적대적 훈련 기반의 텍스트 임베딩 증강

김명섭 · 강필성*

고려대학교 산업경영공학과

{myeongsup_kim, pilsung_kang} @ Korea.ac.kr

2021년 대한산업공학회 춘계학술대회

DSBA

Data Science & Business Analytics

목차

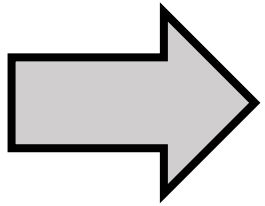
https://www.cleanpng.com/png-bert-ernie-big-bird-count-von-count-oscar-the-grou-1673041/
https://gluebenchmark.com/
https://rajpurkar.github.io/SQuAD-explorer/
https://leaderboard.allenai.org/

<Transformer 기반 언어 모델>

✓ Transformer 기반의 언어 모델의 발달

- Transformer의 구조를 사용한 언어 모델들이 개발되었으며, 다양한 자연어 처리 과업에서 높은 성능을 보이고 있음
- 대표적인 언어 모델인 BERT는 345M개의 매개 변수를 가지고 있으며, 발표 당시 11개의 자연어 처리 과업에서 State-of-the-art의 성능을 기록함

BERT
345M Parameters



NLP Tasks

[1] Vaswani et al., Attention is All You Need, NIPS, 2017
[2] Devlin et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, NAACL, 2019

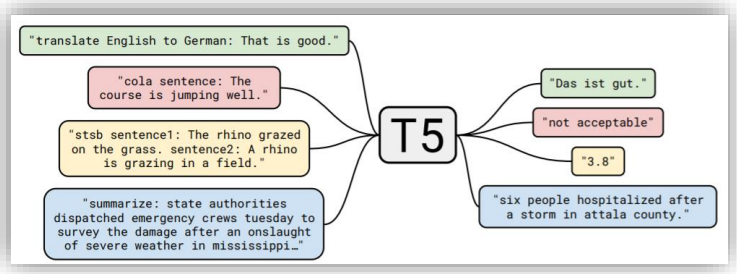
https://www.cleanpng.com/png-bert-ernie-big-bird-count-von-count-oscar-the-grou-1673041/
https://medium.com/dataseries/gpt-3-model-by-openai-the-new-hype-7d112b6ce996
https://arxiv.org/abs/1910.10683

<Transformer 기반 언어 모델>

✓ 대용량 언어 모델의 발전 동향

- BERT의 성공 이후, 지속적으로 매개 변수의 수를 증가시킨 대용량 언어 모델에 관한 연구가 발표됨
- T5*(Text-to-Text Transfer Transformer)*와 GPT-3*(Generative Pre-trained Transformer-3)*의 경우 매우 우수한 성능을 보이고 있으나, 매개 변수의 수가 너무 많아 실제적으로 사용하기 어렵다는 문제점이 발생
- 이를 해결하기 위해 **언어 모델의 매개 변수의 수를 적게 유지하면서 높은 성능을 산출할 수 있도록 하는 연구들이 수행되고 있음**

BERT
345M Parameters



T5
11B Parameters

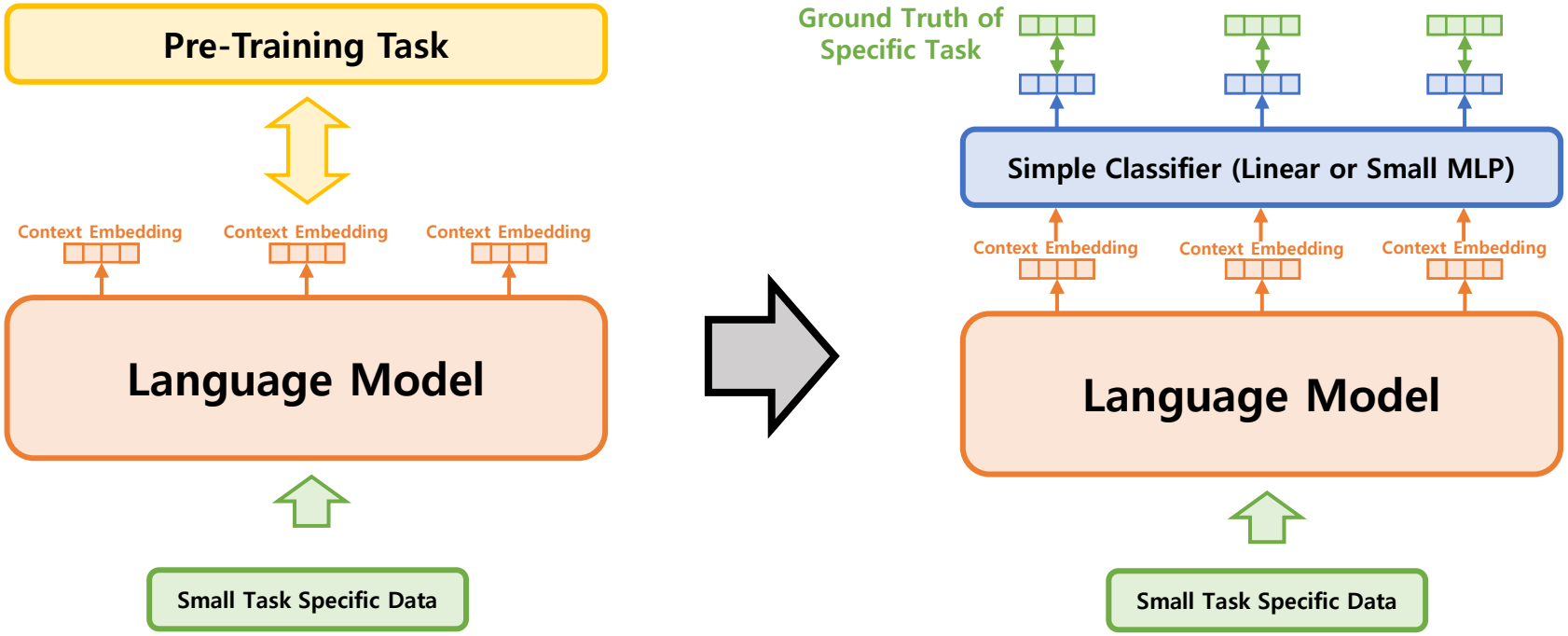
GPT-3
175B Parameters

[3] Raffel et al., Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformers, JMLR, 2020
[4] Brown et al., Language Models are Few-Shot Learners, NeurIPS, 2020

<언어 모델 성능 개선에 관한 연구>

✓ Task-Adaptive Pre-Training (TAPT)

- 일반적으로 언어 모델의 사전 학습에 사용되는 데이터와 미세 조정에 사용되는 데이터의 분포는 상이함
- 과업 데이터를 이용하여 추가적인 사전 학습**(TAPT)**을 수행하고, 이후 목표 과업에 대해 미세 조정을 수행할 경우 유의미한 성능 향상이 나타남



<Step 1: Pre-Training>

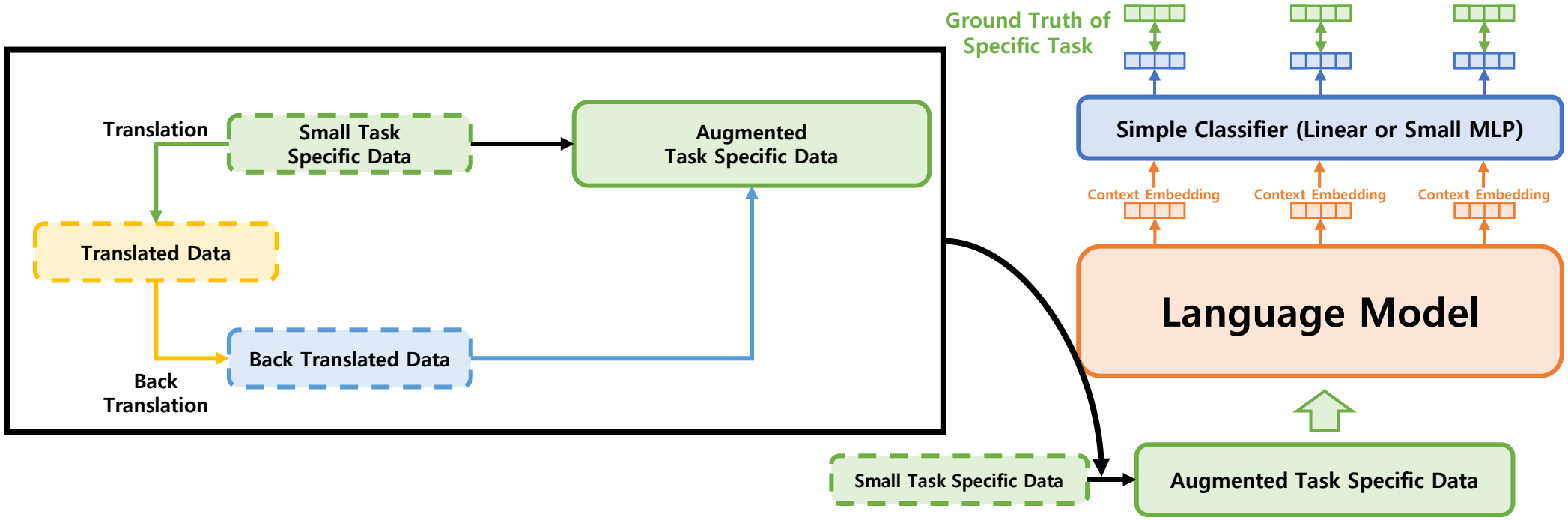
<Step 2: Fine-Tuning>

[5] Gururangan et al., Don't Stop Pretraining: Adapt Language Models to Domains and Tasks, ACL, 2020

<언어 모델 성능 개선에 관한 연구>

✓ Back Translation

- 언어 모델의 Capacity에 비해 Label이 존재하는 데이터의 수가 적기 때문에 과대적합의 위험이 존재
- 이를 방지하기 위해 특정 언어를 다른 언어로 번역한 뒤, 원본 언어로 재 번역하는 과정을 통해 텍스트 데이터를 증강하는 **Back Translation** 방법이 제안되었으며, 유의미한 성능 향상을 보임



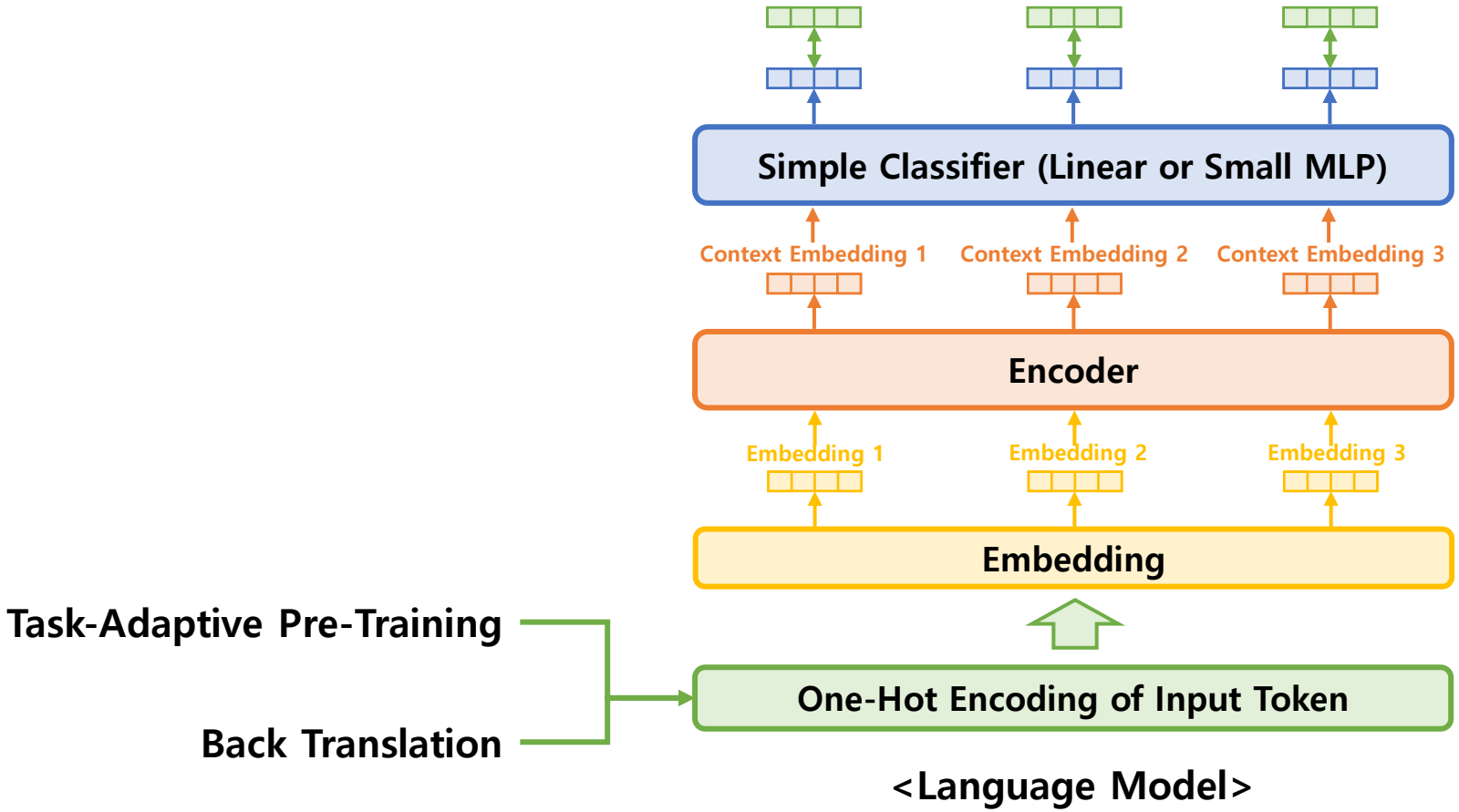
<Step 1: Back Translation>

<Step 2: Fine-Tuning>

[6] Edunov et al., Understanding Back-Translation at Scale, EMNLP, 2018

<언어 모델 성능 개선에 관한 연구>

- 하지만 **TAPT와 Back Translation 모두 자연어 토큰의 조정을 통해 언어 모델의 성능 향상을 시도**하였으며, **언어 모델의 직접적인 입력인 임베딩에 대한 조정은 토큰을 통해 간접적으로 수행되었다는 한계가 존재**



