A Generative Model For Category Text Generation

Yang Li , Quan Pan , Suhang Wang , Tao Yang , Erik Cambria

Information Sciences
March, 2018

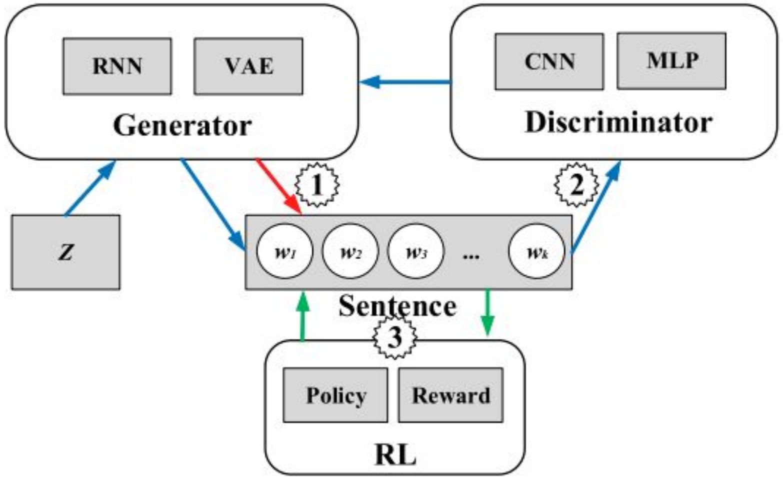
Presenter: Anchit Bhattacharya
Arizona State University

Why?

- Deep learning architecture works better with more labelled data
  - Tends to overfits when training data is small
  - Unable to capture category information
- There is a gap between importance of large data sets and difficulty in obtaining such data

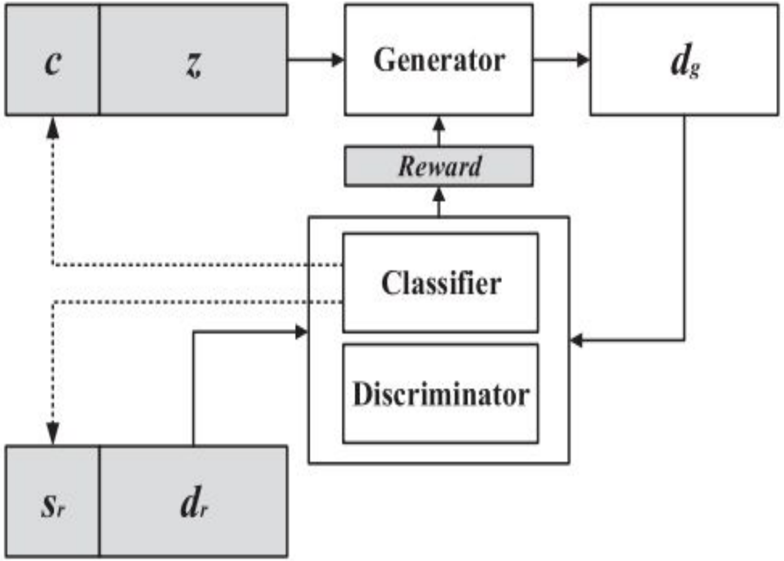
- GAN is an exciting generative architecture to produce new samples, but mostly explored in Image generation
- Most text generation approaches using GAN’s produces unlabelled data, so not suitable for data augmentation for classifiers.
- Novel problem of generating labeled sentences with GANs for Data Augmentation

Sentence Generation Models



- Using an RNN/LSTM model to generate sentence.
- Add a discriminator to play min-max game to update generator
- Use sentence generation as a RL problem, where next token is generated to maximise the long term reward based on a policy.

CS-GAN Model



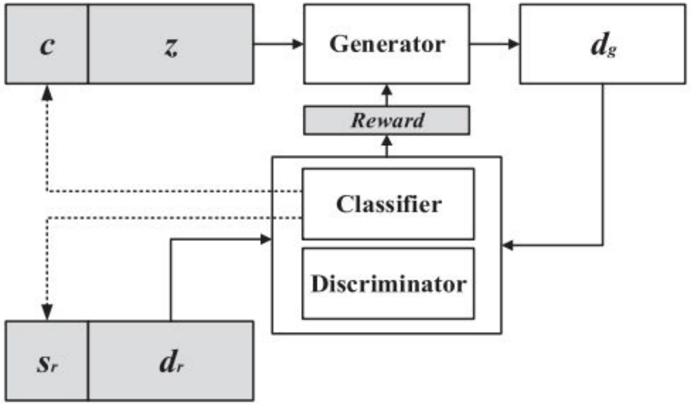
Generator - Generate Category Synthetic Sentence

