GAN-BERT: Generative Adversarial Learning for Robust Text Classification with a Bunch of Labeled Examples

Danilo Croce, Giuseppe Castellucci, Roberto Basili

ACL 2020

1. Introduction

- Pretrained model BERT는 적절한 양의 labeled 데이터를 사용해 fine-tuning을 해야 좋은 결과를 얻음
  - 하지만, 현실에선 적당한 양의 labeled data를 얻기 힘듦

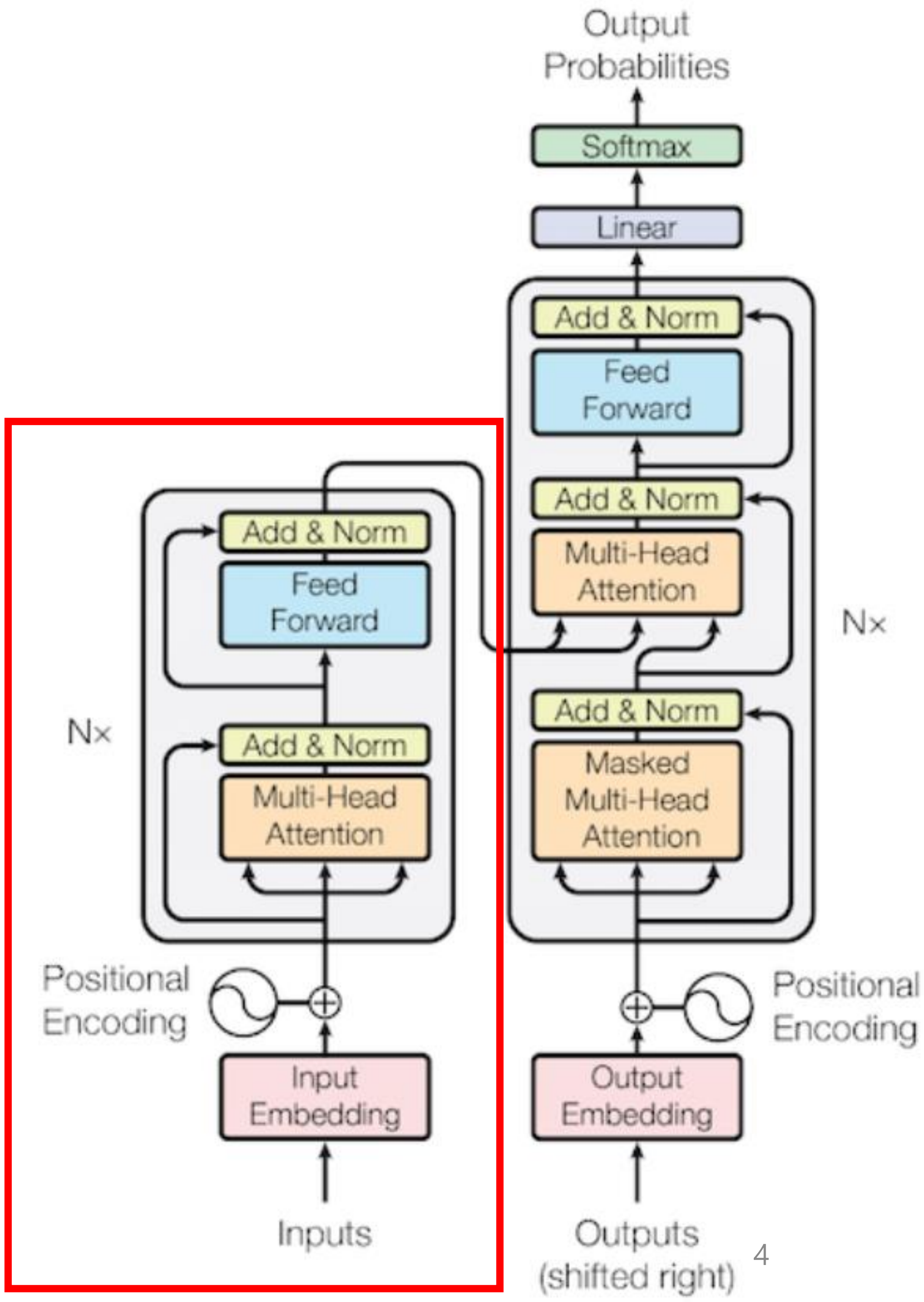
- 적은 labeled data와 많은 unlabeled data로 BERT를 fine-tuning 하기 위해 Semi-Supervised GAN를 이용함