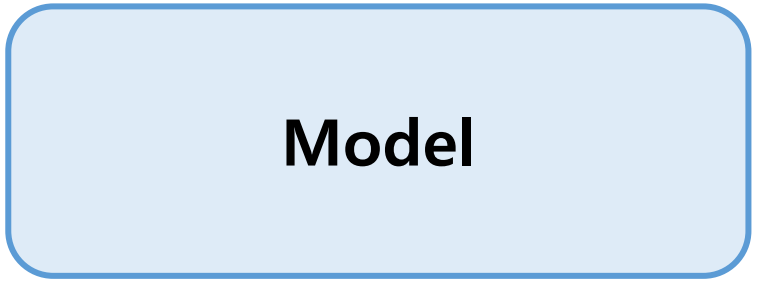
<Language Model>

<적대적 훈련에 관한 연구>

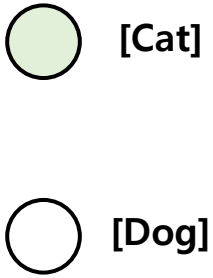
✓ Adversarial Training

- Vision 분야에서 처음으로 제안, 입력 이미지에 미세한 노이즈를 부과하여 사람이 보기에는 원본 이미지와 차이가 없는 적대적 예제(Adversarial Example)를 생성한 후 원본 데이터와 적대적 예제를 함께 학습

<Input Image>



<Training>



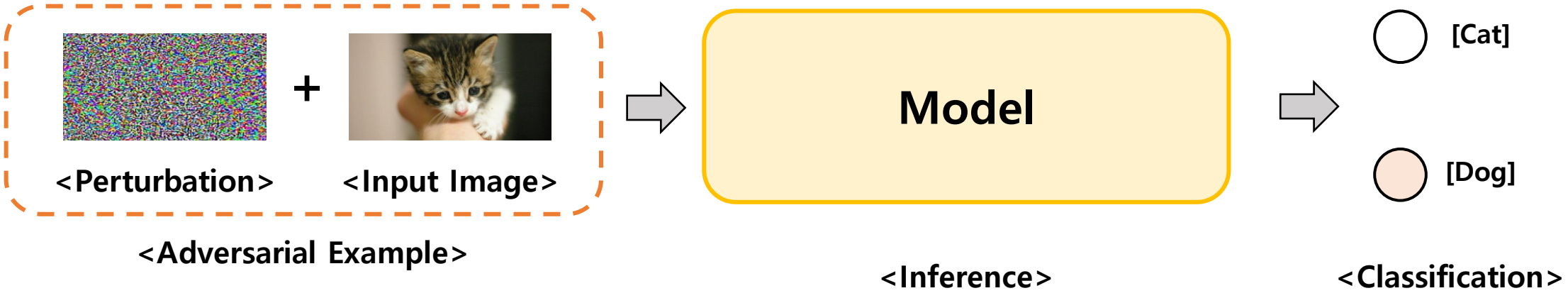
<Classification>

[8] Goodfellow et al., Explaining and Harnessing Adversarial Examples, ICLR, 2015

<적대적 훈련에 관한 연구>

✓ Adversarial Training

- Vision 분야에서 처음으로 제안, 입력 이미지에 미세한 노이즈를 부과하여 사람이 보기에는 원본 이미지와 차이가 없는 적대적 예제(Adversarial Example)를 생성한 후 원본 데이터와 적대적 예제를 함께 학습

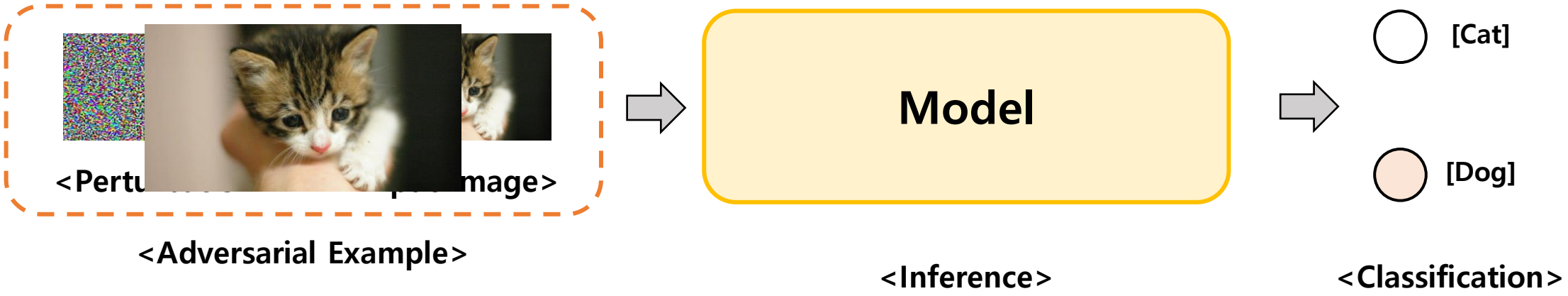


[8] Goodfellow et al., Explaining and Harnessing Adversarial Examples, ICLR, 2015

<적대적 훈련에 관한 연구>

✓ Adversarial Training

- Vision 분야에서 처음으로 제안, 입력 이미지에 미세한 노이즈를 부과하여 사람이 보기에는 원본 이미지와 차이가 없는 적대적 예제(Adversarial Example)를 생성한 후 원본 데이터와 적대적 예제를 함께 학습

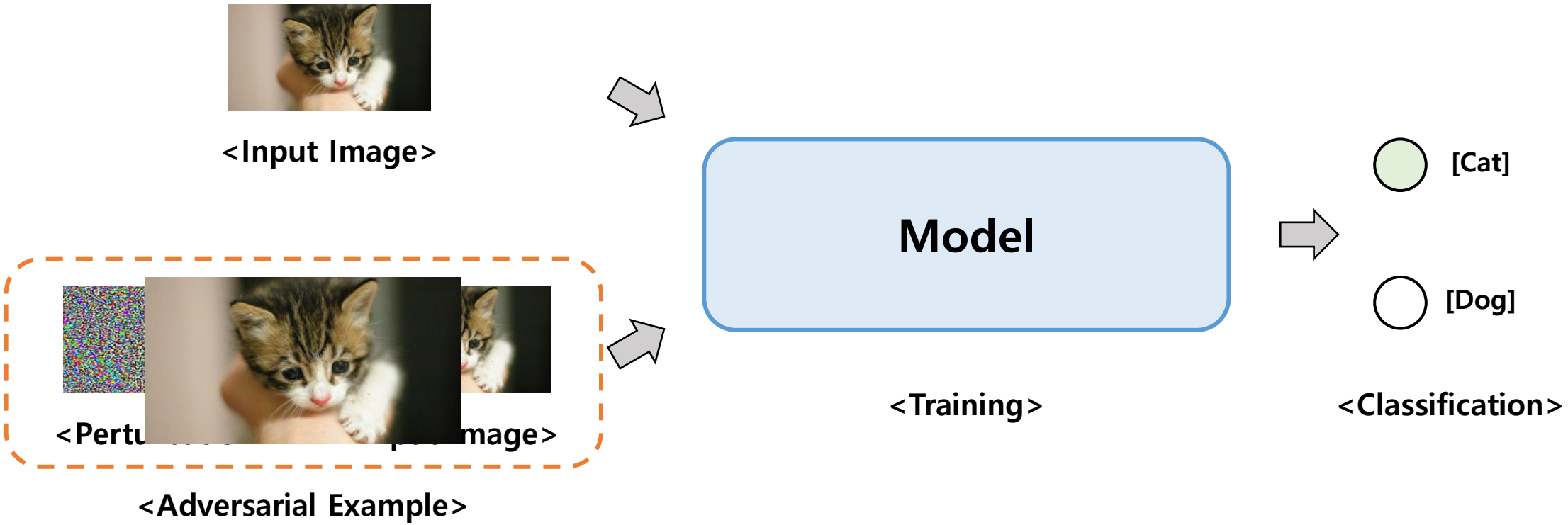


[8] Goodfellow et al., Explaining and Harnessing Adversarial Examples, ICLR, 2015

<적대적 훈련에 관한 연구>

✓ Adversarial Training

- Vision 분야에서 처음으로 제안, 입력 이미지에 미세한 노이즈를 부과하여 사람이 보기에는 원본 이미지와 차이가 없는 적대적 예제(Adversarial Example)를 생성한 후 원본 데이터와 적대적 예제를 함께 학습



[8] Goodfellow et al., Explaining and Harnessing Adversarial Examples, ICLR, 2015

<적대적 훈련에 관한 연구>

✓ Adversarial Training

- Vision 분야에서 처음으로 제안, 입력 이미지에 미세한 노이즈를 부과하여 사람이 보기에는 원본 이미지와 차이가 없는 적대적 예제(Adversarial Example)를 생성한 후 원본 데이터와 적대적 예제를 함께 학습
- **Vision 분야에서 적대적 훈련을 수행할 경우 강건성은 증가**하나, **일반화 성능은 감소**하는 것으로 알려져 있음

<Input Image>

<Adversarial Example>

Model

<Training>

[Cat]

[Dog]

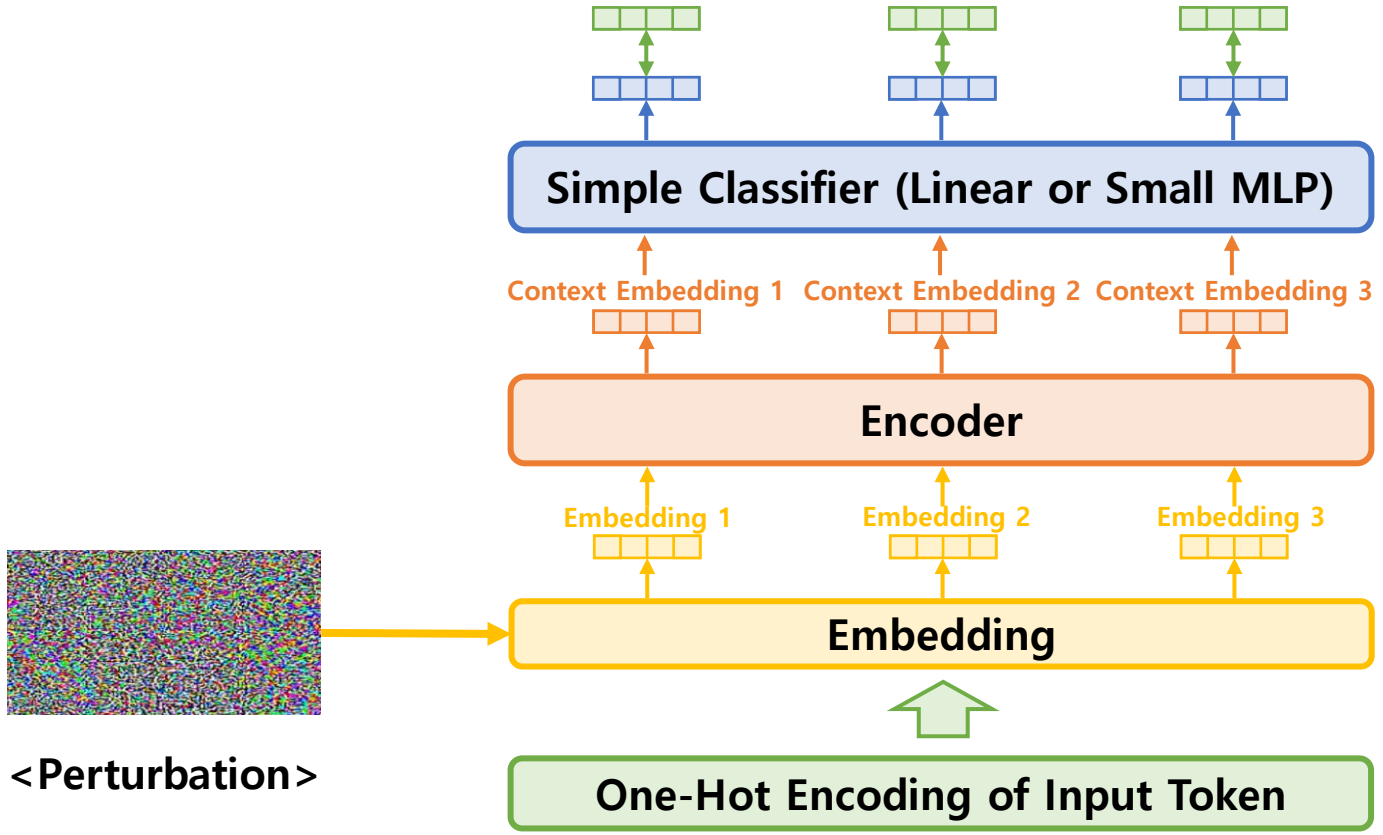
<Classification>

[8] Goodfellow et al., Explaining and Harnessing Adversarial Examples, ICLR, 2015

<적대적 훈련에 관한 연구>

✓ Adversarial Training for NLU

- *Zhu et al.(2020)* 에서는 언어 모델의 임베딩에 대해 적대적 훈련을 수행하는 학습 기법을 제안
- Projected Gradient Descent(PGD)를 기반으로 임베딩에 제한된 크기의 Perturbation을 더하는 방식으로 훈련