Discriminator - To determine the sentence validity

Classifier - Determine Category Accuracy

Generator

- LSTM is used as a generator
- Category Information is added at each generating step



Modified LSTM:-

$$f_t = \sigma\,(W_f[x_t;\, z;\, c] + U_f\, h_{t\text{-}1} + b_f)$$

$$i_t = \sigma\,(W_i[x_t;\, z;\, c] + U_i h_{t-1} + b_i)$$

$$o_t = \sigma\,(W_o[x_t;\, z;\, c] + U_o h_{t-1} + b_o)$$

$$c_t = f_t\, c_{t-1} + it\, \sigma\,(Wc[x_t;\, z;\, c] + U_c h_{t-1} + b_c\,)$$

$$h_t = o_t\, \mathrm{relu}(c_t)$$

Generator(Contd)

- Reinforcement Learning for choosing next token based on long term Reward

$D_{\theta d}$ - probability that the sentence is real

$D_{\theta c}$ - probability that the sentence is in right category

$J(\theta g) = E[R_T\, | w0, c, \theta g] = \sum_{dg \in W} G_{\theta g}\, (d_g | w_0, c) Q^{G\theta g}\, (w_{0,}\, d_g),$

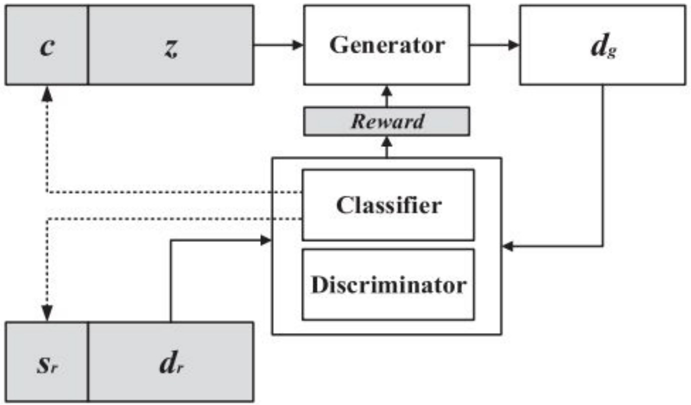
$G_{\theta g}$ - Generated results based on LSTM - $p_G(d_g | w_0, z, c)$

$Q^{G\theta g}\, (w_{0,}\, d_g) = (2D_{\theta d} D_{\theta c}\,)/\, D_{\theta d} + D_{\theta c}$

$Q^{G\theta g}\, (w_{0,}\, d_g)$ - Action Value function during the selection process

Descriptor

- Discriminator and Classifier are CNNs used for Sentence Classification



Classifier Loss:-

$LC = Loss(< C(dr; θc), sr >) + Loss(< C(G(z; θg); θc), ϕ(c) >)$

Discriminator:-

$LD = HEdg\sim(c,z) + HEdr\sim data$

$HEdg\sim(c,z) = Edg\sim(c,z)[-\log(1 - D(G(ϕ(c), z, θg)), θd)]$

$HEdr\sim data = Edr\sim data[-\log(D(sr, dr, θd))]$

SST	POSITIVE	NEGATIVE
	This is really lead movie.	As film and day waste.
	You take the one better.	This movie is mess.
Emotion	LOVE	EMPTY
	Love the warm	Is telling a story?
	Oh thank you my dear	Read the book boring
	RELIEF	ANGER
	Read the book	I hate the hat
	Thanks	Jest the boy
	SURPRISE	NEUTRAL
	Weather today is a gift to us	Have a tea please.
	I hate that happend	So the heat is here
	HAPPINESS	SADNESS
	haha	I hate it
	Enjoy the game	No body is at eating
	FUN	ENTHUSIAMSM
	Yeah	Love the health
	Have the fun	Love the eat
NEWS	SPORT	BUSINESS
	Ball center with two teams	There is a flat race on euro
	In a race, the man wade in river	It is easy a pricing ran in tea
	ENTERTAINMENT	US
	We see it ease a strong man	U.S. tears a ratar in the u.s.
	Not a keen actor	Japan for a real anneal stress
	WORLD	HEALTH
	Hear the trade war end	Done in a dead spar
	Europe end this under the stress	Argue a 1 star new rush
	SCI-TEC	
	Tale in fick nart ear at bet	
	Nets play and in Brate	

Results on Classification

Table 3
The classification results on the different size dataset.

