므로, 부분공간의 원소를 전체 latent space에 그대로 넣는 함수가 됩니다.

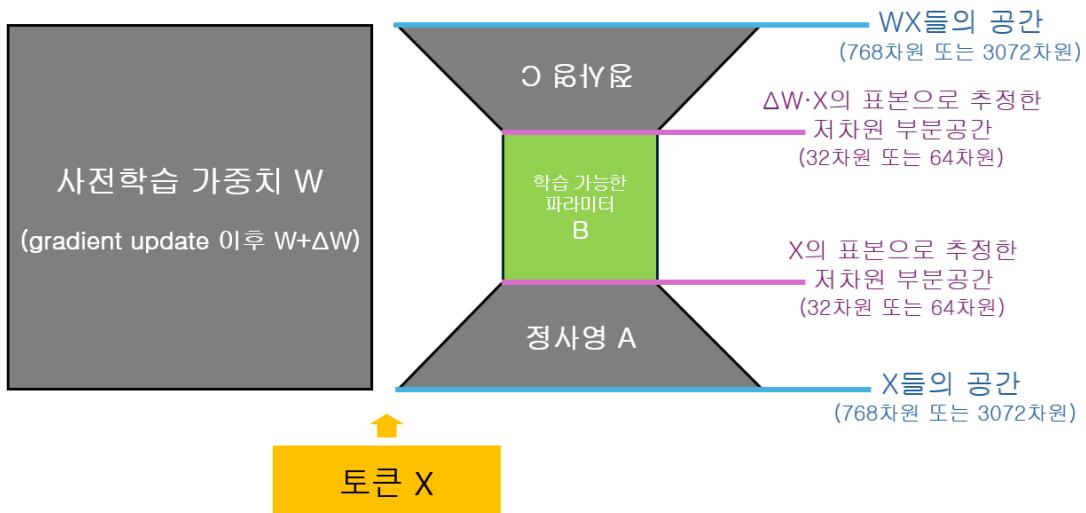


Figure 1: The structure of EIDA.

이렇게 학습과정이  $\Delta W$ 의 저차원 근사를  $B$  부분에 실현하도록 구성하였습니다.

Gradient update  $\Delta W$  자체를 분석 대상으로 삼아서 표본을 수집하게 되면, 일단  $\Delta W$ 가 너무 큰 공간의 원소이고, 그런만큼 안정적인 추정을 위해 요구되는 표본의 수도 커지게 됩니다. 그래서 그보다 훨씬 작은 차원 공간의 원소인 token representation들을 분석 대상으로 하여  $W$ 의 정의역과 공역의 차원 압축을 수행하고, 이를 통해  $\Delta W$ 의 크기를 줄이는 전략을 취하였습니다.

## 4 RoBERTa-base 모델, GLUE SST-2 데이터셋

GLUE SST-2는 영화 리뷰 문장으로 감성 분석을 학습하는 데이터셋입니다. Label 1은 긍정, 0은 부정을 나타냅니다. Label의 종류에 맞춰 num\_labels=2 옵션으로 허깅페이스에 업로드된 RoBERTa-base 모델을 불러와서 학습하였습니다.

### 4.1 토큰 분포 관찰

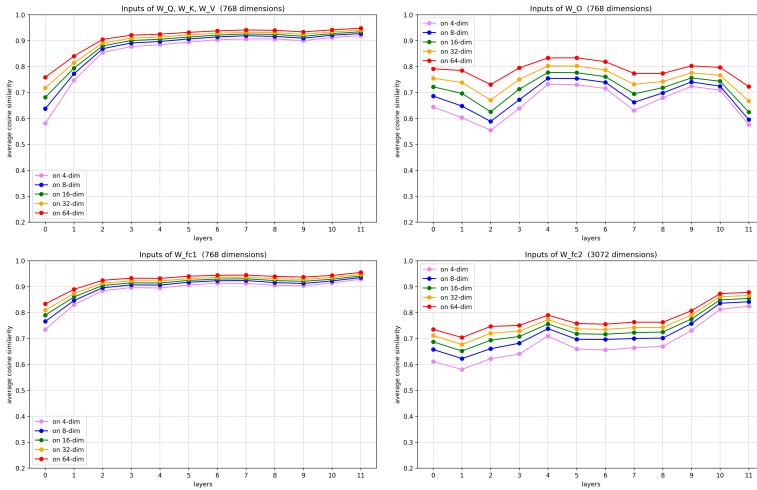
표본 추출은 다음 73개의 가중치의 입력 토큰을 대상으로 하였습니다.

