GAN-BERT

accepted by ACL

Danilo Croce, Giuseppe Castellicucci, Roberto Basili

채형주

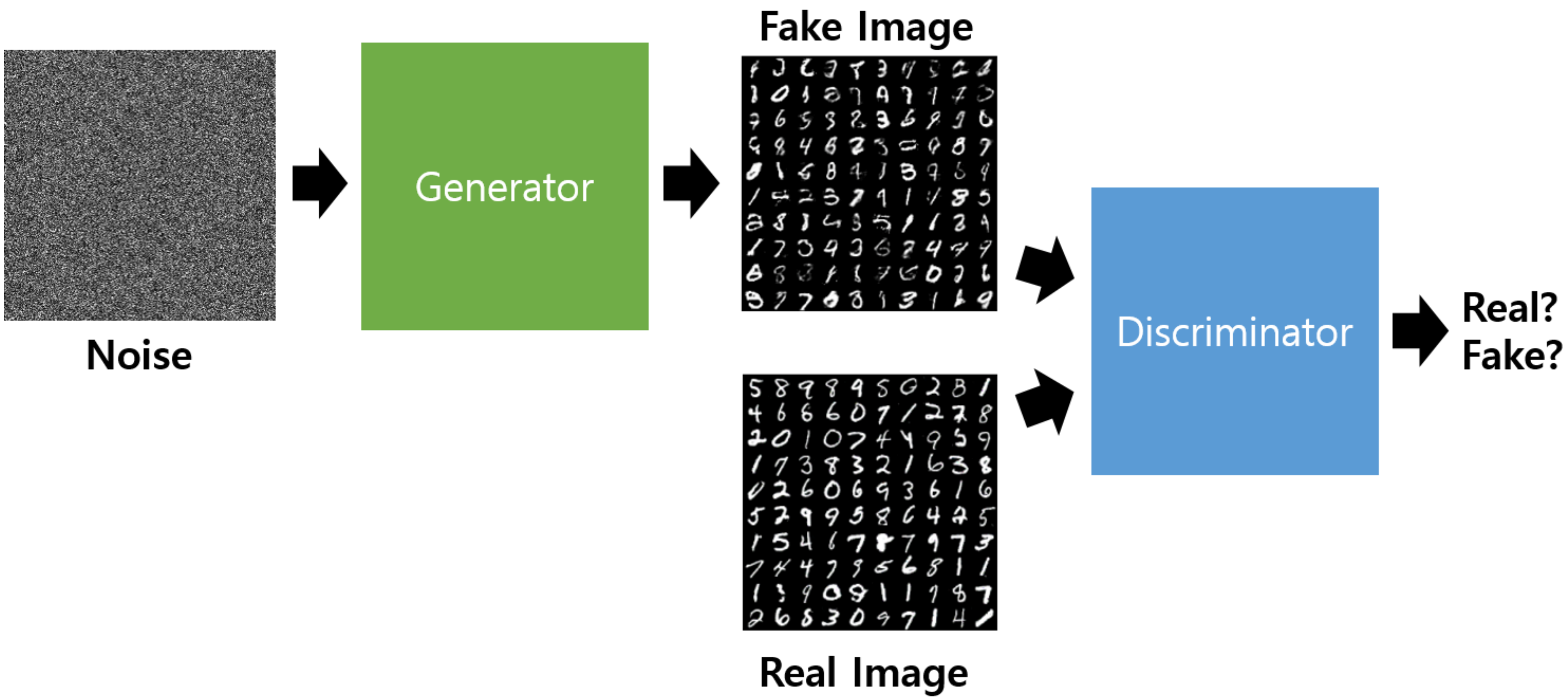
목차

- 제안 배경
- GAN
- SS-GAN
- GAN-BERT
- Result
- Conclusion

제안 배경

- 항상 annotated data는 부족하고, unlabeled data를 직접 labeling하기에는 상당한 비용이 들어간다.
- BERT는 좋은 성능을 내지만 fine-tuning을 위한 많은 양의 labeled data 필요
- BERT를 fine-tuning 할때 200개 이하의 annotated data instances로 학습시킨다면 성능의 상당한 저하가 온다.