Model name	Amazon-5000	Amazon-30000
CNN	84.83%	89.55%
CS-GAN w/o RL&GAN	85.60%	**89.67%**
CS-GAN w/o RL	86.18%	89.54%
CS-GAN	**86.43%**	89.34%
Model name	Emotion-15000	NEWS-15000
CNN	40.75%	72.08%
CS-GAN w/o RL&GAN	39.32%	72.31%
CS-GAN w/o RL	40.14%	72.09%
CS-GAN	**41.52%**	**74.33%**

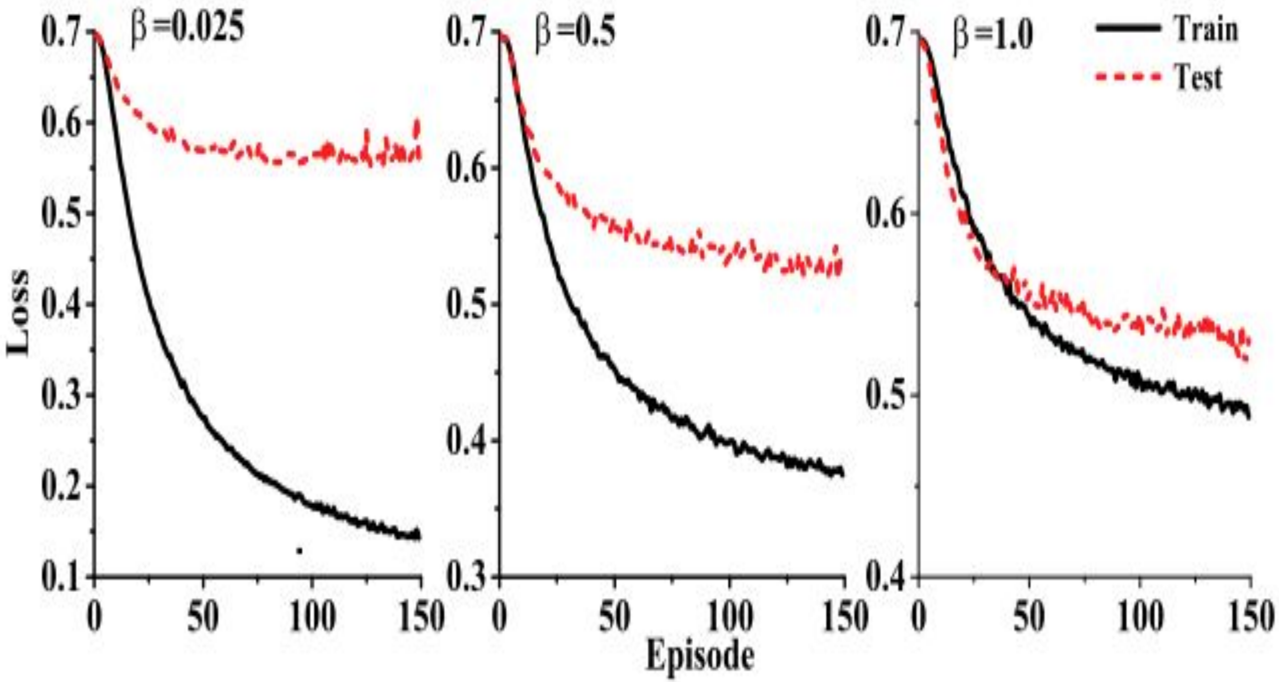
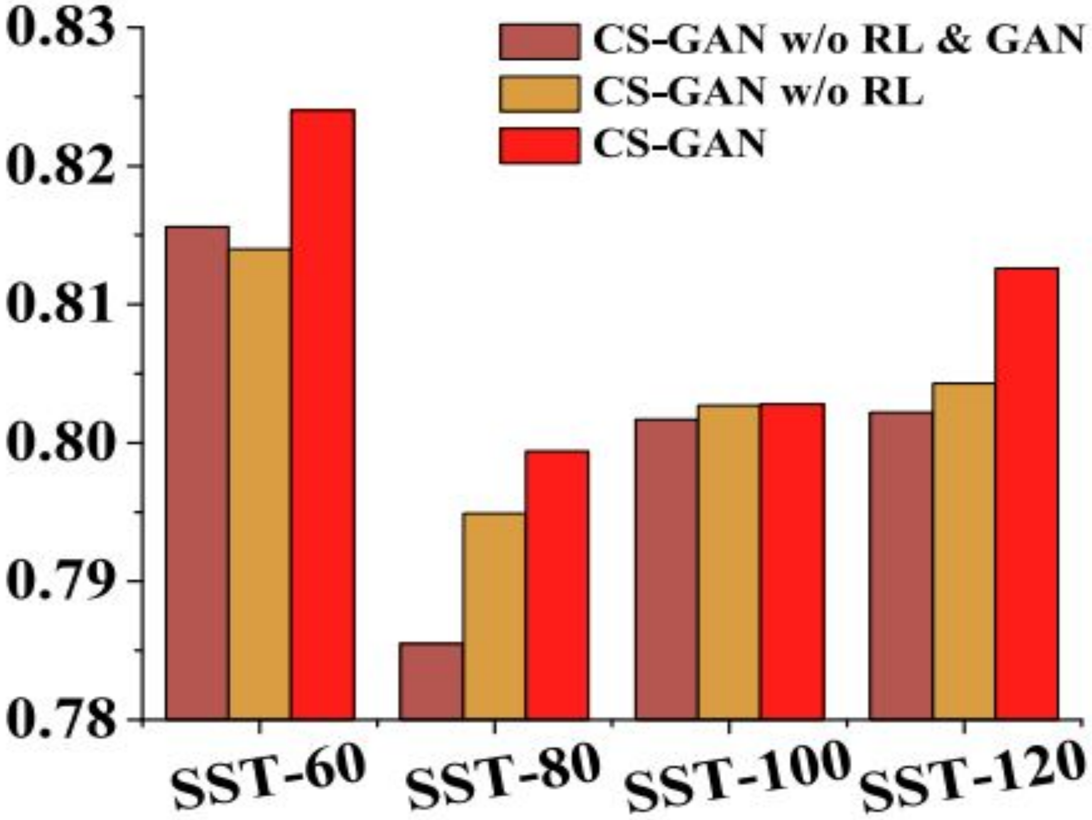
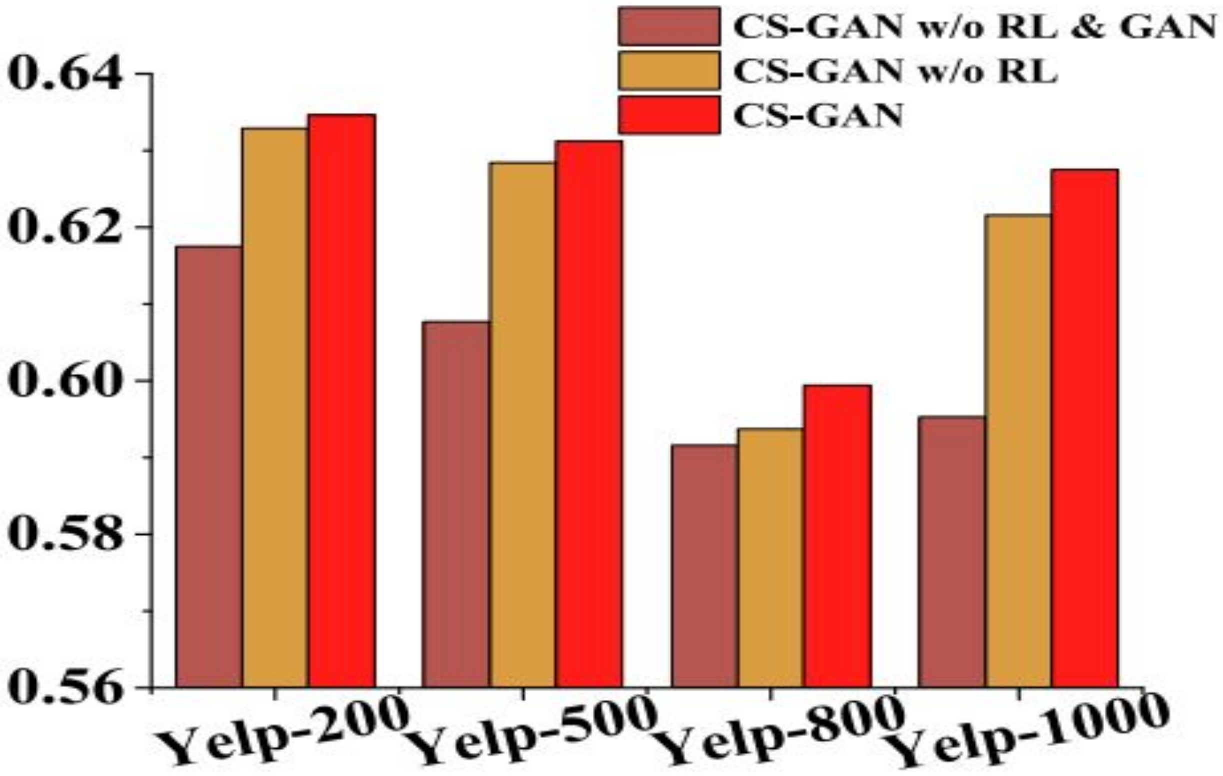