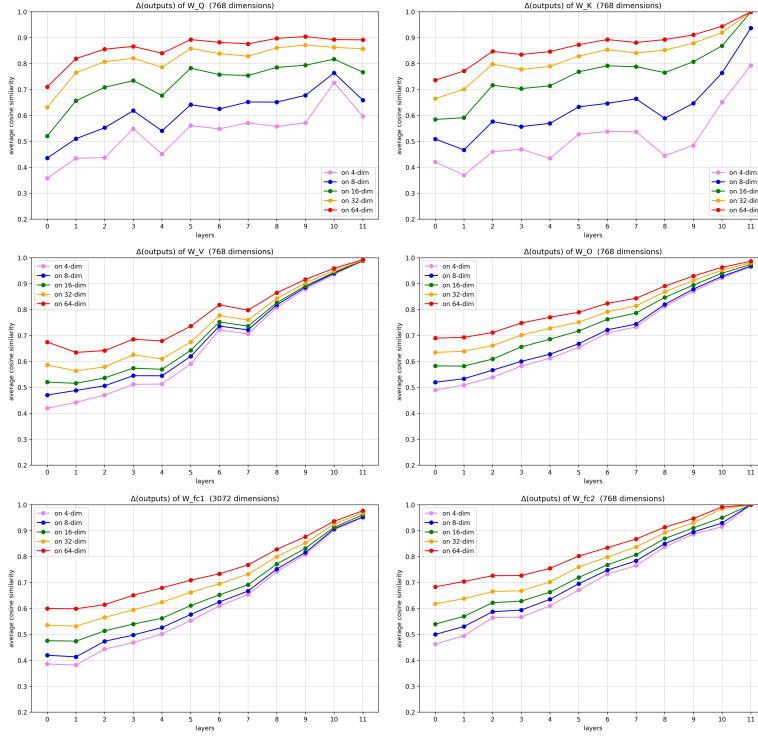
- 12개 encoding layer 각각에서:  $W_Q, W_K, W_V, W_O, W_{fc1}, W_{fc2}$
- 모델 맨 마지막의 classifier 중에서 첫번째 가중치

실험은 다음 과정을 따릅니다.

- SST-2의 train set에서 데이터 256개를 임의로 골라 배치 크기 16으로 모델에 통과시키면서, 가중치의 입력 토큰  $X$ 의 표본을 각 시퀀스 당 2개씩 추출합니다. 한 레이어의  $W_Q, W_K, W_V$ 는 동일한 토큰들을 입력으로 가지므로, 모델 내에서 표본 수집이 이루어지는 장소는 총 49곳입니다. 크기 512의 표본이 49개 얻어집니다.
- 배치가 모델을 다 통과하고 나서  $\Delta W$ 가 얻어지면, 추출된  $X$ 들을 이용해  $\Delta W \cdot X$ 를 계산합니다. 크기 512의 표본이 73개 얻어집니다.
- 가중치의 입력 토큰  $X$ 의 분포 49가지와  $\Delta W \cdot X$ 의 분포 73가지 각각에 대해, 주성분 분석으로 차원 압축을 수행합니다. 이 과정은 토큰의 분포를 4, 8, 16, 32, 64차원 부분공간으로 근사합니다.
- 부분공간 추정에 이용되지 않은 토큰들을 얻기 위해 SST-2의 train set에서 2048개의 데이터를 랜덤하게 새로 고릅니다. 위의 추출과정과 같은 방법으로  $X$ 와  $\Delta W \cdot X$ 의 표본을 얻습니다. 이 토큰들을 추정된 부분공간에 정사영하고 정사영이 원래 토큰과 얼마나 cosine similarity를 가지는지 측정합니다.