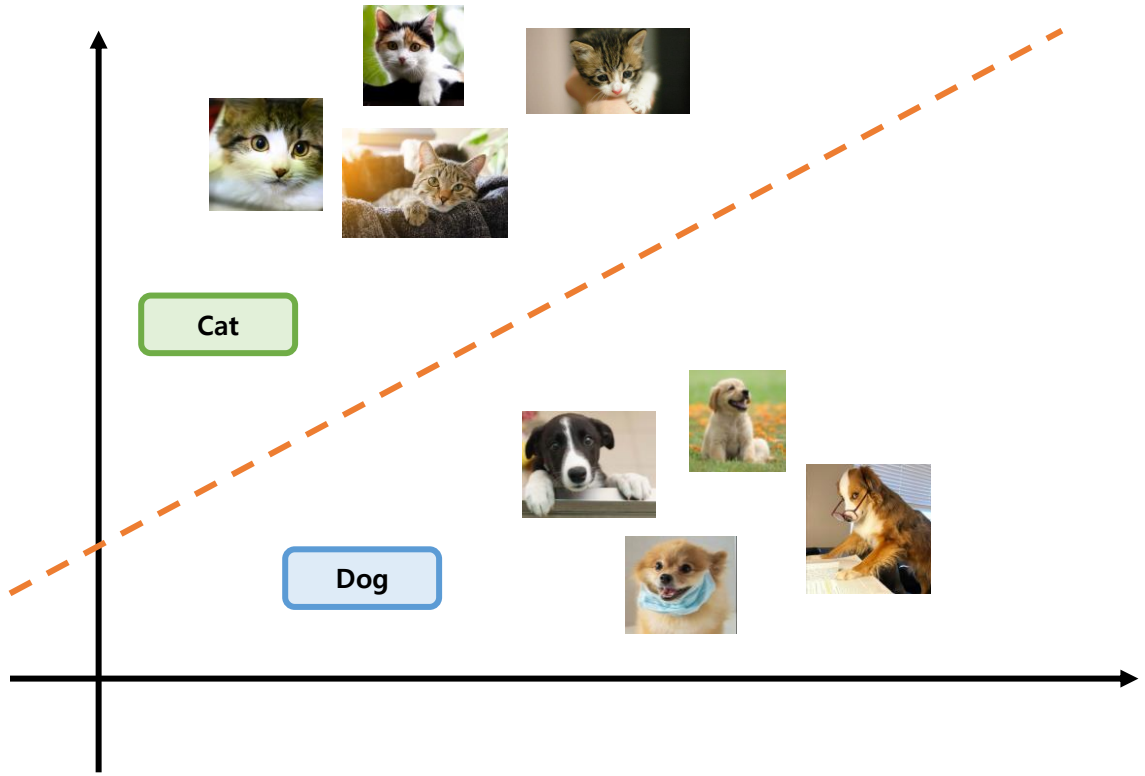


[9] Zhu et al., FreeLB: Enhanced Adversarial Training for Natural Language Understanding, ICLR, 2020

<적대적 훈련에 관한 연구>

✓ Adversarial Training for NLU

- *Zhu et al.(2020)* 에서는 언어 모델의 임베딩에 대해 적대적 훈련을 수행하는 학습 기법을 제안
- Projected Gradient Descent(PGD)를 기반으로 임베딩에 제한된 크기의 Perturbation을 더하는 방식으로 훈련

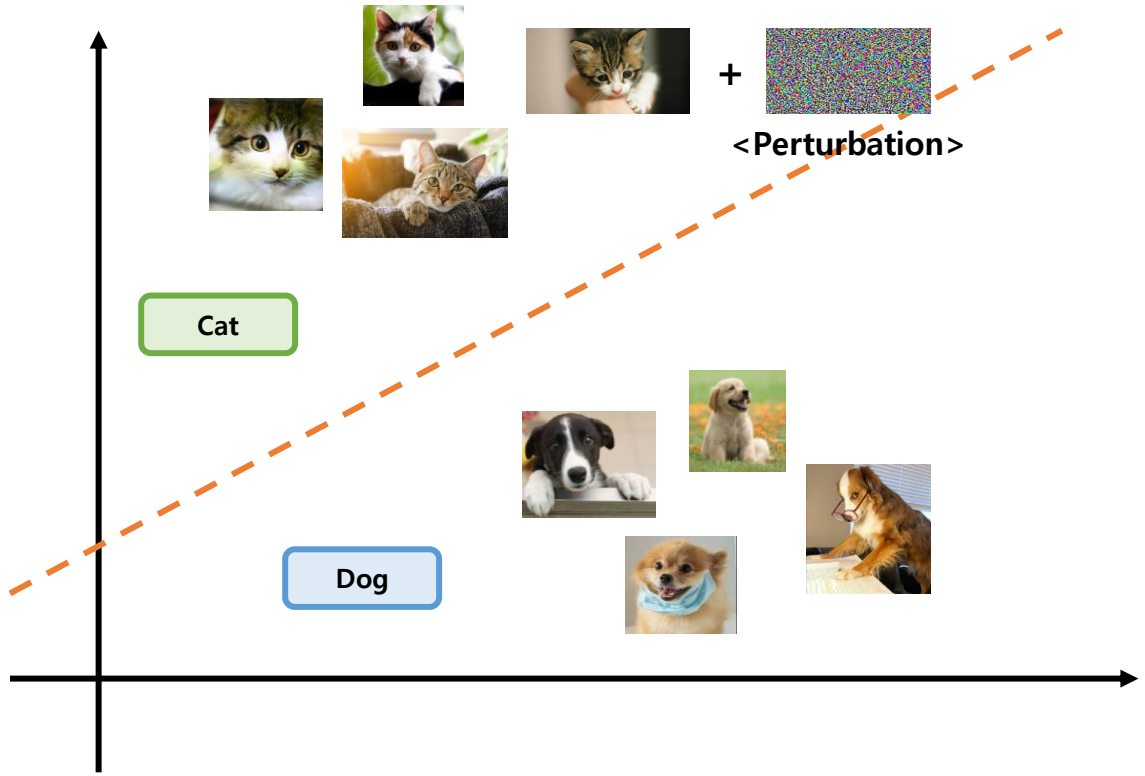


[9] Zhu et al., FreeLB: Enhanced Adversarial Training for Natural Language Understanding, ICLR, 2020
[10] Madry et al., Towards Deep Learning Models Resistant to Adversarial Attacks, ICLR 2018

<적대적 훈련에 관한 연구>

✓ Adversarial Training for NLU

- *Zhu et al.(2020)* 에서는 언어 모델의 임베딩에 대해 적대적 훈련을 수행하는 학습 기법을 제안
- Projected Gradient Descent(PGD)를 기반으로 임베딩에 제한된 크기의 Perturbation을 더하는 방식으로 훈련

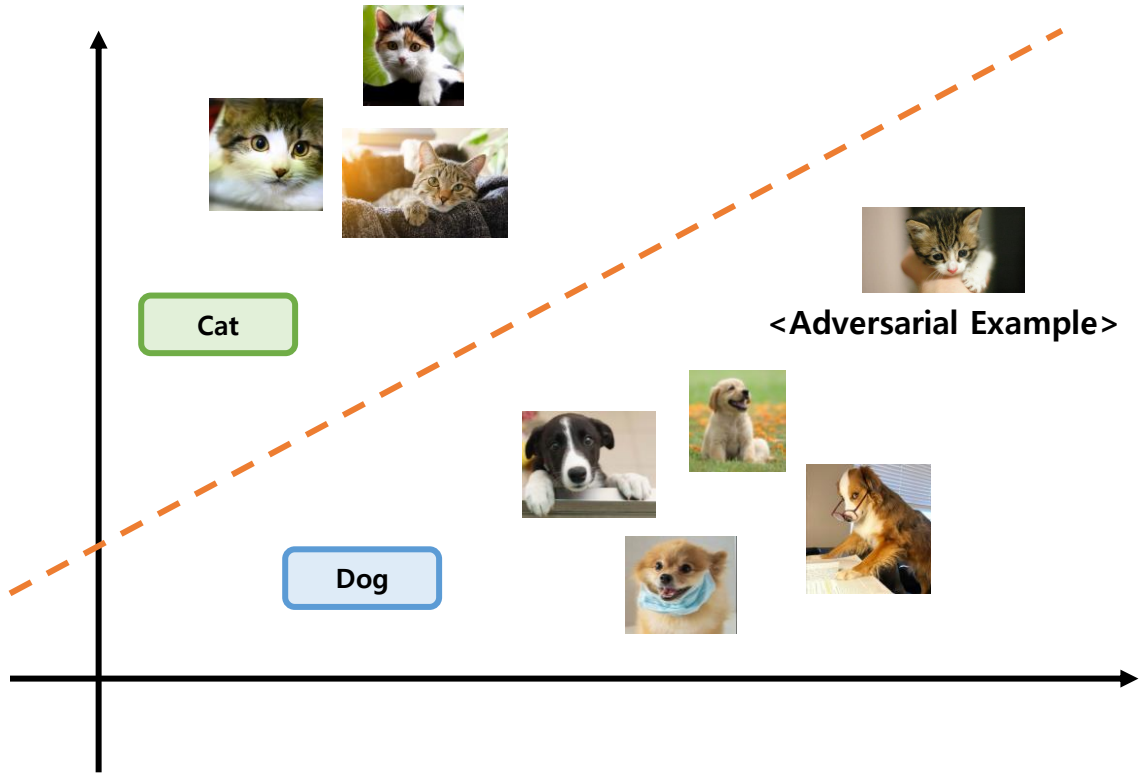


[9] Zhu et al., FreeLB: Enhanced Adversarial Training for Natural Language Understanding, ICLR, 2020
[10] Madry et al., Towards Deep Learning Models Resistant to Adversarial Attacks, ICLR 2018

<적대적 훈련에 관한 연구>

✓ Adversarial Training for NLU

- *Zhu et al.(2020)* 에서는 언어 모델의 임베딩에 대해 적대적 훈련을 수행하는 학습 기법을 제안
- Projected Gradient Descent(PGD)를 기반으로 임베딩에 제한된 크기의 Perturbation을 더하는 방식으로 훈련

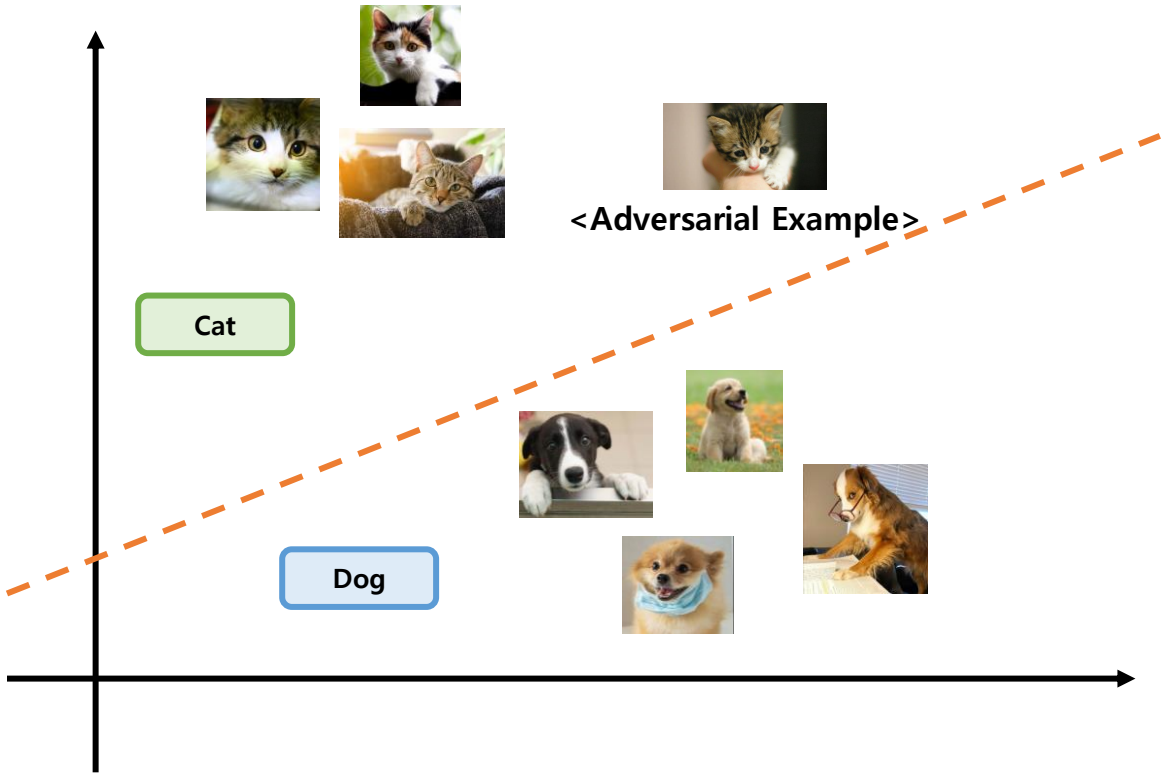


[9] Zhu et al., FreeLB: Enhanced Adversarial Training for Natural Language Understanding, ICLR, 2020
[10] Madry et al., Towards Deep Learning Models Resistant to Adversarial Attacks, ICLR 2018

<적대적 훈련에 관한 연구>

✓ Adversarial Training for NLU

- *Zhu et al.(2020)* 에서는 언어 모델의 임베딩에 대해 적대적 훈련을 수행하는 학습 기법을 제안
- Projected Gradient Descent(PGD)를 기반으로 임베딩에 제한된 크기의 Perturbation을 더하는 방식으로 훈련

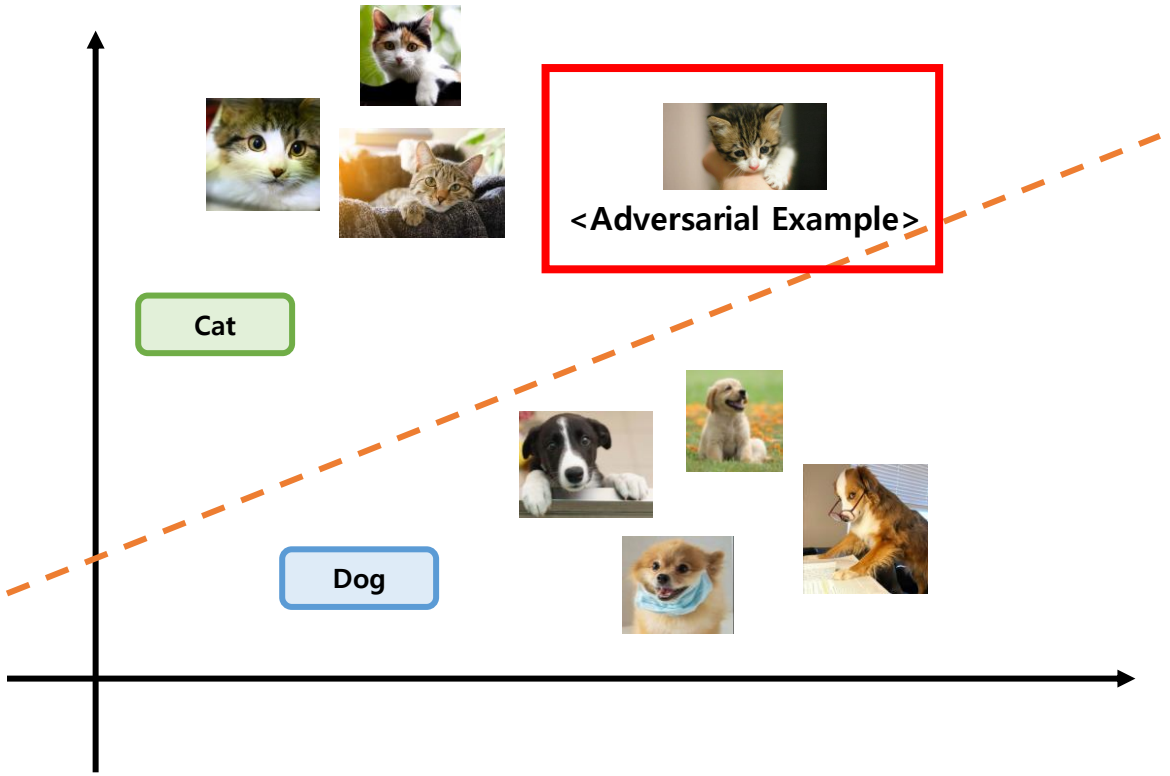


[9] Zhu et al., FreeLB: Enhanced Adversarial Training for Natural Language Understanding, ICLR, 2020
[10] Madry et al., Towards Deep Learning Models Resistant to Adversarial Attacks, ICLR 2018