Fig. 6. The results of the losses from CS-GAN using different numbers of generated data as training data, and β is the generation ratio.


0.83
0.82
0.81
0.80
0.79
0.78
CS-GAN w/o RL & GAN
CS-GAN w/o RL
CS-GAN
SST-60
SST-80
SST-100
SST-120
(a) SST


CS-GAN w/o RL & GAN
CS-GAN w/o RL
CS-GAN
0.64
0.62
0.60
0.58
0.56
Yelp-200
Yelp-500
Yelp-800
Yelp-1000

(b) Yelp

Text Generation Results

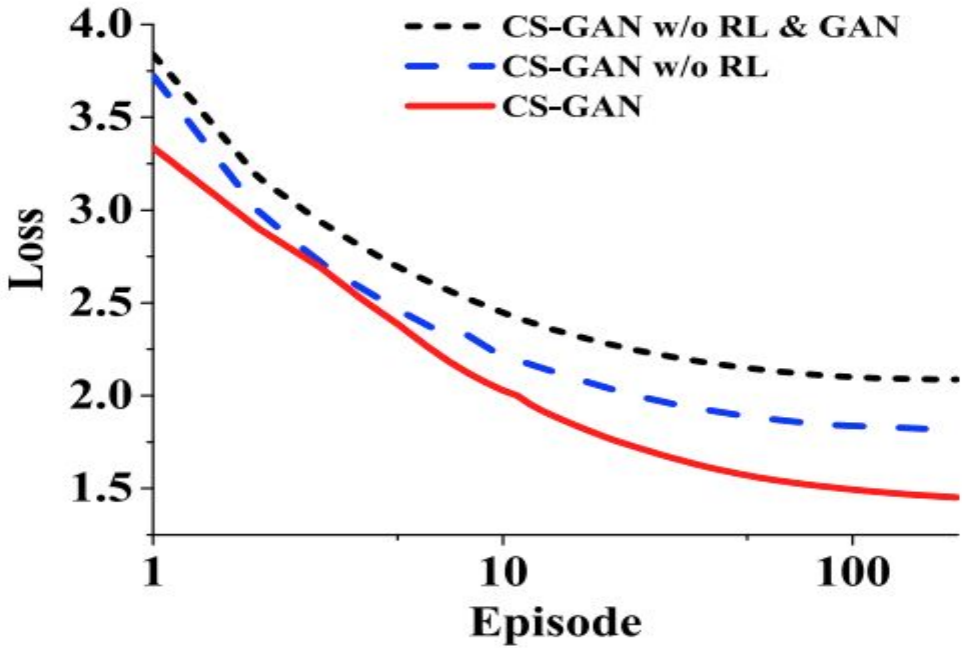


Fig. 4. The results of NLL from model CS-GAN, CS-GAN without RL and CS-GAN without RL & GAN in text data generation.

Conclusion

- The proposed model performs well in supervised learning tasks, especially in multicategory datasets.
- However, the advantage can be weakened when there is large amount of data with little category information.
- Shows, better performance for small labeled dataset with small sentence length.