전체 과정을 5회 반복하고 평균값을 그래프로 표현하였습니다.





## 4.2 어댑터를 이용한 학습 수행

가중치의 입력  $X$ 의 위치 49곳 중, 그래프에 나타난 cosine similarity의 값이 32차원일 때보다 64차원일 때 0.03 이상 더 높은 10곳에서는  $X$ 의 분포를 64차원으로 근사하고, 그렇지 않은 39곳에서는 32차원으로 근사했습니다.  $\Delta W \cdot X$ 의 위치 73곳 중, 그래프에 나타난 cosine similarity의 값이 32차원일 때보다 64차원일 때 0.03 이상 더 높은 51곳에서는  $\Delta W \cdot X$ 의 분포를 64차원으로 근사하고, 그렇지 않은 22곳에서는 32차원으로 근사했습니다.

Fine-tuning은 다음 과정을 따릅니다.

- 모델에서 73가지의  $W$ 가 아닌 부분은 모델의 맨 마지막 classifier의 두 번째 파라미터(가중치  $2 \times 768$ , 편향 2)를 제외하고 모두 고정합니다.
- SST-2의 train set에서 데이터 256개를 임의로 골라 현재 상태의 모델에 통과시켜 토큰 표본 추출을 수행합니다. (49+73개 latent space 각각에서 512개의 토큰 수집)
- 가중치 73개 각각에 대해, 수집된 표본으로 주성분 분석을 수행하여 정의역을  $X$ 의 분포가 집중된 부분공간에 정사영하는 함수  $A$ 와, 공역을  $\Delta W \cdot X$ 의 분포가 집중된 부분공간에 정사영하는 함수  $C$ 를 얻습니다.
- $A$ 와  $C$  사이에 0으로 초기화된 학습 가능한 파라미터  $B$ 를 두어 어댑터를 구성합니다. 각 가중치  $W$ 를  $W + C^T \cdot B \cdot A$ 로 바꾸고, SST-2의 train set으로 1에폭의 학습을 수행합니다. 여기서  $C^T \cdot B \cdot A$ 의 형태는 입력  $X$ 에  $X \cdot W^T + b$  형태로 작용하는 torch.nn.Linear 파라미터의 방식을 고려해서 적은 것입니다. learning rate 2e-4, weight decay 0.1, FP16 혼합 정밀도를 사용하였습니다. 1에폭의 학습은 warmup 구간 10%와 이후 decay 구간 90%의 linear schedule을 따르도록

했습니다.

- 어댑터  $C^T \cdot B \cdot A$ 를 가중치  $W$ 에 더하고 모델을 저장합니다.

이 과정을 4회 반복하는 것으로 4에폭의 학습을 수행하였습니다. 학습 결과는 다음과 같습니다.

GLUE SST-2 dev set score (Accuracy)	
RoBERTa-base	94.8
RoBERTa-base with LoRA	95.1
RoBERTa-base with EIDA	93.4

Table 1: Comparison with baselines on the GLUE SST-2 dev set.

## 5 GPT2 모델, E2E NLG Challenge 데이터셋

E2E NLG Challenge는 모델이 식당에 대한 정보를 담고 있는 입력인 meaning representation을 받아서 자연어 문장 출력인 human reference를 내놓도록 훈련하는 데이터셋입니다. 예시는 다음과 같습니다.

- meaning representation : “name : The Wrestlers | Type : coffee shop | food : Italian | price : more than £ 30 | area : riverside | family friendly : yes | near : Raja Indian Cuisine”
- human reference : “A coffee shop that is kid friendly near Raja Indian Cuisine named The Wrestlers in the riverside area has a price range of more than £ 30 that serves Italian food .”

### 5.1 토큰 분포 관찰

표본 추출은 다음 72개의 가중치의 입력 토큰을 대상으로 하였습니다.