<적대적 훈련에 관한 연구>

✓ Adversarial Training for NLU

- *Zhu et al.(2020)* 에서는 언어 모델의 임베딩에 대해 적대적 훈련을 수행하는 학습 기법을 제안
- Projected Gradient Descent(PGD)를 기반으로 임베딩에 제한된 크기의 Perturbation을 더하는 방식으로 훈련



[9] Zhu et al., FreeLB: Enhanced Adversarial Training for Natural Language Understanding, ICLR, 2020
[10] Madry et al., Towards Deep Learning Models Resistant to Adversarial Attacks, ICLR 2018

<적대적 훈련에 관한 연구>

✓ Adversarial Training for NLU

- *Zhu et al.(2020)* 에서는 언어 모델의 임베딩에 대해 적대적 훈련을 수행하는 학습 기법을 제안
- Projected Gradient Descent(PGD)를 기반으로 임베딩에 제한된 크기의 Perturbation을 더하는 방식으로 훈련

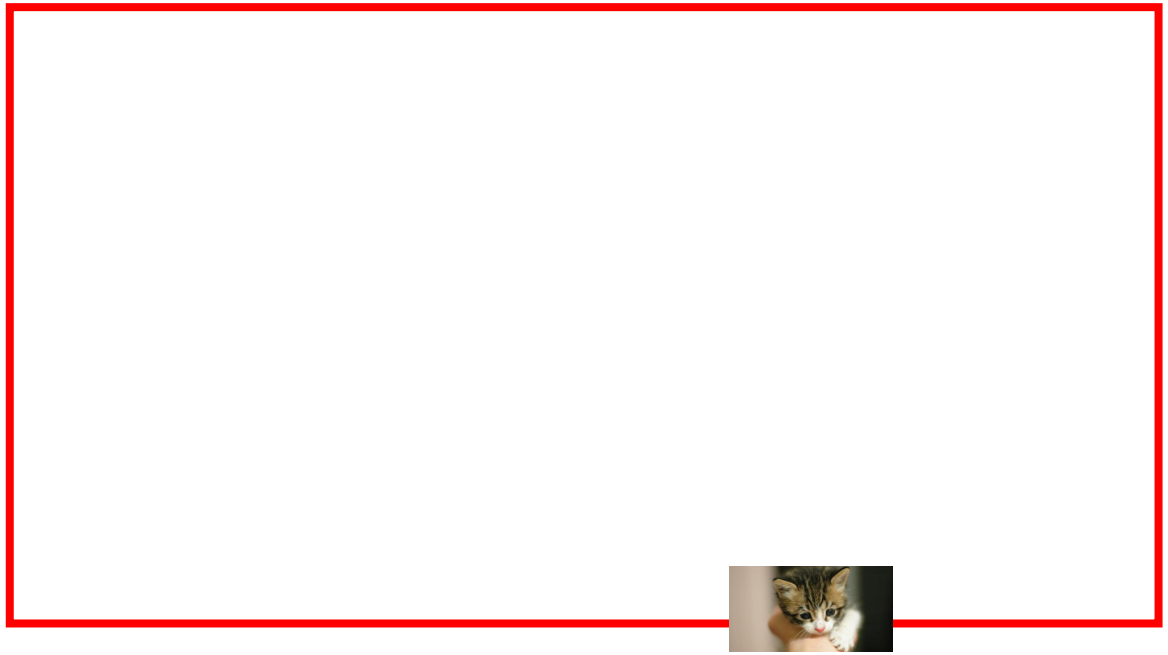
<Adversarial Example>

[9] Zhu et al., FreeLB: Enhanced Adversarial Training for Natural Language Understanding, ICLR, 2020
[10] Madry et al., Towards Deep Learning Models Resistant to Adversarial Attacks, ICLR 2018

<적대적 훈련에 관한 연구>

✓ Adversarial Training for NLU

- *Zhu et al.(2020)* 에서는 언어 모델의 임베딩에 대해 적대적 훈련을 수행하는 학습 기법을 제안
- Projected Gradient Descent(PGD)를 기반으로 임베딩에 제한된 크기의 Perturbation을 더하는 방식으로 훈련



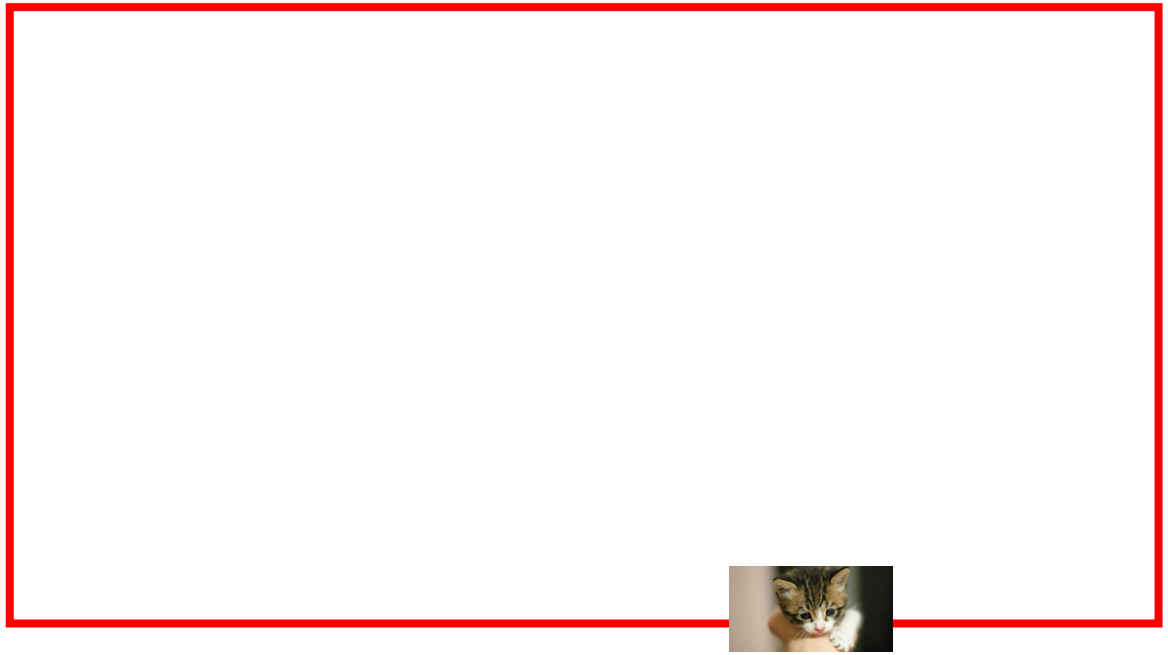
<Adversarial Example>

[9] Zhu et al., FreeLB: Enhanced Adversarial Training for Natural Language Understanding, ICLR, 2020
[10] Madry et al., Towards Deep Learning Models Resistant to Adversarial Attacks, ICLR 2018

<적대적 훈련에 관한 연구>

✓ Adversarial Training for NLU

- *Zhu et al.(2020)* 에서는 언어 모델의 임베딩에 대해 적대적 훈련을 수행하는 학습 기법을 제안
- Projected Gradient Descent(PGD)를 기반으로 임베딩에 제한된 크기의 Perturbation을 더하는 방식으로 훈련
- Vision 분야와는 달리, **NLP 분야에서는 적대적 훈련을 수행**할 경우 **일반화 성능이 향상**되는 것으로 보고됨



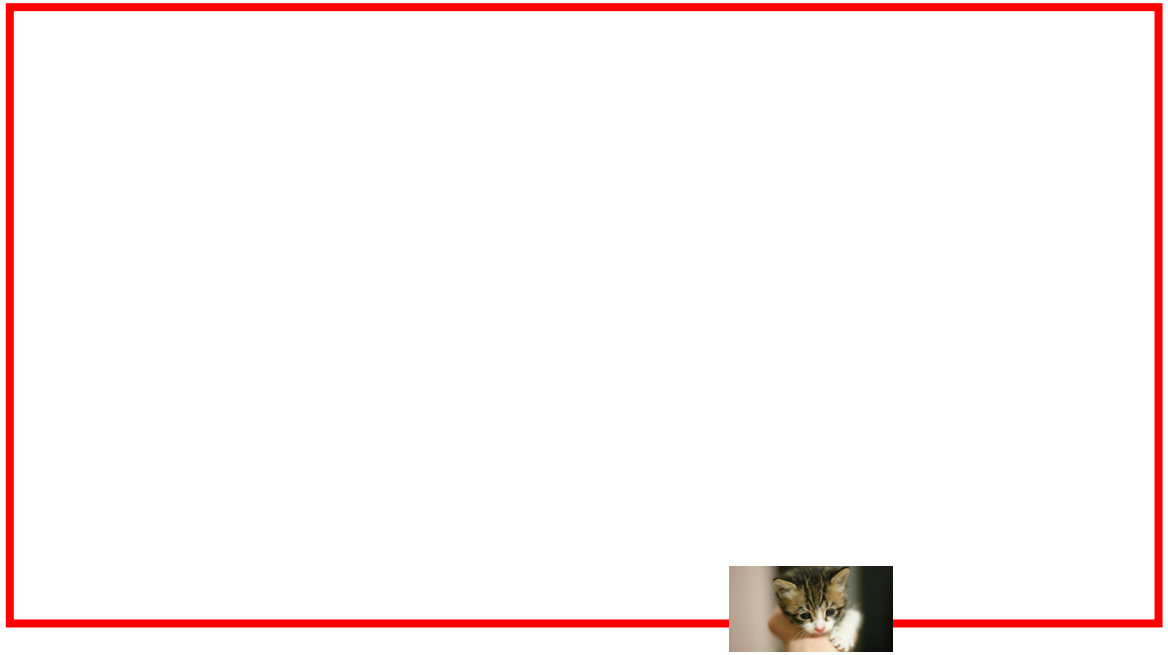
<Adversarial Example>

[9] Zhu et al., FreeLB: Enhanced Adversarial Training for Natural Language Understanding, ICLR, 2020
[10] Madry et al., Towards Deep Learning Models Resistant to Adversarial Attacks, ICLR 2018

<적대적 훈련에 관한 연구>

✓ Adversarial Training for NLU

- *Zhu et al.(2020)* 에서는 언어 모델의 임베딩에 대해 적대적 훈련을 수행하는 학습 기법을 제안
- Projected Gradient Descent(PGD)를 기반으로 임베딩에 제한된 크기의 Perturbation을 더하는 방식으로 훈련
- 하지만, 제안된 방법론은 **일시적으로 적대적 예제를 생성하여 학습을 수행**하며, **한번 사용된 적대적 예제는 다음 번 학습에 참여할 수 없다는 한계**가 존재



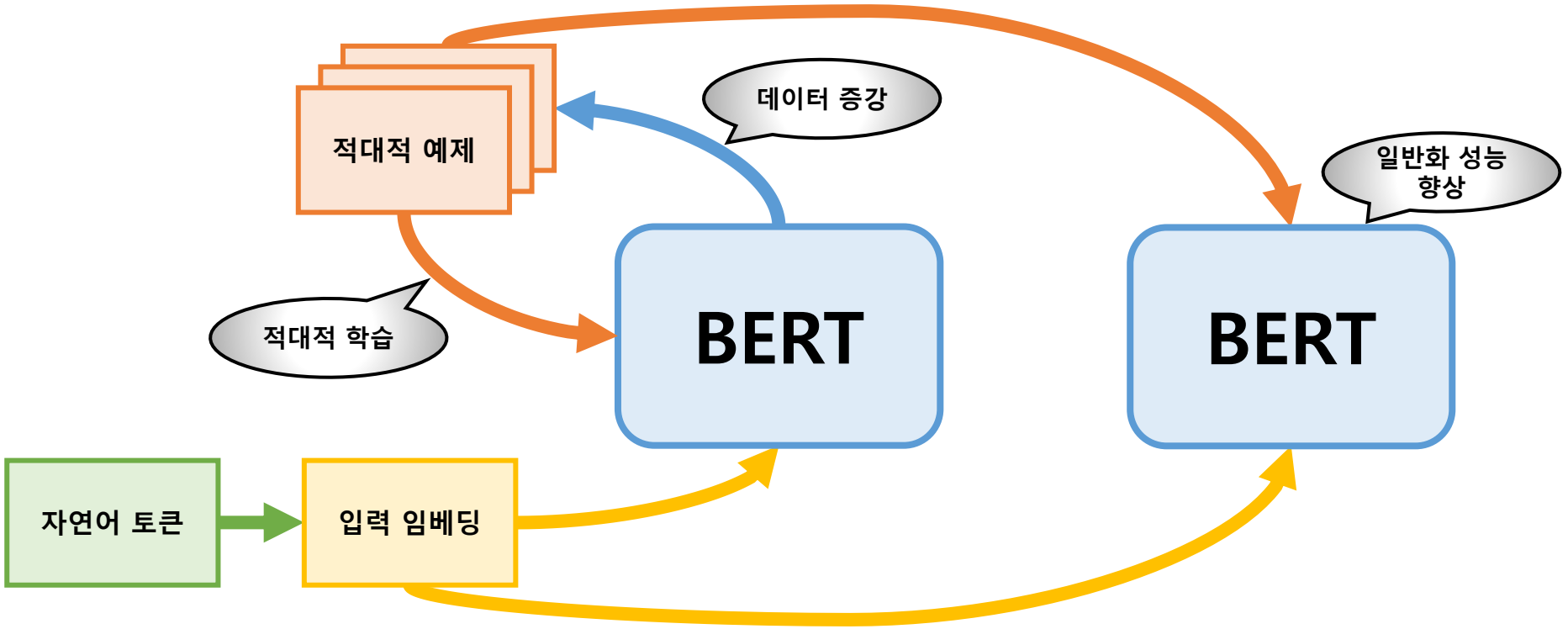
<Adversarial Example>

[9] Zhu et al., FreeLB: Enhanced Adversarial Training for Natural Language Understanding, ICLR, 2020
[10] Madry et al., Towards Deep Learning Models Resistant to Adversarial Attacks, ICLR 2018