GAN

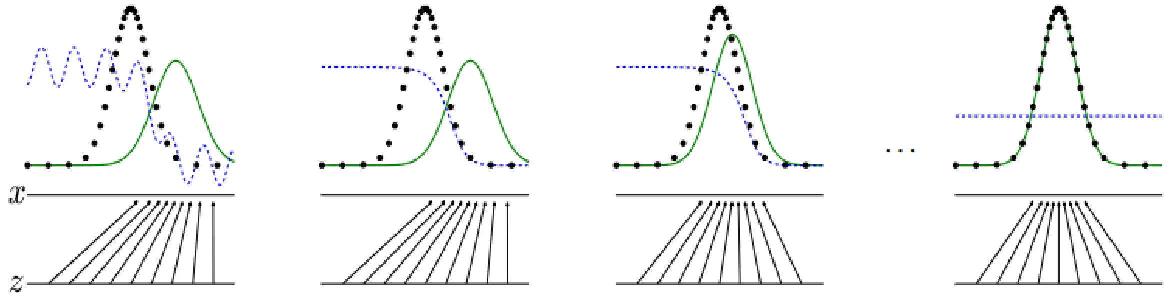


Fake Image
Noise
Generator
Discriminator
Real?
Fake?
Real Image

GAN : to find Nash Equilibrium

	Real Data	Fake Data
D == 1(진짜라고 판단)		Descriminator update
D == 0(가짜라고 판단)	Descriminator update	Generator update

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))]. \quad (1)$$



(a) (b) (c) (d)

4.1 Global Optimality of $p_g = p_{\text{data}}$

We first consider the optimal discriminator D for any given generator G.

Proposition 1. *For G fixed, the optimal discriminator D is*

$$D^*_G(\boldsymbol{x}) = \frac{p_{data}(\boldsymbol{x})}{p_{data}(\boldsymbol{x}) + p_g(\boldsymbol{x})}$$

SS-GAN : semi supervised GAN

- 기존 GAN은 내쉬 균형으로 수렴이 잘 안되는 문제가 있음 (non-convex, large parameters)
- feature matching
$$||\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}\mathbf{f}(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}\mathbf{f}(G(\boldsymbol{z}))||_2^2$$
- semi-supervised learning

$$\begin{aligned}
L &= -\mathbb{E}_{\boldsymbol{x},y\sim p_{\text{data}}(\boldsymbol{x},y)}[\log p_{\text{model}}(y|\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{x}\sim G}[\log p_{\text{model}}(y=K+1|\boldsymbol{x})] \\
&= L_{\text{supervised}} + L_{\text{unsupervised}}, \text{ where} \\
L_{\text{supervised}} &= -\mathbb{E}_{\boldsymbol{x},y\sim p_{\text{data}}(\boldsymbol{x},y)} \log p_{\text{model}}(y|\boldsymbol{x}, y < K+1) \\
L_{\text{unsupervised}} &= -\{\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})} \log[1 - p_{\text{model}}(y=K+1|\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim G} \log[p_{\text{model}}(y=K+1|\boldsymbol{x})]\},
\end{aligned}$$

SS-GAN : semi supervised GAN

Figure 3: *(Left)* samples generated by model during semi-supervised training. Samples can be clearly distinguished from images coming from MNIST dataset. *(Right)* Samples generated with minibatch discrimination. Samples are completely indistinguishable from dataset images.

Model	Number of incorrectly predicted test examples for a given number of labeled samples			
	20	50	100	200
DGN [21]			333 ± 14	
Virtual Adversarial [22]			212	
CatGAN [14]			191 ± 10	
Skip Deep Generative Model [23]			132 ± 7	
Ladder network [24]			106 ± 37	
Auxiliary Deep Generative Model [23]			96 ± 2	
Our model	1677 ± 452	221 ± 136	93 ± 6.5	90 ± 4.2
Ensemble of 10 of our models	1134 ± 445	142 ± 96	86 ± 5.6	81 ± 4.3

GAN-BERT

$$L_{D_{\text{supervised}}} = -\mathbb{E}_{x,\, y \sim p_{\text{data}}} \log[p_{\text{model}}(y = i|x,\, i < k+1)]$$