- 12개 decoding block 각각에서:  $W_Q, W_K, W_V, W_O, W_{fc1}, W_{fc2}$

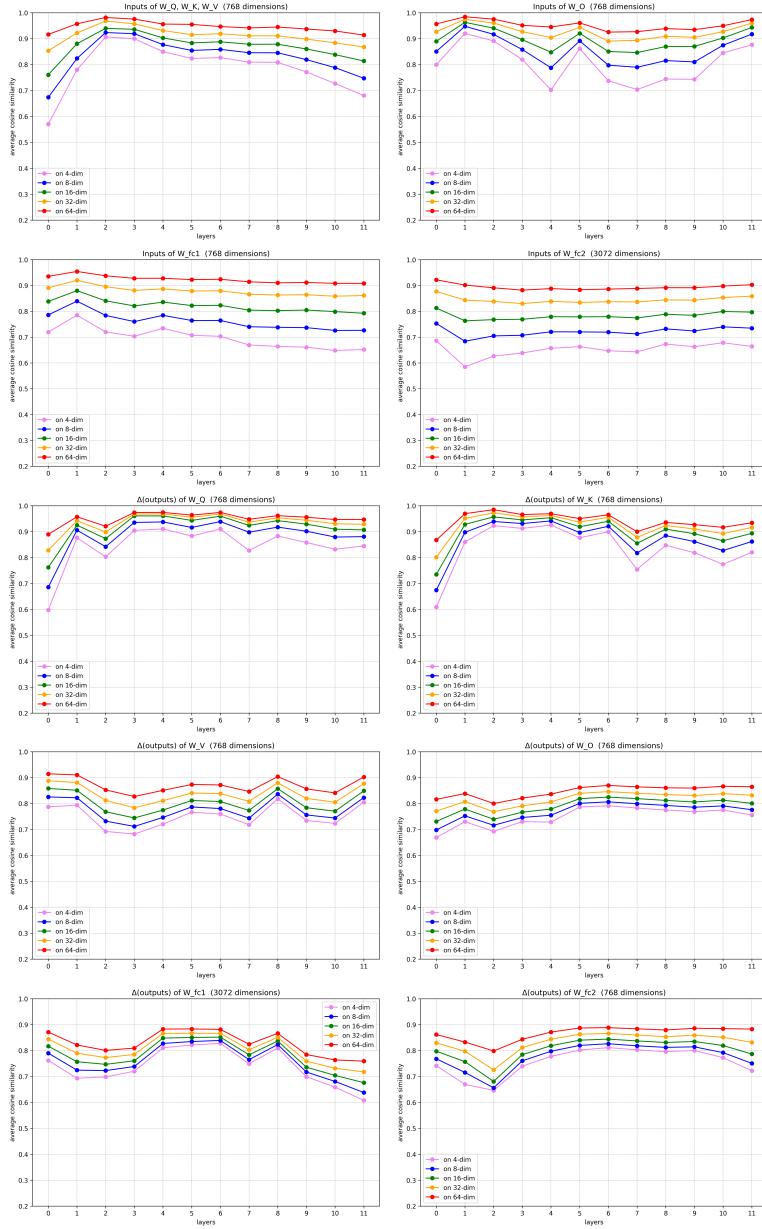
실험은 다음 과정을 따릅니다.

- E2E NLG Challenge의 train set에서 데이터 128개를 임의로 골라 배치 크기 2로 모델에 통과시키면서, 가중치에 입력되는 모든 토큰  $X$ 를 0.75%의 확률로 추출합니다. 이렇게 하면 한 데이터가 문장을 생성하는 auto-regressive decoding loop를 거치는 동안 평균적으로 8개의 토큰을 수집하게 됩니다. 한 블록의  $W_Q, W_K, W_V$ 는 동일한 토큰들을 입력으로 가지므로, 모델 내에서 표본 수집이 이루어지는 장소는 총 48곳입니다. 평균 크기 1024의 표본이 48개 얻어집니다.
- 배치가 모델을 다 통과하고 나서  $\Delta W$ 가 얻어지면, 추출된  $X$ 들을 이용해  $\Delta W \cdot X$ 를 계산합니다. 평균 크기 1024의 표본이 72개 얻어집니다.
- 가중치의 입력 토큰  $X$ 의 분포 48가지와  $\Delta W \cdot X$ 의 분포 72가지 각각에 대해, 주성분 분석으로 차원 압축을 수행합니다. 이 과정은 토큰의 분포를 4, 8, 16, 32, 64차원 부분공간으로 균사합니다.