<연구 목적>

✓ 적대적 훈련 기반의 텍스트 임베딩 증강

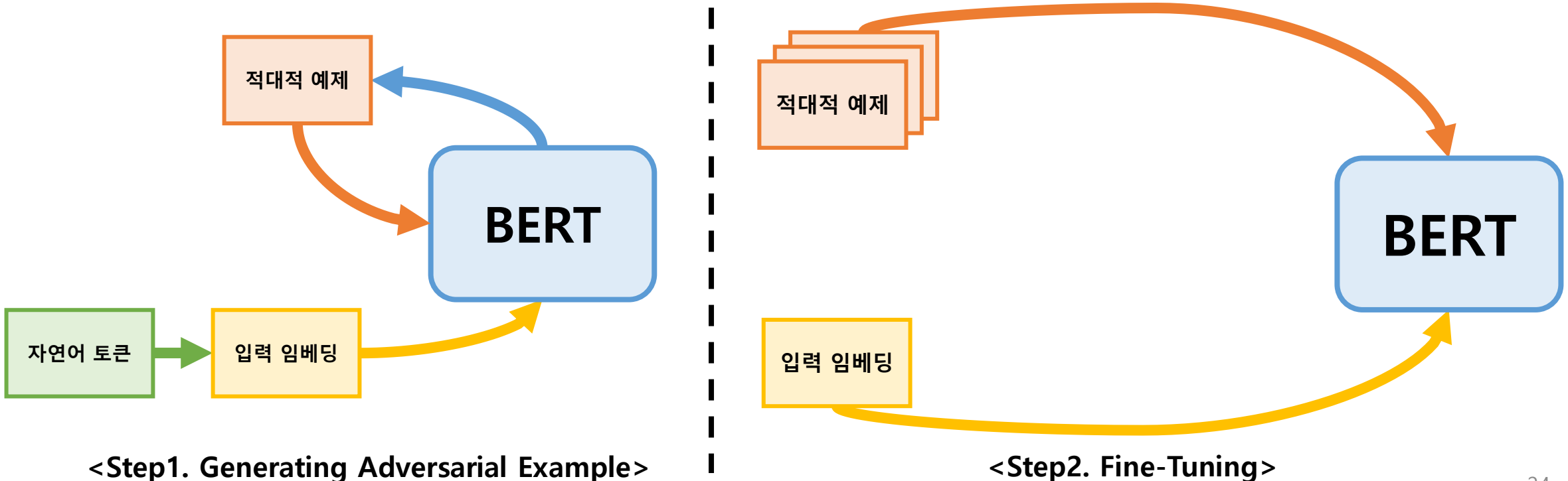
- 언어 모델의 성능 향상에 관한 선행 연구들은 자연어 토큰의 조정을 통해 간접적으로 일반화 성능을 개선
- 임베딩을 조정한 선행 연구에서는 한번 사용된 적대적 예제를 이후의 학습에서 사용할 수 없음
- 본 연구에서는 **적대적 훈련을 기반으로 자연어 임베딩을 증강**하여 **언어 모델 성능을 향상**시키는 방법을 제안



<적대적 훈련 기반의 텍스트 임베딩 증강>

✓ **제안된 모델은 2개의 Step으로 학습을 수행**

- Step 1. 목표 과업 데이터를 이용한 적대적 훈련으로 적대적 예제를 생성
- Step 2. 적대적 예제와 원본 데이터를 이용하여 미세 조정

