SentiGAN: Generating Sentimental Texts via Mixture Adversarial Networks

IJCAI2018 Distinguished Paper

Ke Wang, Xiaojun Wan

Institute of Computer Science and Technology, Peking University
The MOE Key Laboratory of Computational Linguistics, Peking University

Reporter:ziyang

Motivation

- the texts generated by GAN usually suffer from the problems of poor quality, lack of diversity and mode collapse.

Contribution / Bright spot

- propose a novel framework SentiGAN：multiple generators and one multi-class discriminator.
- propose a new penalty based objective to make each generator produce diversified texts of a specific sentiment label.
- outperforms the existing models in both the sentiment accuracy and quality of generated texts. (Use several metrics i.e. fluency, novelty, diversity, intelligibility to measure the quality of generated texts)
- The main intuition is that since text sentiment classification is very strong, we can use the classifier to guide the generation of sentimental texts.

Model

the objective of the i-th generator:

$$L(X) = G_i(X_{t+1}|S_t;\theta_g^i) \cdot V_{D_i}^{G_i}(S_t, X_{t+1}) \qquad (1)$$

$$\begin{aligned} J_{G_i}(\theta_g^i) &= \mathbb{E}_{X\sim P_{g_i}}[L(X)] \\ &= \sum_{t=0}^{t=|X|-1} G_i(X_{t+1}|S_t;\theta_g^i) \cdot V_{D_i}^{G_i}(S_t, X_{t+1}) \end{aligned} \qquad (2)$$

penalty function for the i-th generator :

$$V_{D_i}^{G_i}(S_{t-1}, X_t) = \begin{cases} \frac{1}{N}\sum_{n=1}^{N}(1 - D_i(X_{1:t}^n;\theta_d)) & t < |X| \\ 1 - D_i(X_{1:t};\theta_d) & t = |X| \end{cases} \qquad (3)$$

objective function of the discriminator:

$$\begin{aligned} J_D(\theta_d) = &- \mathbb{E}_{X\sim P_g} log D_{k+1}(X;\theta_d) \\ &- \sum_{i=1}^{k} \mathbb{E}_{X\sim P_{r_i}} log D_i(X;\theta_d) \end{aligned} \qquad (5)$$