다.

- 부분공간 추정에 이용되지 않은 토큰들을 얻기 위해 E2E NLG Challenge의 train set에서 128개의 데이터를 새로 고릅니다. 위의 추출과정과 같은 방법으로  $X$ 와  $\Delta W \cdot X$ 의 표본을 얻습니다. 이 토큰들을 추정된 부분공간에 정사영하고 정사영이 원래 토큰과 얼마나 cosine similarity를 가지는지 측정합니다.

전체 과정을 5회 반복하고 평균값을 그래프로 표현하였습니다.



## 5.2 어댑터를 이용한 학습 수행

가중치의 입력  $X$ 의 위치 48곳 중, 그래프에 나타난 cosine similarity의 값이 32차원일 때보다 64차원일 때 0.03 이상 더 높은 35곳에서는  $X$ 의 분포를 64차원으로 근사하고, 그렇지 않은 13곳에서는 32차원으로 근사했습니다.  $\Delta W \cdot X$ 의 위치 72곳 중, 그래프에 나타난 cosine similarity의 값이 32차원일 때보다 64차원일 때 0.03 이상 더 높은 25곳에서는  $\Delta W \cdot X$ 의 분포를 64차원으로 근사하고, 그렇지 않은 47곳에서는 32차원으로 근사했습니다.

Fine-tuning은 다음 과정을 따릅니다.

- 모델에서 72가지의  $W$ 가 아닌 부분은 모두 고정합니다.
- E2E NLG Challenge의 train set에서 데이터 128개를 임의로 골라 현재 상태의 모델에 통과 시켜 토큰 표본 추출을 수행합니다. (48+72개 latent space 각각에서 평균적으로 1024개의 토큰 수집)
- 가중치 72개 각각에 대해, 수집된 표본으로 주성분 분석을 수행하여 정의역을  $X$ 의 분포가 집중된 부분공간에 정사영하는 함수  $A$ 와, 공역을  $\Delta W \cdot X$ 의 분포가 집중된 부분공간에 정사영하는 함수  $C$ 를 얻습니다.
- $A$ 와  $C$  사이에 0으로 초기화된 학습 가능한 파라미터  $B$ 를 두어 어댑터를 구성합니다. 각 가중치  $W$ 를  $W + A \cdot B \cdot C^T$ 로 바꾸고, E2E NLG Challenge의 train set으로 1에폭의 학습을 수행합니다. 여기서  $A \cdot B \cdot C^T$ 의 형태는 입력  $X$ 에  $b + X \cdot W$  형태로 작용하는 GPT-2의 Conv1D 파라미터의 방식을 고려해서 적은 것입니다. learning rate 5e-5, weight decay 0.01, FP16 혼합 정밀도를 사용하였습니다. 1에폭의 학습은 warmup 구간 15%와 이후 decay 구간 85%의 linear schedule을 따르도록 했습니다.
- 어댑터  $A \cdot B \cdot C^T$ 를 가중치  $W$ 에 더하고 모델을 저장합니다.

이 과정을 16회 반복하는 것으로 16에폭의 학습을 수행하였습니다. 모델이 test set에 대한 예측 문장을 생성할 때는 weight 10의 beam search를 이용하도록 했습니다. 결과는 다음과 같습니다.

	BLEU	NIST	MET	ROU	CIDE
GPT-2 Medium (355M) with LoRA	70.4	8.85	46.8	71.8	2.53
GPT-2 (124M) with EIDA	67.4	8.53	45.1	70.1	2.43

Table 2: Comparison with baselines on the E2E NLG Challenge test set.