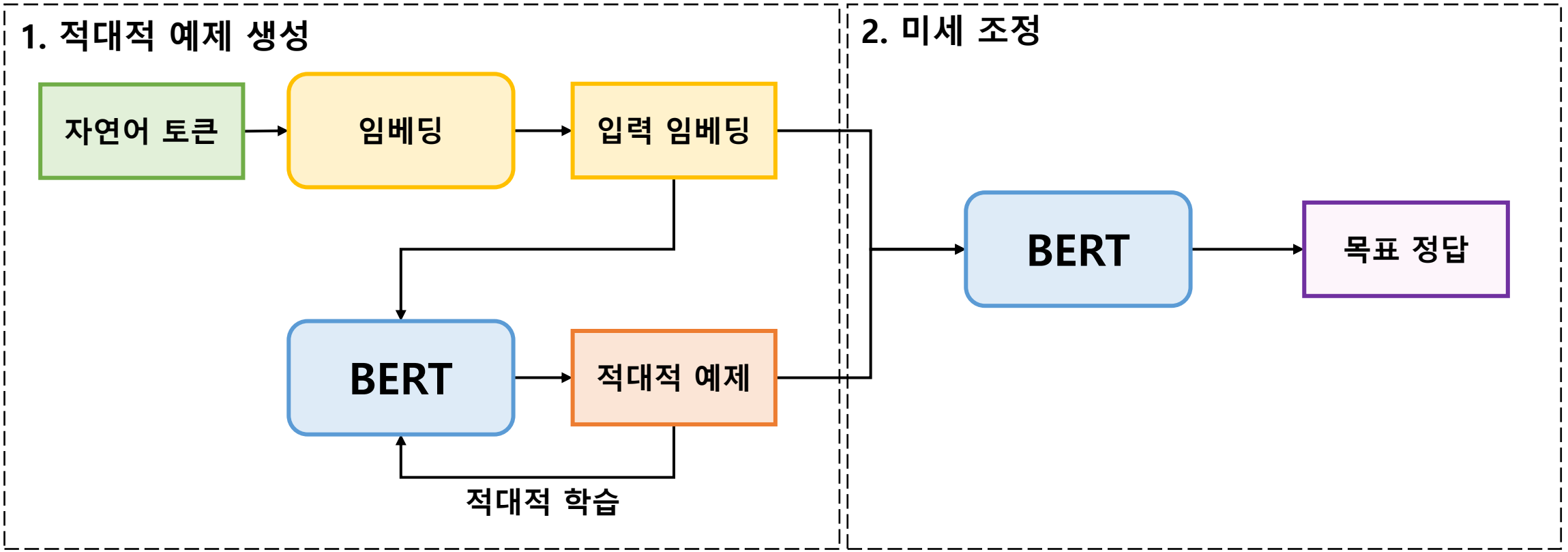


<Step1. Generating Adversarial Example>

<Step2. Fine-Tuning>

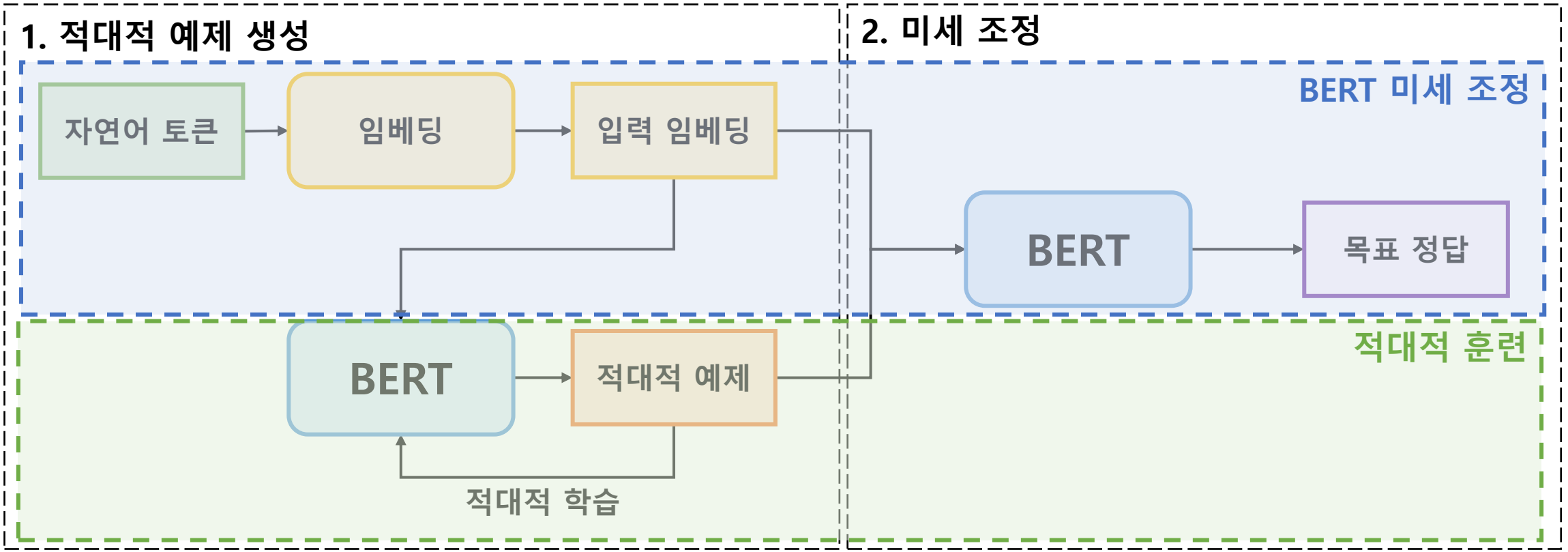
<적대적 훈련 기반의 텍스트 임베딩 증강>

✓ 연구 흐름도



<적대적 훈련 기반의 텍스트 임베딩 증강>

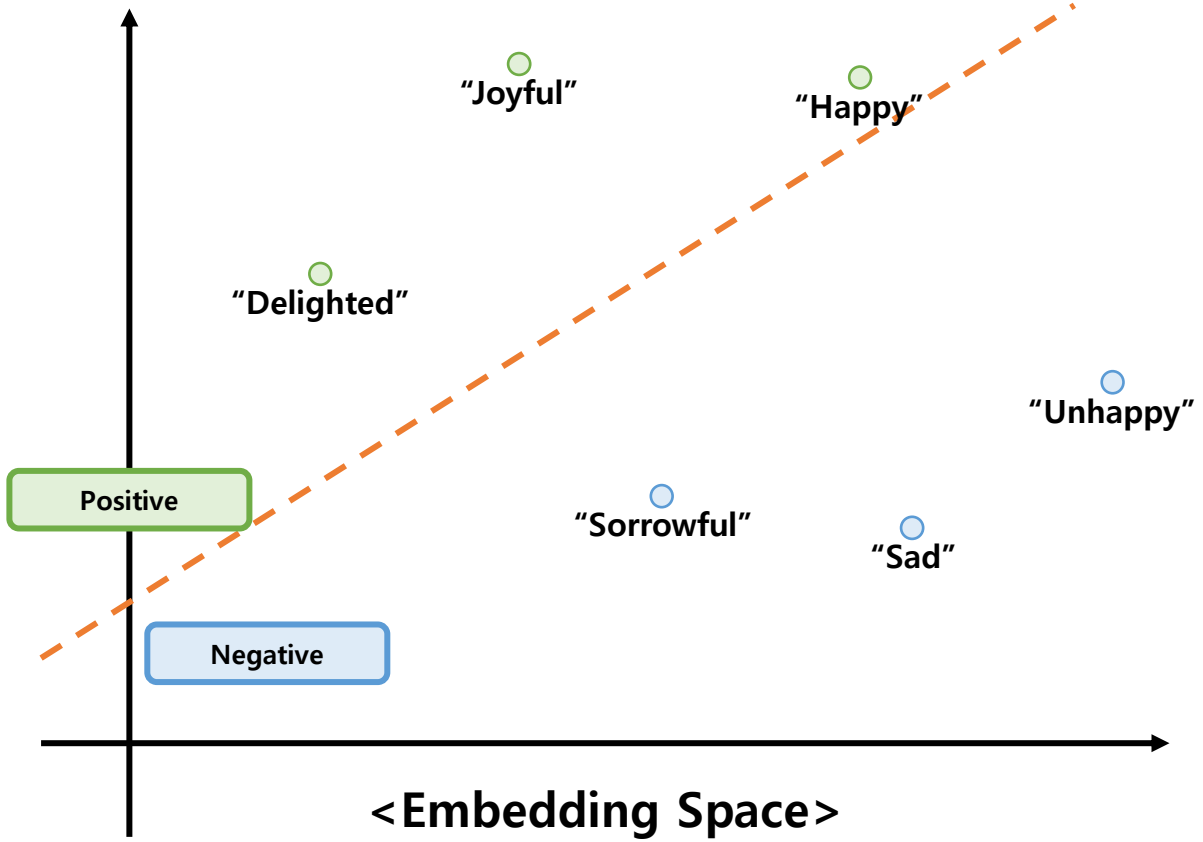
✓ 연구 흐름도



<적대적 훈련 기반의 텍스트 임베딩 증강>

✓ Step 1. 적대적 훈련을 통한 적대적 예제 생성

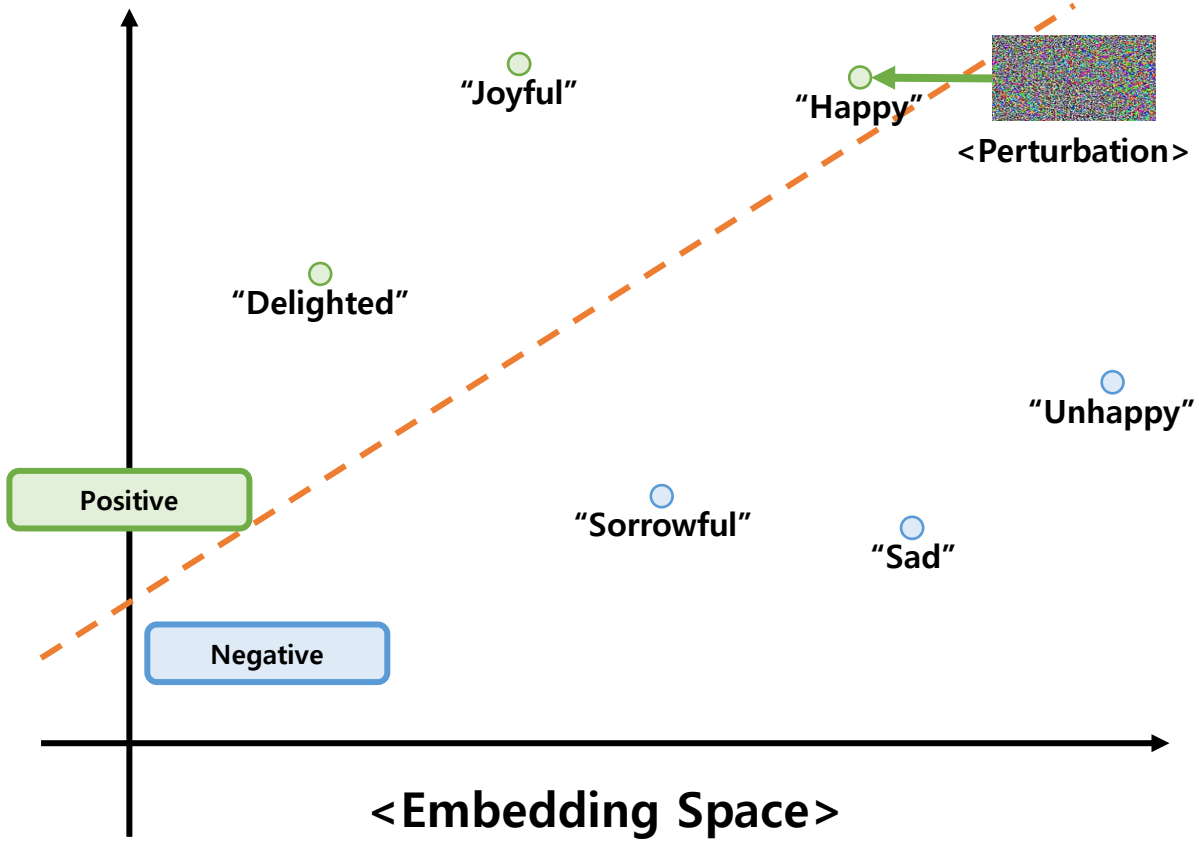
- PGD기반 적대적 훈련을 수행, 임베딩을 대상으로 언어 모델의 Loss를 증가시키도록 Perturbation을 학습
- 생성된 적대적 예제를 이용하여 언어 모델의 Loss를 감소시키는 방향으로 언어 모델 매개 변수를 학습



<적대적 훈련 기반의 텍스트 임베딩 증강>

✓ Step 1. 적대적 훈련을 통한 적대적 예제 생성

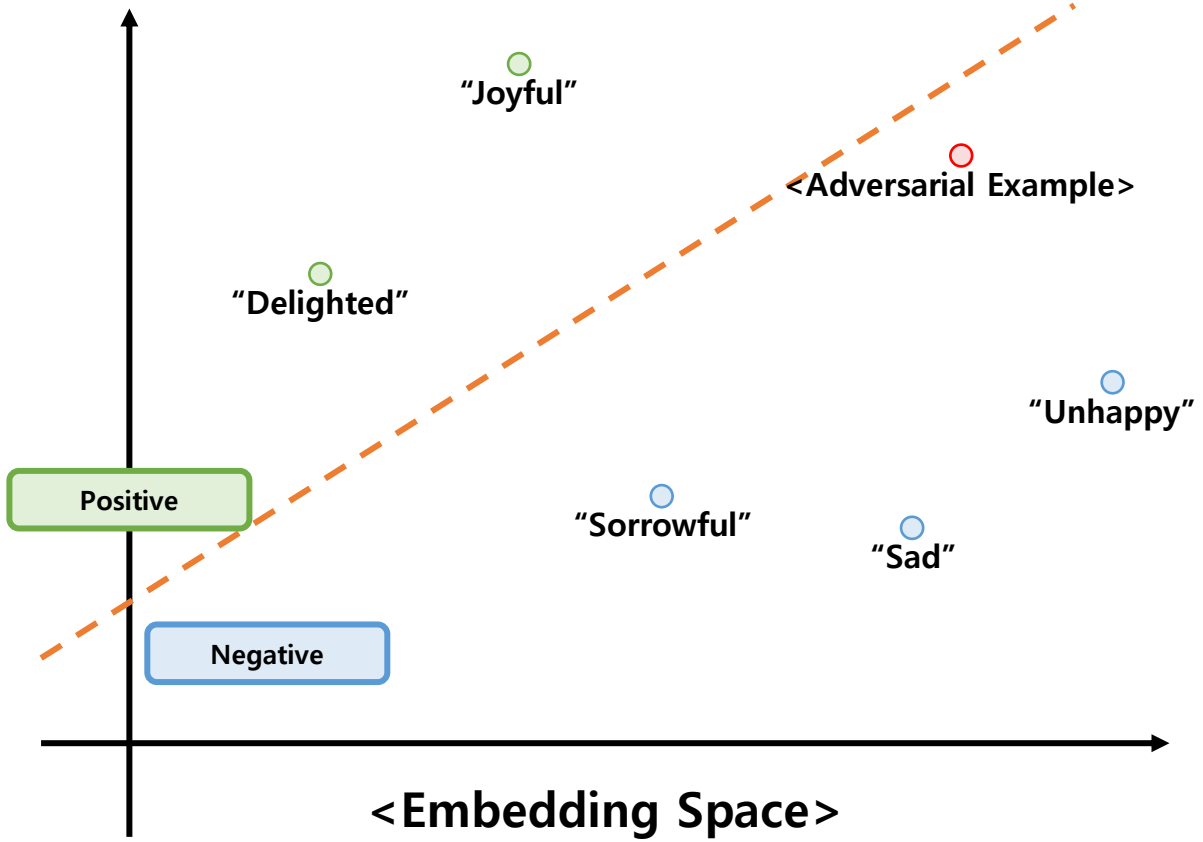
- PGD기반 적대적 훈련을 수행, 임베딩을 대상으로 언어 모델의 Loss를 증가시키도록 Perturbation을 학습
- 생성된 적대적 예제를 이용하여 언어 모델의 Loss를 감소시키는 방향으로 언어 모델 매개 변수를 학습



<적대적 훈련 기반의 텍스트 임베딩 증강>

✓ Step 1. 적대적 훈련을 통한 적대적 예제 생성

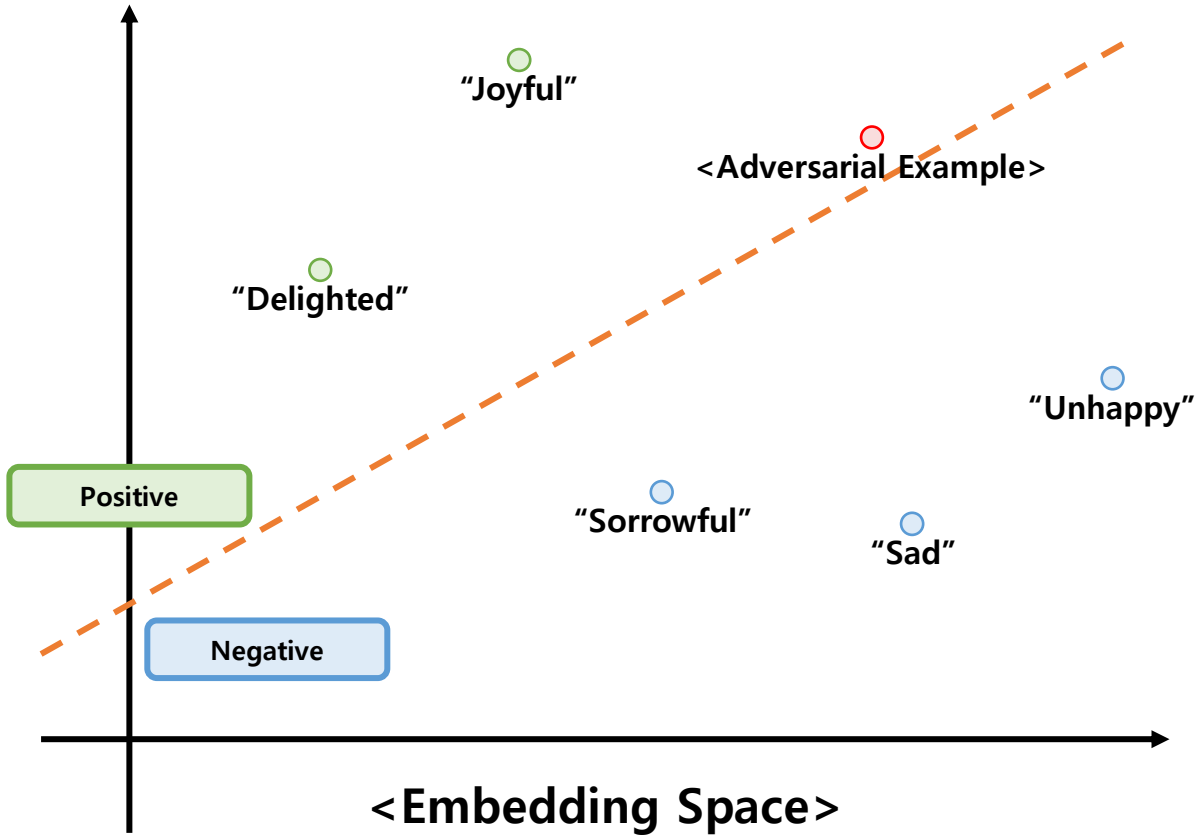
- PGD기반 적대적 훈련을 수행, 임베딩을 대상으로 언어 모델의 Loss를 증가시키도록 Perturbation을 학습
- 생성된 적대적 예제를 이용하여 언어 모델의 Loss를 감소시키는 방향으로 언어 모델 매개 변수를 학습



<적대적 훈련 기반의 텍스트 임베딩 증강>

✓ Step 1. 적대적 훈련을 통한 적대적 예제 생성

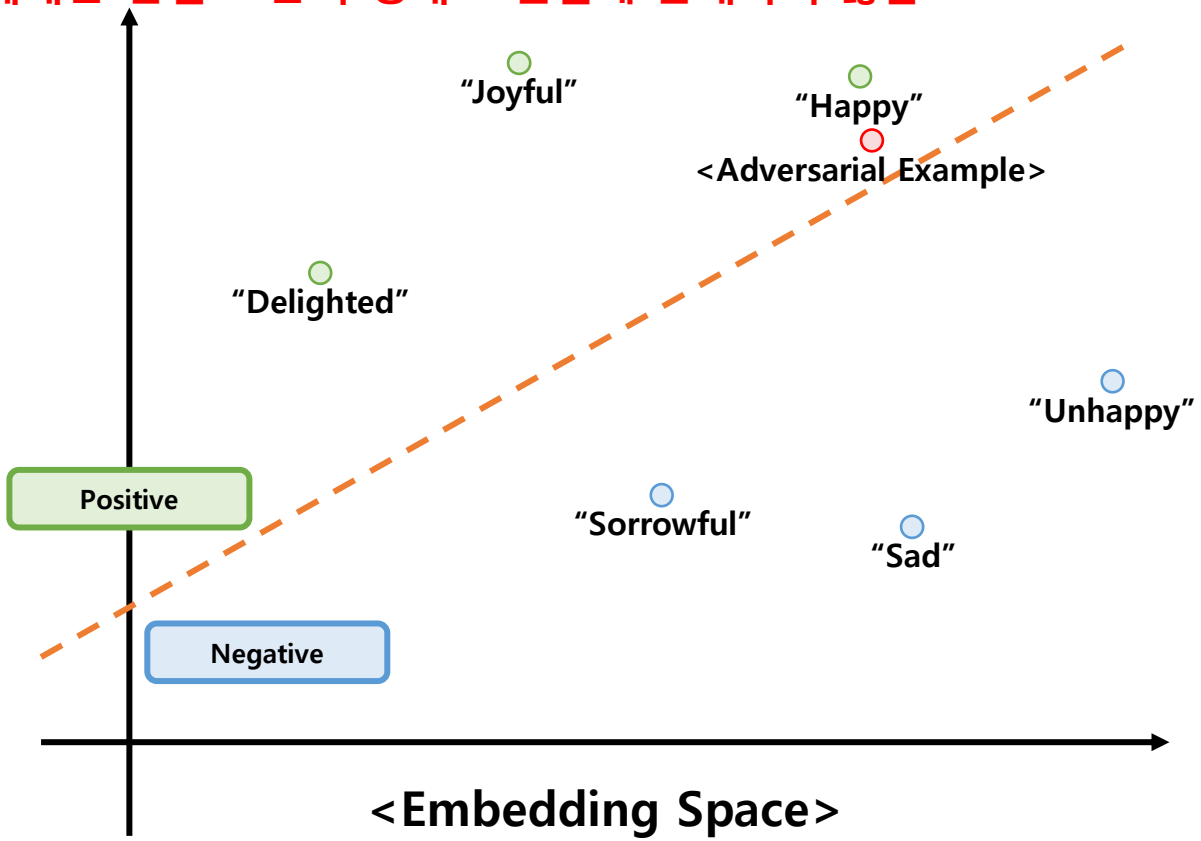
- PGD기반 적대적 훈련을 수행, 임베딩을 대상으로 언어 모델의 Loss를 증가시키도록 Perturbation을 학습
- 생성된 적대적 예제를 이용하여 언어 모델의 Loss를 감소시키는 방향으로 언어 모델 매개 변수를 학습