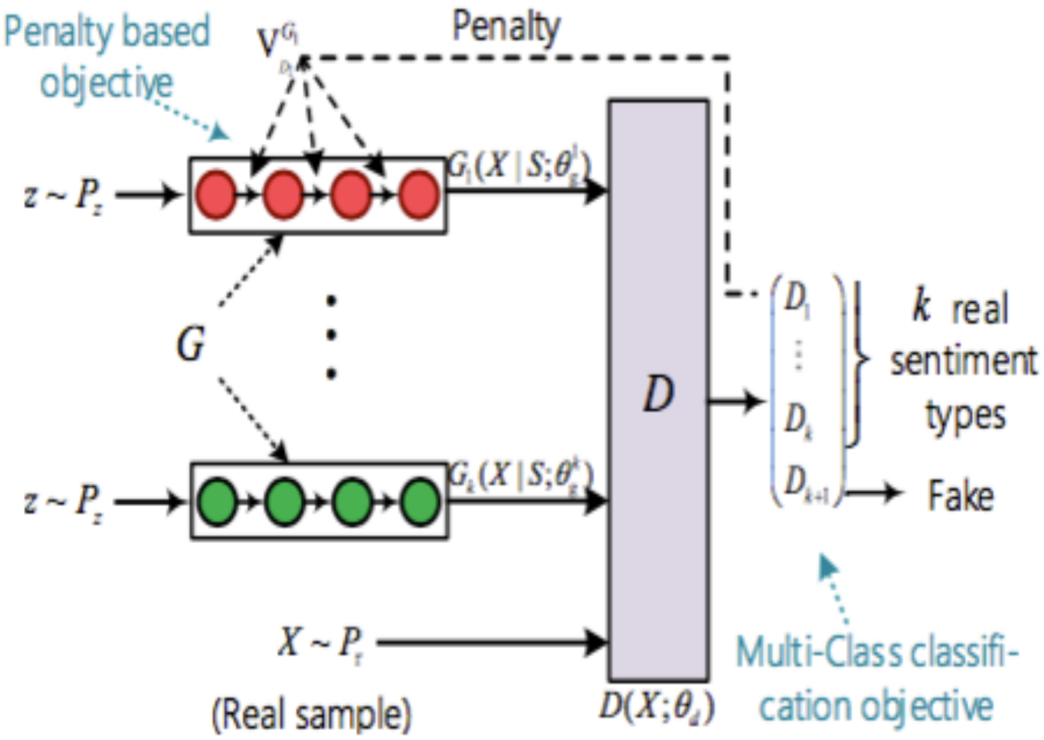

Figure 1: The framework of SentiGAN with k generators and one multi-class discriminator.

Training

Algorithm 1 The adversarial training process in SentiGAN

Input: Input noise, z; Generators, $\{G_i(X|S;\theta_g^i)\}_{i=1}^{i=k}$; Discriminator, $D(X;\theta_d)$; Real text dataset with k types of sentiment, $T = \{T_1, ..., T_k\}$;

Output: Well trained generators, $\{G_i(X|S;\theta_g^i)\}_{i=1}^{i=k}$;

Initialize $\{G_i\}_{i=1}^{i=k}$, D with random weights;
Pre-train $\{G_i\}_{i=1}^{i=k}$ using MLE on T;
Generate fake texts $F = \{F_i\}_{i=1}^{i=k}$ using $\{G_i\}_{i=1}^{i=k}$;
Pre-train $D(X;\theta_d)$ using $\{T_1, ..., T_k, F\}$;
repeat
    for g-steps **do**
        for i in $1 \sim k$ **do**
            Generate fake texts using $G_i(z;\theta_g^i)$;
            Calculate penalty $V_{D_i}^{G_i}$ by Eq (3) ;
            Update $G_i(z;\theta_g^i)$ by minimizing Eq (2);
        end for
    end for
    for d-steps **do**
        Generate fake texts $F = \{F_i\}_{i=1}^{i=k}$ using $\{G_i(X|S;\theta_g^i)\}_{i=1}^{i=k}$;
        Update $D(X;\theta_d)$ using $\{T_1, ..., T_k, F\}$ by minimizing Eq (5);
    end for
until SentiGAN converges
return ;

theoretical analysis of Penalty-Based Objective

$$J_G(X) = \begin{cases} \mathbb{E}_{X\sim P_g}[-log(D(X;\theta_d))] & \textbf{GAN} \\ \mathbb{E}_{X\sim P_g}[-log(G(X|S;\theta g)D(X;\theta_d))] & \textbf{SeqGAN} \\ \mathbb{E}_{X\sim P_g}[G(X|S;\theta_g)V(X)] & \textbf{SentiGAN} \end{cases} \quad (7)$$

1.can be considered as a measure of wasserstein distance： provides meaningful gradients, even when the distributions of P and P do not overlap.

wasserstein distance : $$W(P_r, P_g) = \frac{1}{K} sup_{||L||_L \le K} \mathbb{E}_{X\sim P_r}[L(X)] - \mathbb{E}_{X\sim P_g}[L(X)]. \quad (8)$$

2.forces the generator to prefer a smaller G(X |S ; θg). Thus it results in the generation of diversified samples, rather than repetitive but “good” samples.

$$\begin{aligned} G(X|S;\theta_g)V(X) &= G(X|S;\theta_g)(1 - D(X;\theta_d) \\ &= G(X|S;\theta_g) - G(X|S;\theta_g)D(X;\theta_d) \end{aligned} \quad (9)$$

Experiments

- Evaluate:
- 1.sentiment accuracy of the generated texts
- 2. the quality of generated texts (i.e., fluency, novelty, diversity, intelligibility)

sentiment accuracy of the generated texts

Evaluator: state-of-the-art sentiment classifier [Hu *et al.*, 2016] achieves an accuracy of 90% on the SST.