2. Background

- BERT
- GAN
- Improved GAN : Feature Matching
- Semi-Supervised GAN

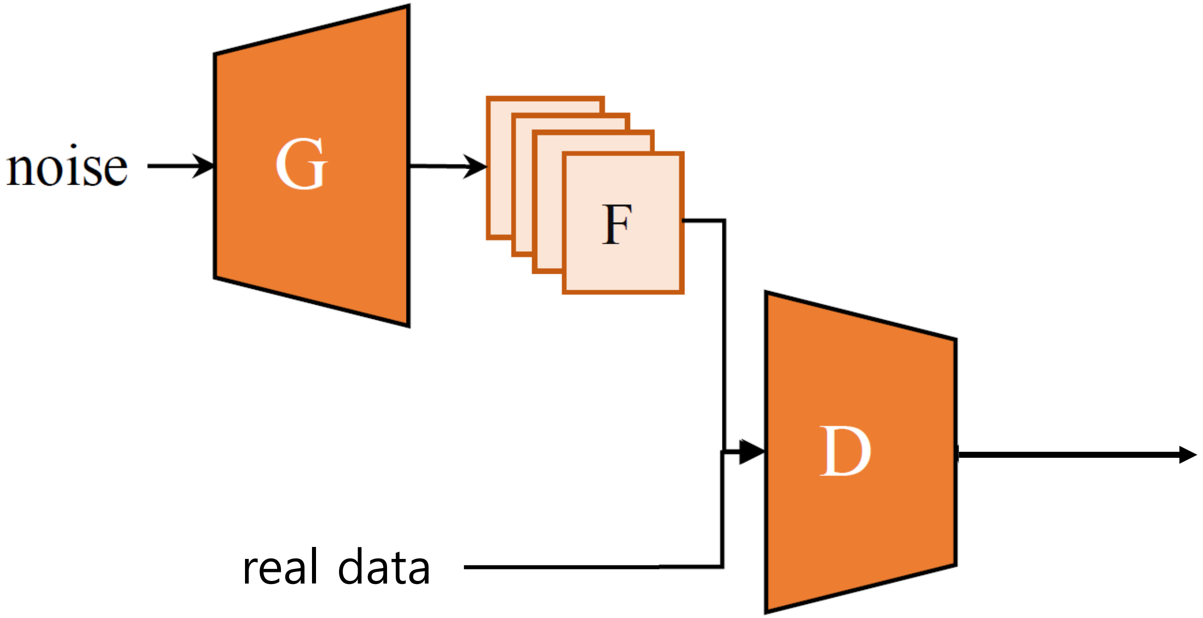
2. Background

- BERT
  - Transformer based pretrained model
  - Encoder로만 구성됨
  - Contextual word embedding
    - 같은 단어라도 문맥에 따라 표현방법이 바뀜



2. Background

- GAN
  - training without labels (unsupervised learning)



  - Generator : try to generate fake data to be real
  - Discriminator : try to discriminate fake data

2. Background

- Improved GAN : Feature Matching
  - Original generator loss

$$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log\left(1 - D\left(G\left(z^{(i)}\right)\right)\right).$$

  - Discriminator를 overtraining하는 것을 방지하기 위해 generator에 새로운 objective function 추가

$$||\mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}} \mathbf{f}(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{z} \sim p_{\boldsymbol{z}}(\boldsymbol{z})} \mathbf{f}(G(\boldsymbol{z}))||_2^2$$

    - f(x) : activations on an intermediate layer of the discriminator
    - x : real data, z : noise

  - Generator가 real data와 비슷한 데이터를 만들 수 있게 함

2. Background

- Semi-Supervised GAN
  - standard multi-class classifier (supervised learning)
    - Cross-entropy on probability

2. Background

- Semi-Supervised GAN
  - Adding samples from the GAN generator G, labeling with a new "generated" class y = K + 1

$$p_{\text{model}}(y = K + 1 \mid \boldsymbol{x})$$

$$L = -\mathbb{E}_{\boldsymbol{x},y\sim p_{\text{data}}(\boldsymbol{x},y)}[\log p_{\text{model}}(y|\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{x}\sim G}[\log p_{\text{model}}(y = K+1|\boldsymbol{x})]$$

$$= L_{\text{supervised}} + L_{\text{unsupervised}}$$

$$L_{\text{supervised}} = -\mathbb{E}_{\boldsymbol{x},y\sim p_{\text{data}}(\boldsymbol{x},y)} \log p_{\text{model}}(y|\boldsymbol{x}, y < K+1)$$