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

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

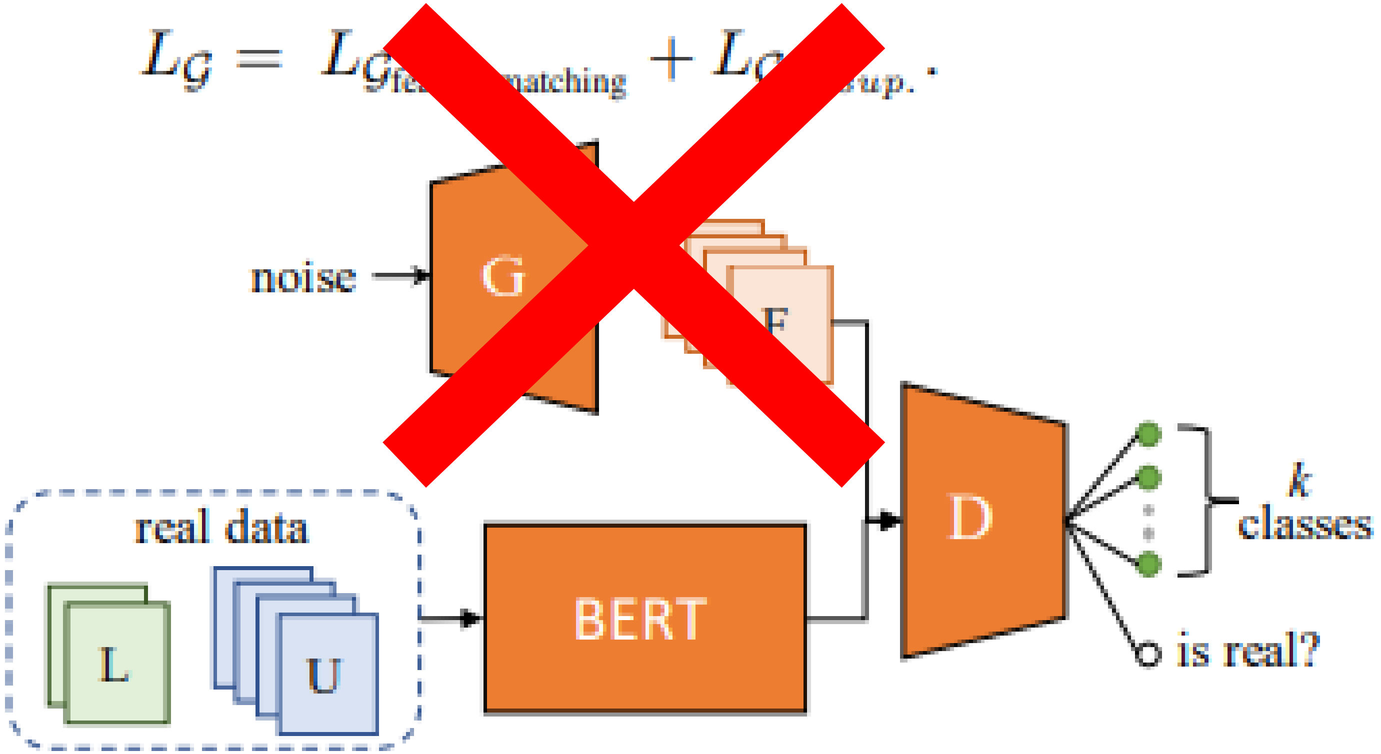
$$L_{\mathcal{G}_{unsup.}} = -\mathbb{E}_{x\sim\mathcal{G}} \log[1 - p_m(\hat{y} = y|x, y = k+1)]$$

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

- sentence classification task
- Generator
  - input : 100-d noise vector
  - MLP
  - output : h_fake (768-d)

- Descriminator
  - input : h* = hCLS
  - MLP
  - output : k+1 vector of logit(trough softmax layer)