Accuracy	MR	BR	CR
Real Data	0.892	0.874	0.846
RNNLM	0.622	0.595	0.552
SeqGAN	0.717	0.684	0.632
VAE	0.751	0.721	0.643
SentiGAN(k=1)	0.803	0.750	0.731
C-GAN	0.822	0.773	0.762
S-VAE	0.831	0.793	0.727
SentiGAN(k=2)	**0.885**	**0.841**	**0.803**

Table 1: Comparison of sentiment accuracy of generated sentences. The real data is the training corpus.

Fluency

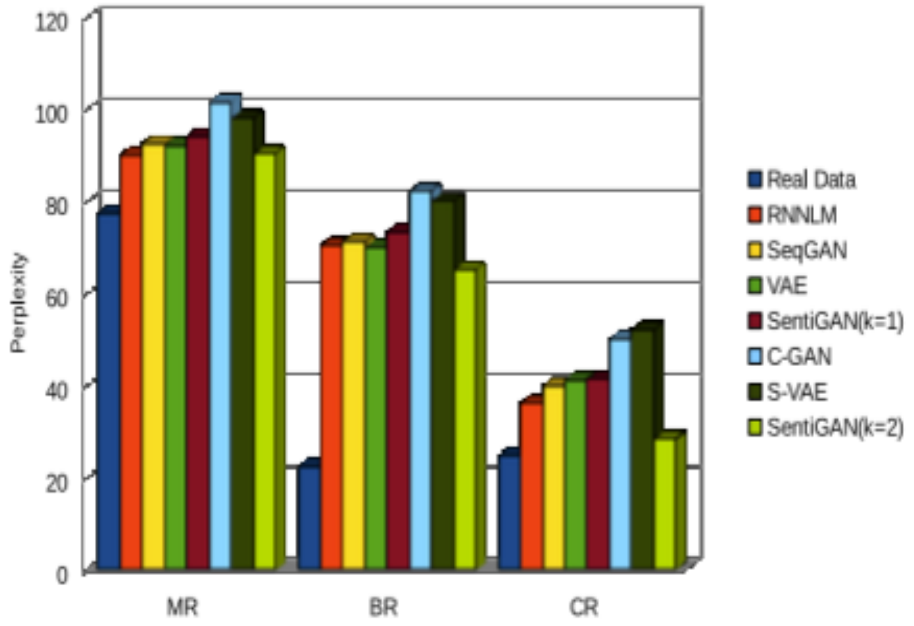

Figure 2: Comparison of fluency (Perplexity) of generated sentences (Lower perplexity means better fluency).

Intelligibility

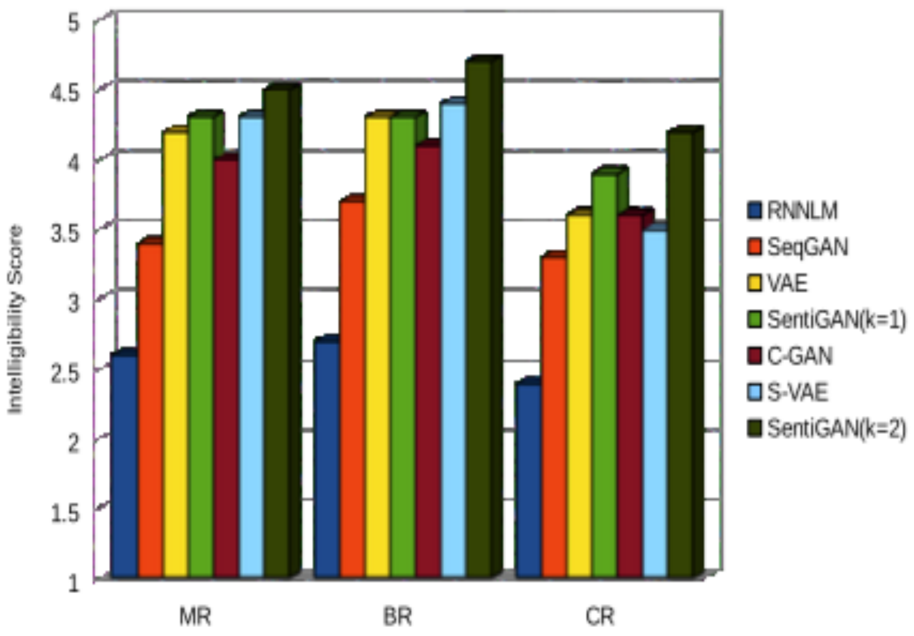

Figure 3: Comparison of intelligibility of generated sentences by human evaluation.