2. Background

- Semi-Supervised GAN
  - Adding samples from the GAN generator G, labeling with a new "generated" class y = K + 1

$$p_{\text{model}}(y = K+1 \mid \boldsymbol{x})$$

$$L = -\mathbb{E}_{\boldsymbol{x},y\sim p_{\text{data}}(\boldsymbol{x},y)}[\log p_{\text{model}}(y|\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{x}\sim G}[\log p_{\text{model}}(y = K+1|\boldsymbol{x})]$$

$$= L_{\text{supervised}} + L_{\text{unsupervised}}$$

$$L_{\text{supervised}} = -\mathbb{E}_{\boldsymbol{x},y\sim p_{\text{data}}(\boldsymbol{x},y)} \log p_{\text{model}}(y|\boldsymbol{x}, y < K+1)$$

$$L_{\text{unsupervised}} = -\{\underline{\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})} \log[1 - p_{\text{model}}(y = K+1|\boldsymbol{x})]} + \mathbb{E}_{\boldsymbol{x}\sim G} \log[p_{\text{model}}(y = K+1|\boldsymbol{x})]\}$$

Real data 중 unlabeled data가 k+1 label로 분류되지 않게 함

2. Background

- Semi-Supervised GAN
  - Adding samples from the GAN generator G, labeling with a new "generated" class y = K + 1

$$p_{\text{model}}(y = K + 1 \mid \boldsymbol{x})$$

$$L = -\mathbb{E}_{\boldsymbol{x},y\sim p_{\text{data}}(\boldsymbol{x},y)}[\log p_{\text{model}}(y|\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{x}\sim G}[\log p_{\text{model}}(y = K+1|\boldsymbol{x})]$$

$$= L_{\text{supervised}} + L_{\text{unsupervised}}$$

$$L_{\text{supervised}} = -\mathbb{E}_{\boldsymbol{x},y\sim p_{\text{data}}(\boldsymbol{x},y)} \log p_{\text{model}}(y|\boldsymbol{x}, y < K+1)$$

$$L_{\text{unsupervised}} = -\{\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})} \log[1 - p_{\text{model}}(y = K+1|\boldsymbol{x})] + \underline{\mathbb{E}_{\boldsymbol{x}\sim G} \log[p_{\text{model}}(y = K+1|\boldsymbol{x})]}\}$$

fake data가 k+1 label로 분류되게 함

2. Background

- Semi-Supervised GAN
  - Adding samples from the GAN generator G, labeling with a new "generated" class y = K + 1

$$p_{\text{model}}(y = K + 1 \mid \boldsymbol{x})$$

$$L = -\mathbb{E}_{\boldsymbol{x},y\sim p_{\text{data}}(\boldsymbol{x},y)}[\log p_{\text{model}}(y|\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{x}\sim G}[\log p_{\text{model}}(y = K+1|\boldsymbol{x})]$$