- BERT is updated when D is updated



GAN-BERT

- model specification

  noise vector : from N(0, 1)

  h : 768-d

  activation function G, D : leaky relu

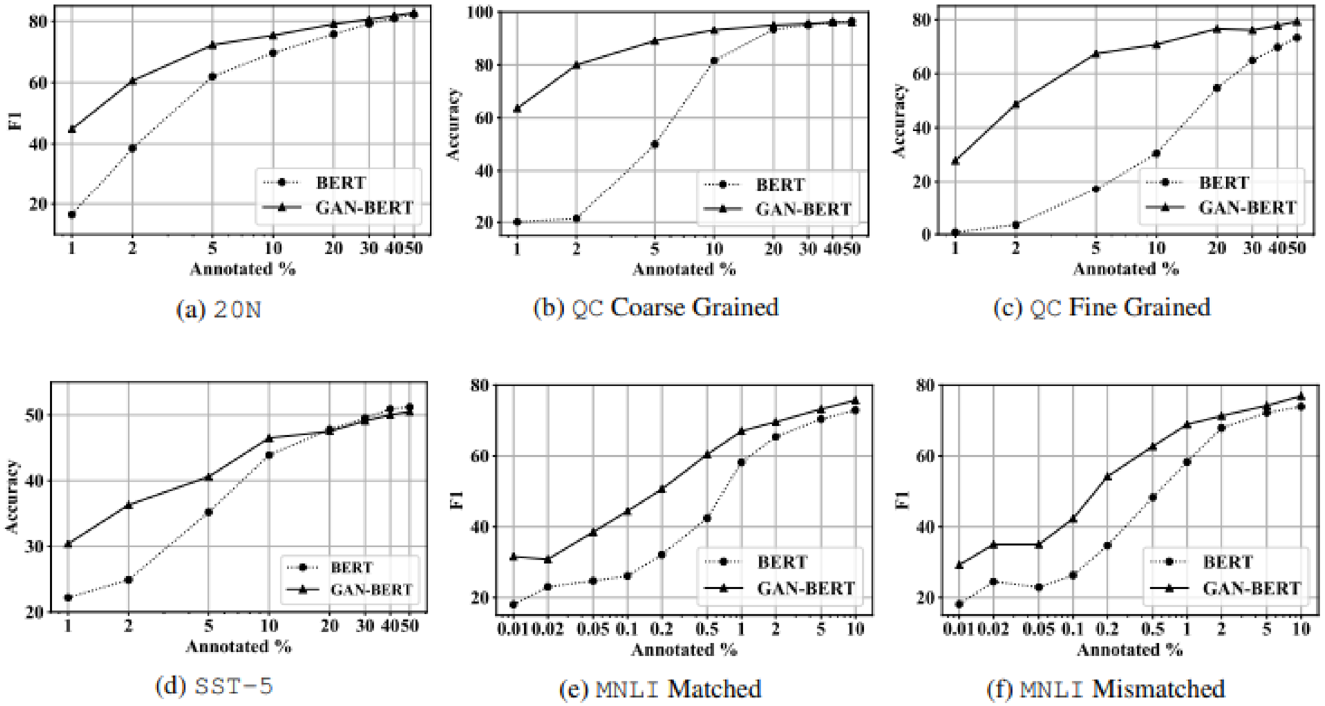
  dropout : 0.1

  annotated data 를 0.1% 또는 1%부터 사용하기 시작하고, 점점 늘려가면서 BERT와 차이를비교

  가능하면 사용한 annotated data의 100배에 해당하는 unlabeled data를 제공

Results

- Dataset
  - 20 NEWS Group(20N)
  - Question Classification(QC) – [fine grained / coarse grained]
  - Sentiment Analysis(SST-5)
  - MNLI



(a) 20N (b) QC Coarse Grained (c) QC Fine Grained

(d) SST-5 (e) MNLI Matched (f) MNLI Mismatched

Conclusion

- 안좋은 훈련환경에서 BERT를 능가하였다!
- labeled data개수가 적을 때에 GAN을 이용하여 성능을 비약적으로 향상시켰다.
- GPT-2, DistilBERT같은 다른 모델에도 적용가능하고, 다양한 task에도 적용할 수 있는 가능성을 열어주었다.
- BERT의 pre-training단계에도 직접적으로 적용해보고 싶다.