Novelty

$$Novelty(S_i) = 1 - \max\{\varphi(S_i, C_j)\}_{j=1}^{j=|C|}$$

Methods	MR	BR	CR
RNNLM	0.267	0.283	0.399
SeqGAN	0.298	0.328	0.437
VAE	0.287	0.347	0.417
SentiGAN(k=1)	0.344	0.409	0.479
C-GAN	0.368	0.398	0.482
S-VAE	0.328	0.369	0.437
SentiGAN(k=2)	**0.395**	**0.427**	**0.549**

Table 2: Comparison of the novelty of generated sentences.

Diversity

$$Diversity(S_i) = 1 - \max\{\varphi(S_i, S_j)\}_{j=1}^{j=|S|, j\neq i}$$

Methods	MR	BR	CR
Real Data	0.753	0.705	0.741
RNNLM	0.691	0.677	0.663
SeqGAN	0.641	0.636	0.619
VAE	0.661	0.658	0.620
SentiGAN(k=1)	0.711	0.687	0.668
C-GAN	0.726	0.688	0.680
S-VAE	0.692	0.687	0.649
SentiGAN(k=2)	**0.741**	**0.713**	**0.708**

Table 3: Comparison of the diversity of generated sentences.

Validation of Penalty-Based Objective

Method	MLE	SeqGAN	RankGAN	SentiGAN(k=1)
NLL	9.038	8.736	8.247	**6.924**

Table 5: The performance comparison of different methods on the synthetic data in terms of the negative log-likelihood (NLL) scores.

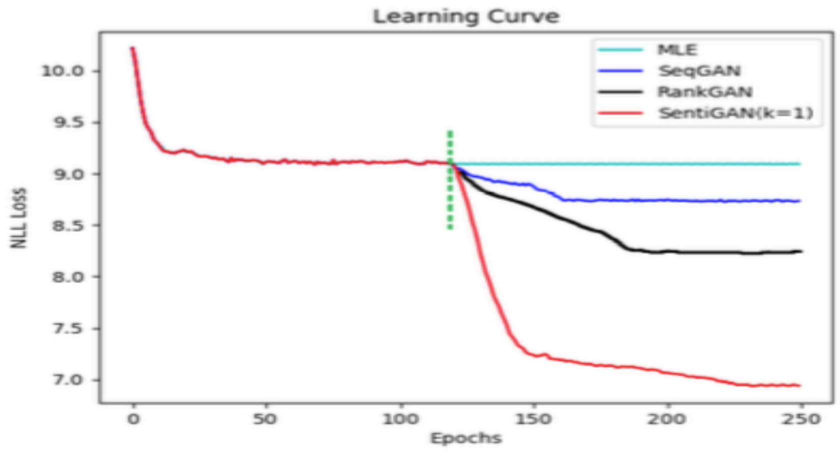

Figure 4: The illustration of learning curves. Dotted line is the end of pre-training.

	SentiGAN(k=2)	C-GAN
Positive	a fantastic finally , simply perfect masterpiece.	give it credit , this is our 's brilliant . ***(Unreadable)***
	one of the greatest movies i have ever seen.	good , bloody fun movie
	funny and entertaining , just an emotionally idea but it was pretty good.	makes me smile every time to get on alien . ***(Unreadable)***
	the best comedy is a science fiction , captain is like a comic legend.	powerfully moving ! ***(Very short)***
Negative	one of the most disturbing and sickening movies i have ever seen.	very bad comedy. ***(Very short)***
	a story which fails to rise above its disgusting source material .	a mere shadow of its predecessors
	the comedy is nonexistent .	a timeless classic western dog ... ***(Wrong sentiment)***
	this is a truly bad movie .	one of those history movie traps

Table 4: Examples sentences generated by SentiGAN and Conditional GAN trained on MR.

some thoughts

- Diversity
- Only focus on generating short sentences (length $\leqslant$ 15 words)
- Classifier: Benefit? Limit?