$$= L_{\text{supervised}} + L_{\text{unsupervised}}$$

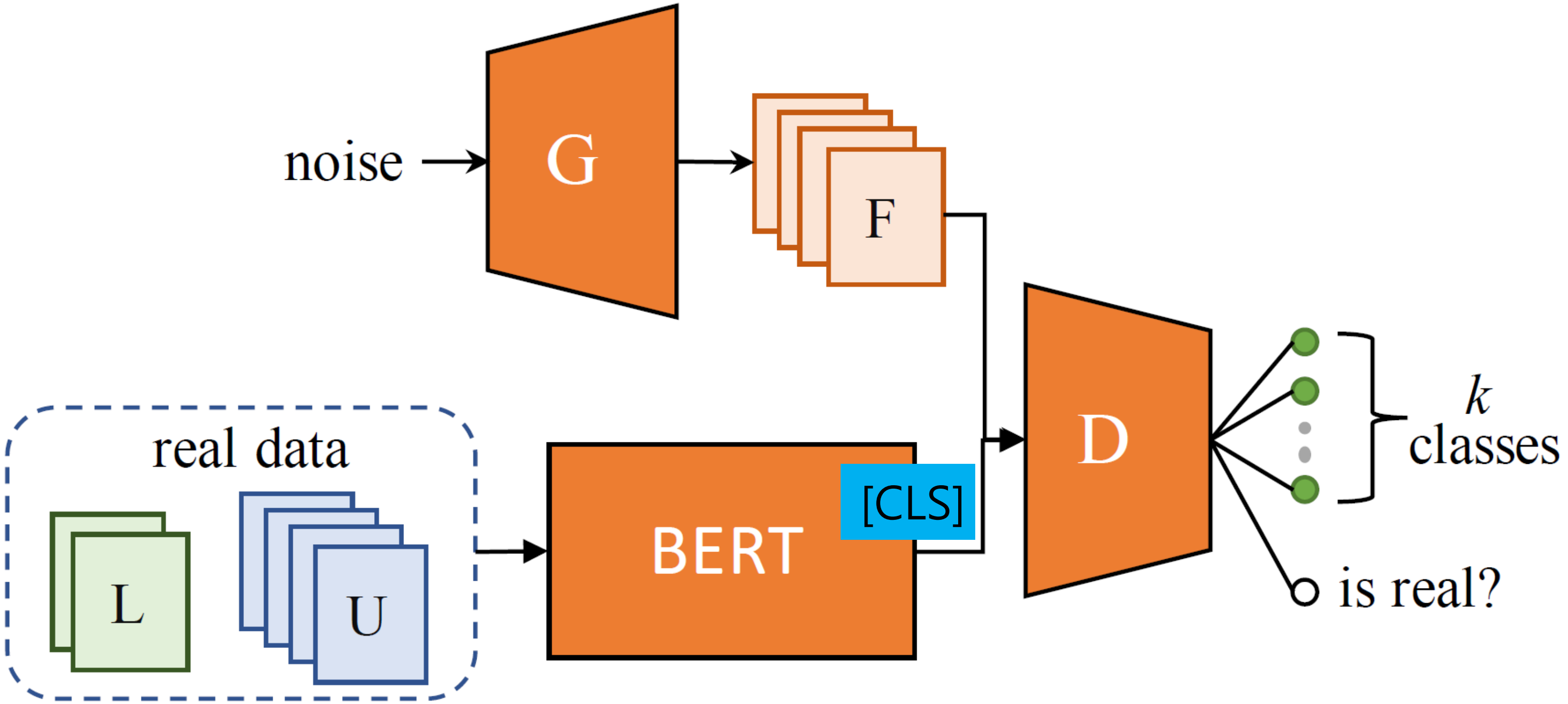
$$L_{\text{supervised}} = -\mathbb{E}_{\boldsymbol{x},y\sim p_{\text{data}}(\boldsymbol{x},y)} \log p_{\text{model}}(y|\boldsymbol{x}, y < K+1)$$

$$L_{\text{unsupervised}} = -\{\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})} \log[1 - p_{\text{model}}(y = K+1|\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim G} \log[p_{\text{model}}(y = K+1|\boldsymbol{x})]\}$$

  - cross-entropy loss -> *L*supervised & *L*unsupervised
  - replace $1 - p_{\text{model}}(y = K + 1 \mid \boldsymbol{x})$ to *D(x)*

$$L_{\text{unsupervised}} = -\{\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})} \log D(\boldsymbol{x}) + \mathbb{E}_{z\sim\text{noise}} \log(1 - D(G(\boldsymbol{z})))\}$$

3. Methodology



3. Methodology

- Discriminator Loss

$$= L_{\text{supervised}} + L_{\text{unsupervised}}$$

$$L_{\text{supervised}} = -\mathbb{E}_{\boldsymbol{x},y\sim p_{\text{data}}(\boldsymbol{x},y)} \log p_{\text{model}}(y|\boldsymbol{x}, y < K+1)$$

$$L_{\text{unsupervised}} = -\{\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})} \log[1 - p_{\text{model}}(y = K+1|\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim G} \log[p_{\text{model}}(y = K+1|\boldsymbol{x})]\}$$