<적대적 훈련 기반의 텍스트 임베딩 증강>

✓ Step 1. 적대적 훈련을 통한 적대적 예제 생성

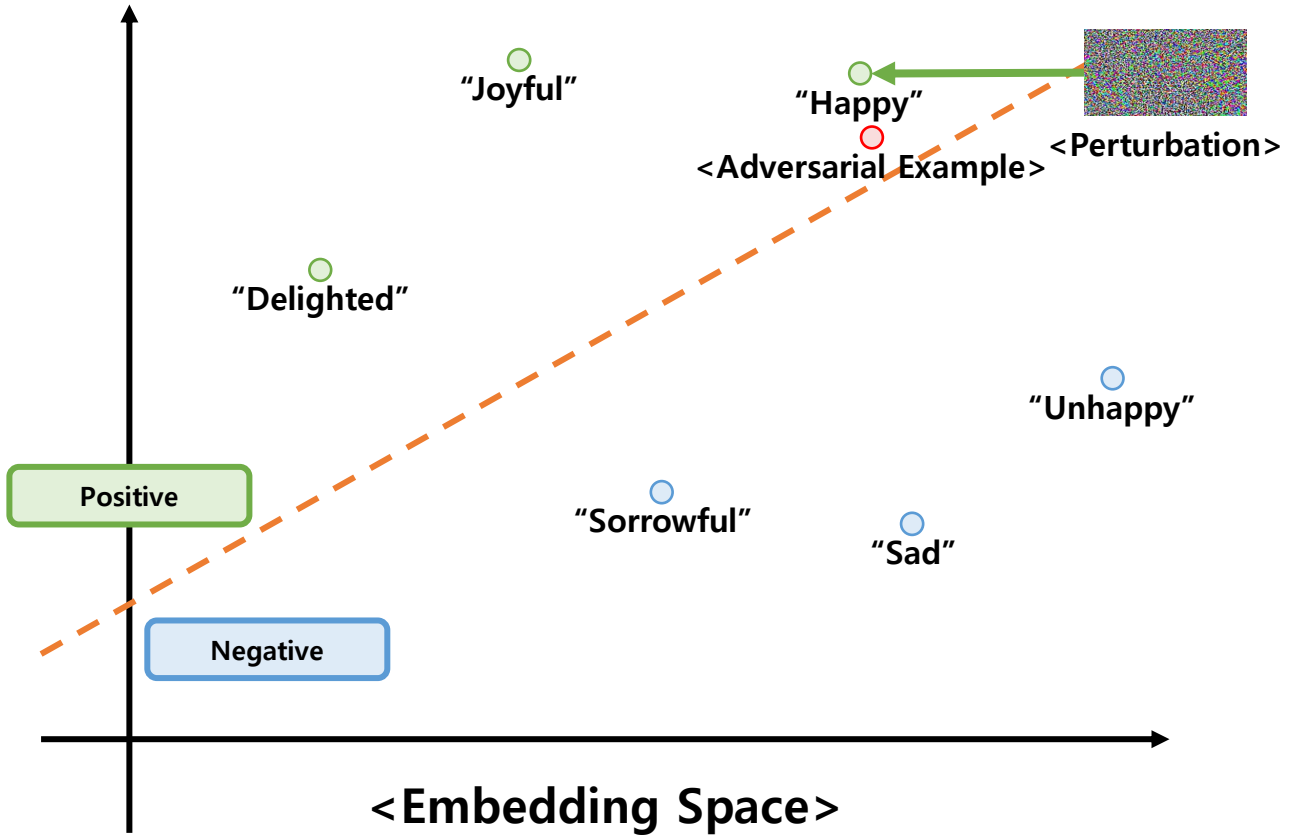
- 적대적 훈련 과정을 반복하여 임베딩 공간 내에서 적대적 예제들을 생성
- 생성된 적대적 예제들을 텍스트 임베딩 증강으로 활용
- 이 때, **적대적 예제는 단일 토큰의 형태로 현실에 존재하지 않음**



<적대적 훈련 기반의 텍스트 임베딩 증강>

✓ Step 1. 적대적 훈련을 통한 적대적 예제 생성

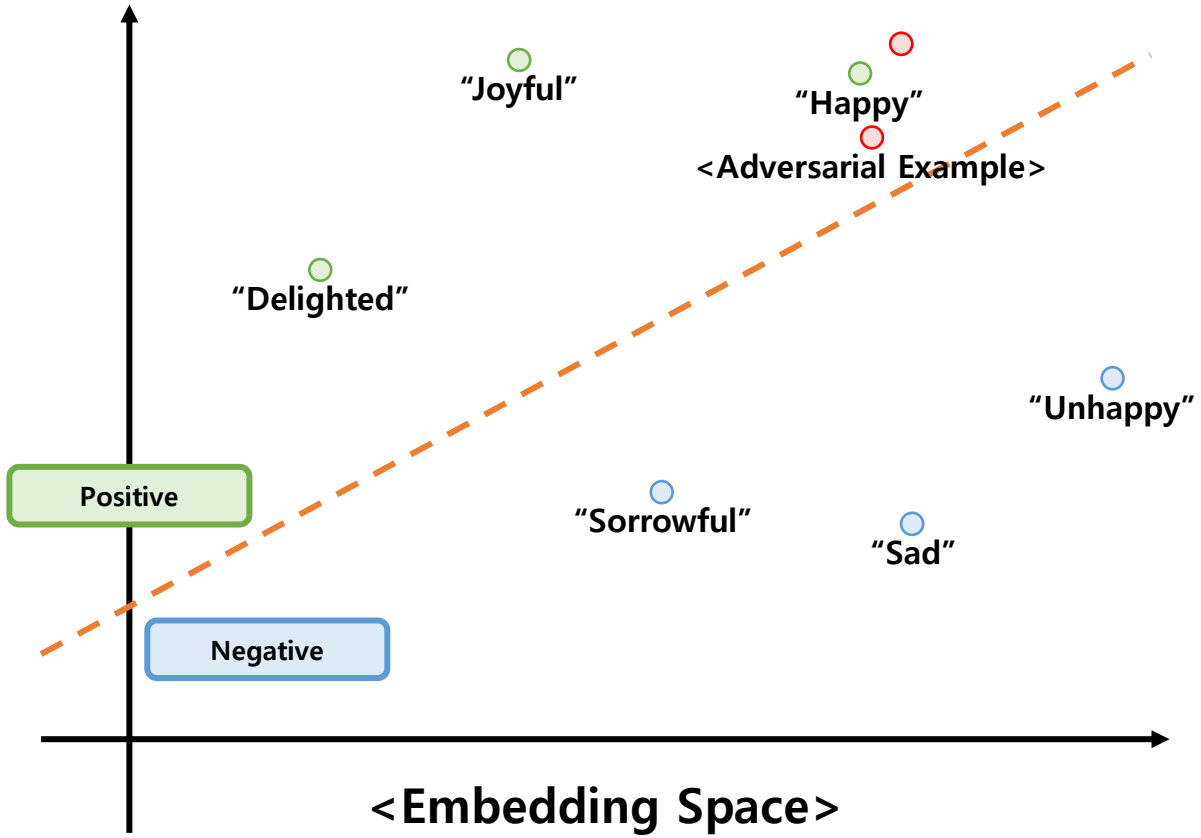
- 적대적 훈련 과정을 반복하여 임베딩 공간 내에서 적대적 예제들을 생성
- 생성된 적대적 예제들을 텍스트 임베딩 증강으로 활용



<적대적 훈련 기반의 텍스트 임베딩 증강>

✓ Step 1. 적대적 훈련을 통한 적대적 예제 생성

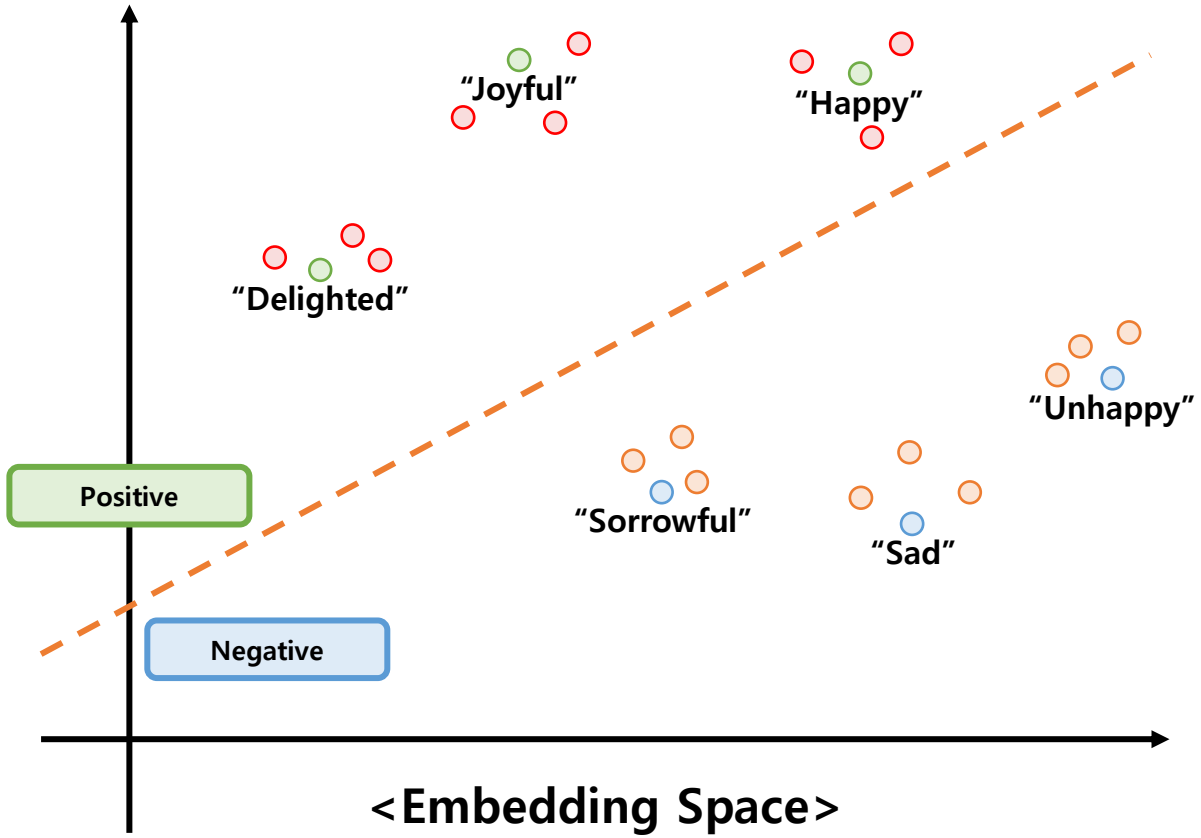
- 적대적 훈련 과정을 반복하여 임베딩 공간 내에서 적대적 예제들을 생성
- 생성된 적대적 예제들을 텍스트 임베딩 증강으로 활용



<적대적 훈련 기반의 텍스트 임베딩 증강>

✓ Step 1. 적대적 훈련을 통한 적대적 예제 생성

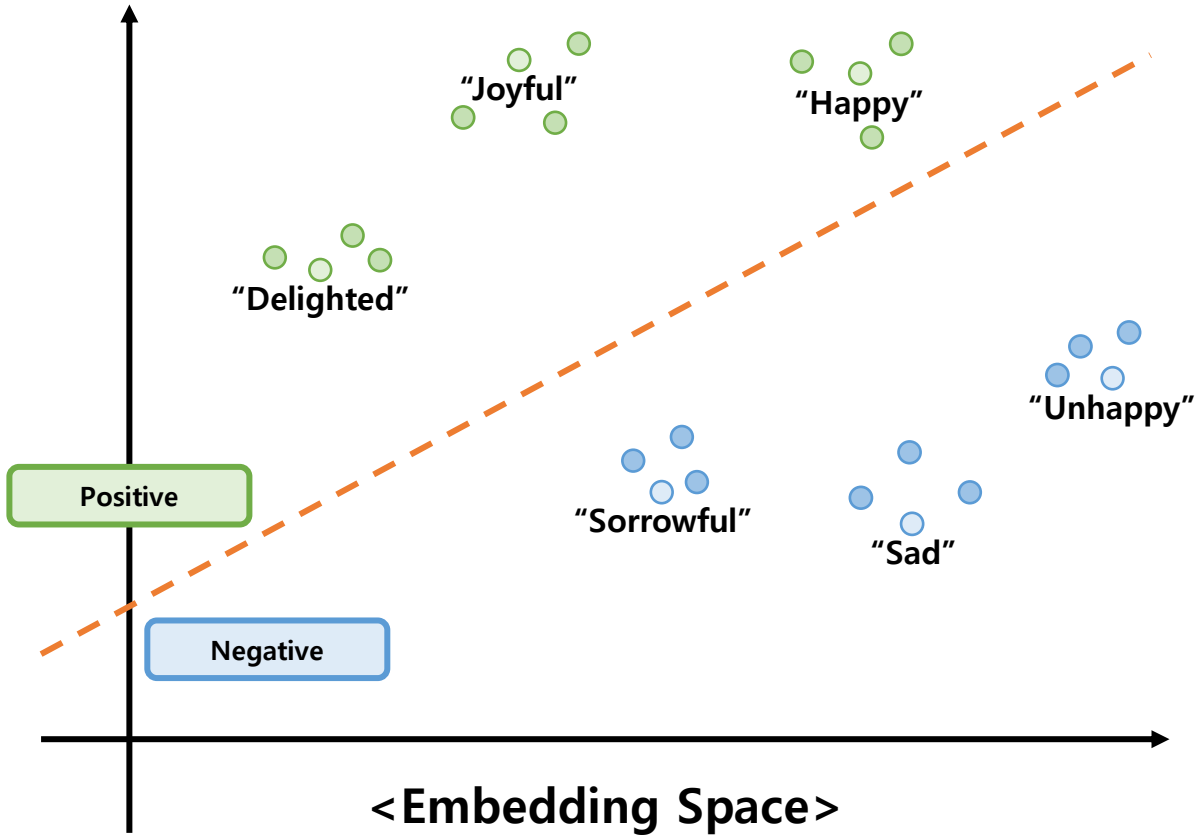
- 적대적 훈련 과정을 반복하여 임베딩 공간 내에서 적대적 예제들을 생성
- 생성된 적대적 예제들을 텍스트 임베딩 증강으로 활용



<적대적 훈련 기반의 텍스트 임베딩 증강>

✓ Step 1. 적대적 훈련을 통한 적대적 예제 생성

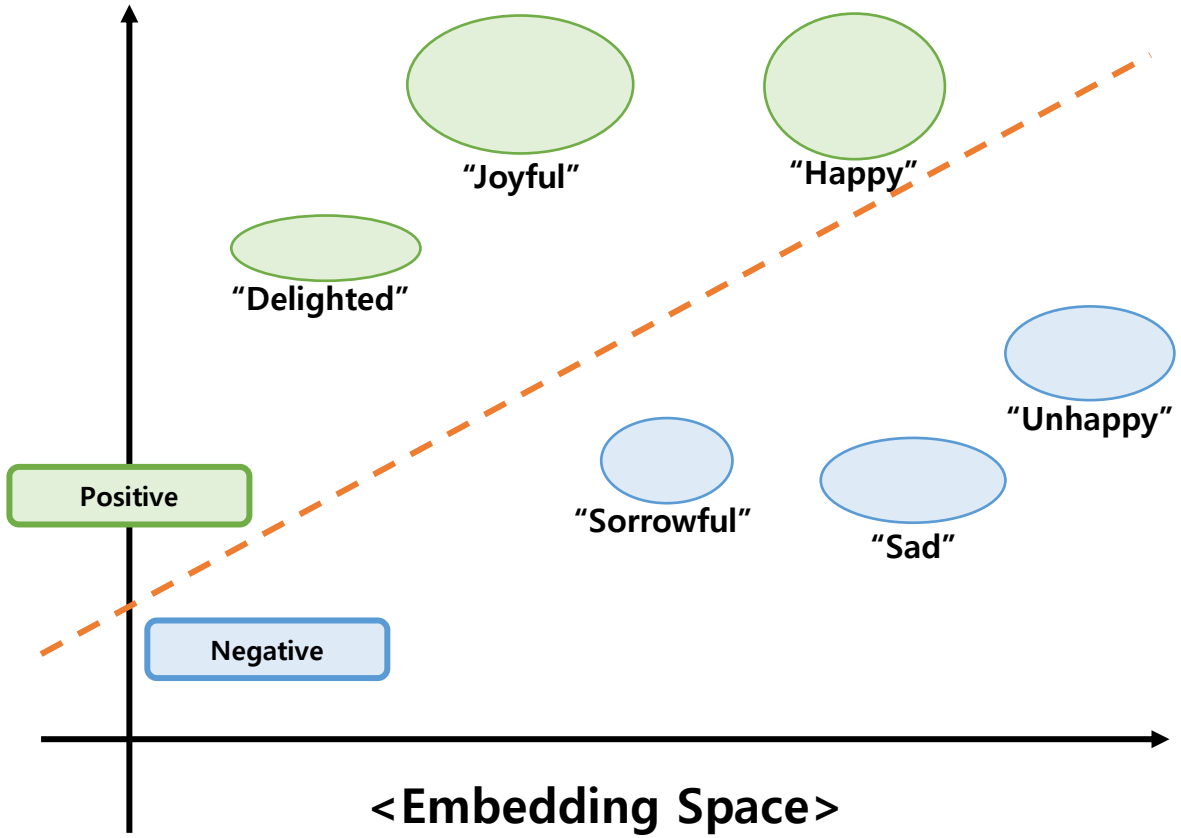
- PGD 기반 적대적 훈련을 통해 생성된 적대적 예제들은 원본 임베딩을 중심으로 매우 작은 공간 내에서 생성
- 따라서 생성된 적대적 예제(임베딩 증강)의 정답은 원본 자연어 토큰과 동일



<적대적 훈련 기반의 텍스트 임베딩 증강>

✓ Step 1. 적대적 훈련을 통한 적대적 예제 생성

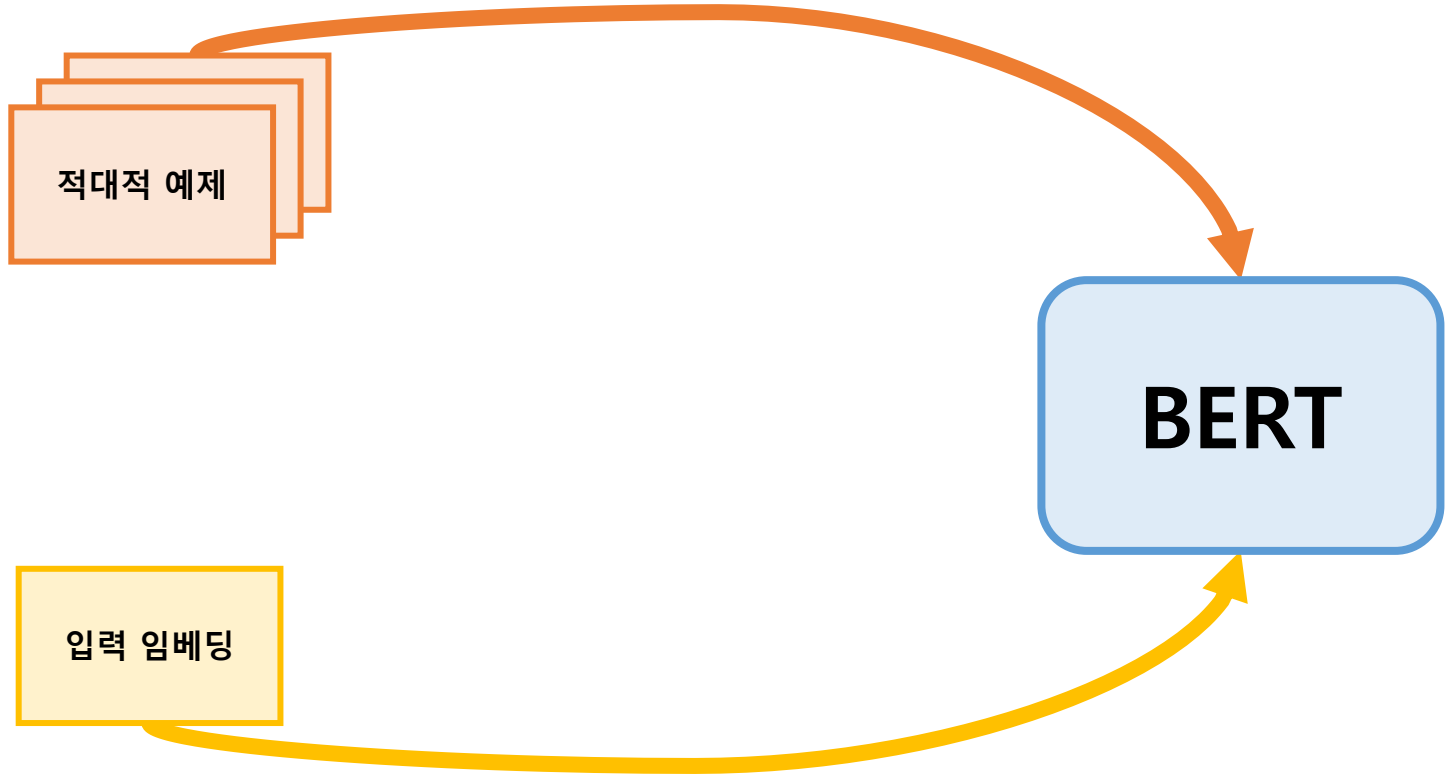
- 생성된 임베딩들은 현실에는 존재하지 않으나, 임베딩 공간 내에서 다양한 동의어들에 대해 생성된 임베딩 사이의 의미적 공간을 메우는 역할을 수행, 결과적으로 보다 강건하고 일반화된 임베딩을 형성할 수 있음



<적대적 훈련 기반의 텍스트 임베딩 증강>

✓ Step 2. 적대적 예제와 원본 데이터를 이용한 미세 조정

- 증강된 텍스트 임베딩과 원본 입력 임베딩을 함께 사용하여 목표 과업에 대해 미세 조정을 수행

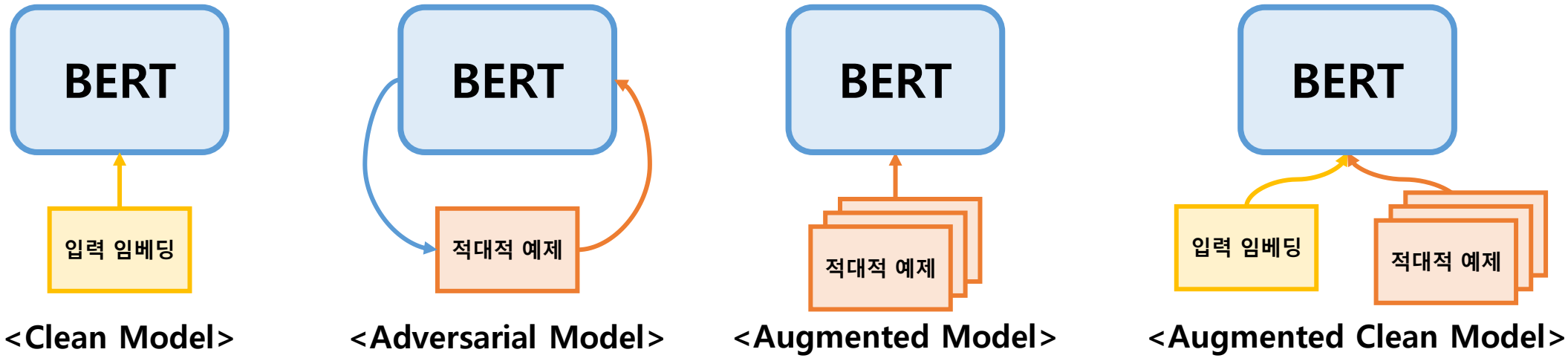


<Fine-Tuning>

<실험 방법>

✓ **실험 방법** (Backbone & Baseline)

- 총 4개의 모델에 대해 성능을 비교



<실험 방법>

✓ **실험 방법** (Backbone & Baseline)

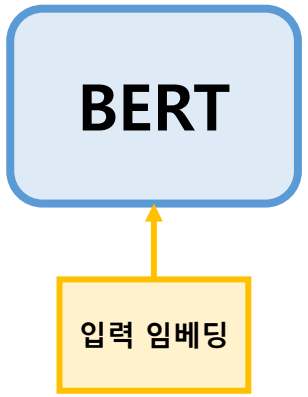
- 총 4개의 모델에 대해 성능을 비교

✓ **Clean Model** (Backbone)

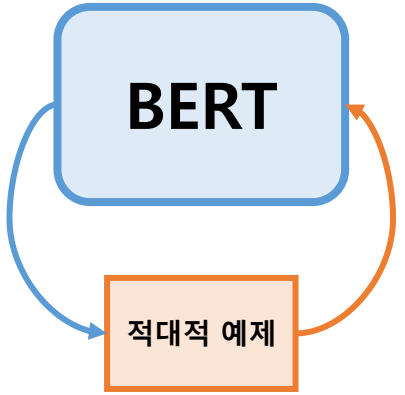
- 원본 입력 임베딩 만을 사용하여 미세 조정을 수행한 모델

✓ **Adversarial Model** (Baseline)

- 미세 조정 과정에서 적대적 훈련을 수행한 모델



<Clean Model>



<Adversarial Model>

<실험 방법>

✓ **실험 방법** (Proposed)

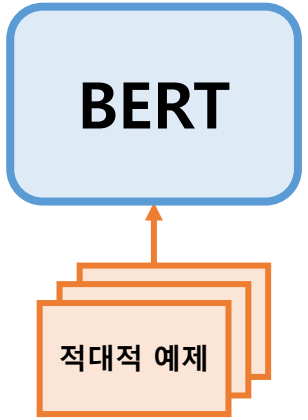
- 총 4개의 모델에 대해 성능을 비교

✓ **Augmented Model** (Proposed)

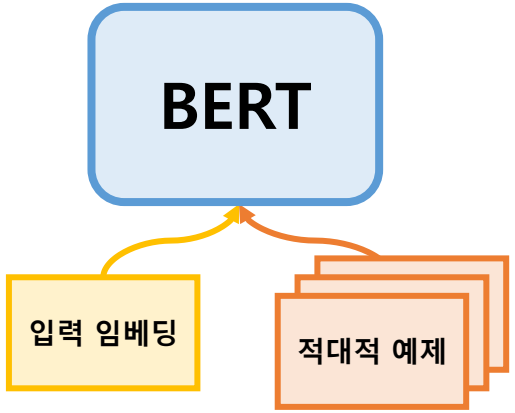
- Adversarial Model을 학습하는 과정에서 생성된 적대적 예제를 이용하여 미세 조정을 수행한 모델

✓ **Augmented Clean Model** (Proposed)

- Augmented Model에 Clean Data로 추가적인 미세 조정을 수행한 모델



<Augmented Model>



<Augmented Clean Model>

<실험 결과>

Dataset	Model	Test Accuracy	Increase to Backbone	Increase to Baseline	Note
IMDB	**Clean** **($BERT_{base}$)**	92.68	-	-	Backbone
	Adversarial **(PGD)**	92.99	+0.31%p	-	Baseline
	Augmented **(PGD)**	**93.17**	**+0.49%p**	**+0.18%p**	**Proposed**
	Augmented Clean **(PGD)**	92.95	+0.27%p	-0.04%p	**Proposed**

<실험 결과>

Dataset	Model	Dev Accuracy	Increase to Backbone	Increase to Baseline	Note
SST-2	**Clean** **($BERT_{base}$)**	91.97	-	-	Backbone
	Adversarial **(PGD)**	91.74	-0.23%p	-	Baseline
	Augmented **(PGD)**	**92.66**	**+0.69%p**	**+0.92%p**	**Proposed**
	Augmented Clean **(PGD)**	91.85	-0.12%p	+0.11%p	**Proposed**

<결론>

✓ 결론

- **적대적 훈련을 이용하여 2 Step으로 자연어 임베딩을 증강하는 방법을 제안함**
  - Step 1. Target Task 데이터를 이용한 적대적 훈련으로 적대적 예제를 생성
  - Step 2. 적대적 예제와 원본 데이터를 이용하여 미세 조정

- IMDB 데이터를 기준으로 FreeLB 논문에서 제안된 **PGD 적대적 훈련 방법으로 산출된 성능을 상회하는 성능**을 확인하였으며, 별도의 **매개 변수 추가 없이 언어 모델의 성능을 향상**시킬 수 있음

- 일시적으로 적대적 예제를 생성하여 학습에 적용하는 방법보다, **적대적 훈련을 수행하며 생성된 적대적 예제를 저장한 뒤, 별도의 모델에 반복 학습하는 방법**이 **성능적으로 우수함**을 확인

<향후 연구>

✓ Parameter Tuning을 통해 추가적으로 성능이 향상되는지 확인할 예정

- 시간의 제약으로 인해 실험된 모든 성능은 Parameter의 탐색이 면밀하게 수행되지 않은 결과에 해당
- 추후 다양한 Parameter에 대해 실험을 수행할 예정

✓ 다른 데이터에도 제안된 임베딩 증강의 효과가 유효한지에 관한 실험 수행 예정

- 현재 실험 결과는 IMDB, SST-2 데이터만을 사용하여 산출되었음, 보다 다양한 데이터셋에 대해 실험을 수행할 예정

Thank You