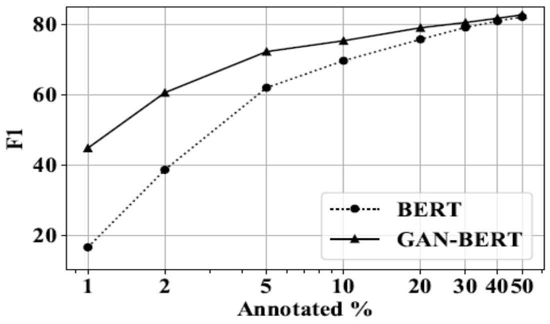
- Generator Loss

$$= L_{\mathcal{G}_{\text{feature matching}}} + L_{\mathcal{G}_{unsup.}}$$

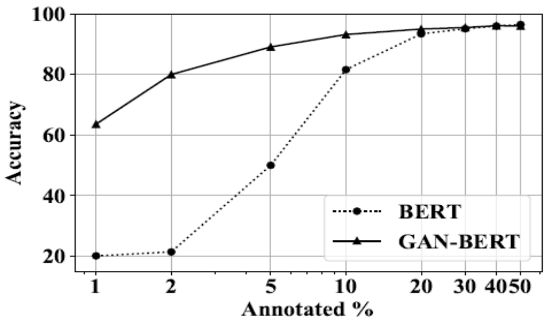
$$L_{G_{\text{feature matching}}} = \left\|\mathbb{E}_{x\sim p_d} f(x) - \mathbb{E}_{x\sim \mathcal{G}} f(x)\right\|_2^2$$

$$L_{\mathcal{G}_{unsup.}} = -\mathbb{E}_{x\sim\mathcal{G}} \log[1 - p_m(\hat{y} = y|x, y = k+1)]$$

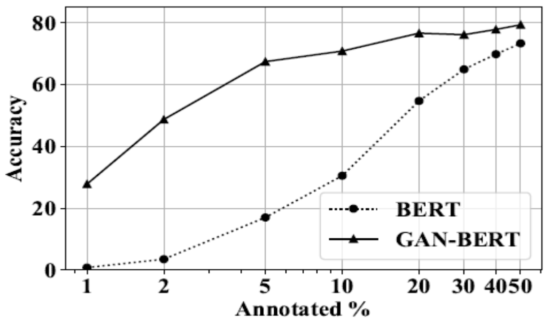
4. Results



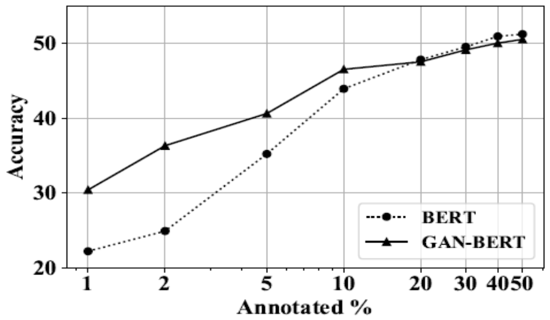
(a) 20N



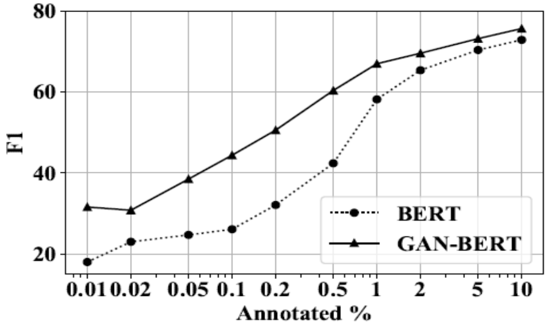
(b) QC Coarse Grained



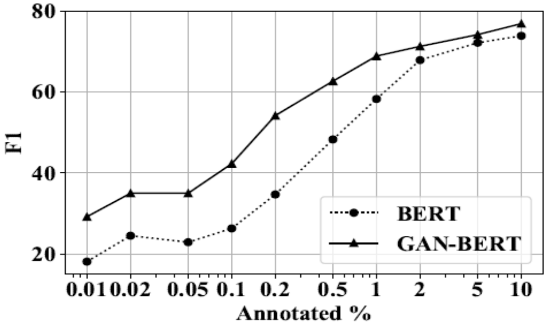
(c) QC Fine Grained



(d) SST-5



(e) MNLI Matched



(f) MNLI Mismatched

- Changed labeled data numbers
- Compared performances with BERT
- Dataset(Task)
  - 20N(news group classification), QC(question classification), SST5(sentiment classification), MNLI(textual entailment classification)

5. Conclusion

- 적은 labeled data와 많은 unlabeled data로 semi-supervised GAN을 사용해 BERT를 fine-tuning을 하여 좋은 결과를 얻음

Q&A