Towards Controlled Generation of Text

Zhiting Hu Zichao Yang Xiaodan Lang **Ruslan Salakhutdinov** Eric P.Xing

Carnegie Mellon University

ICML 2017

Presenter: Anchit Bhattacharya

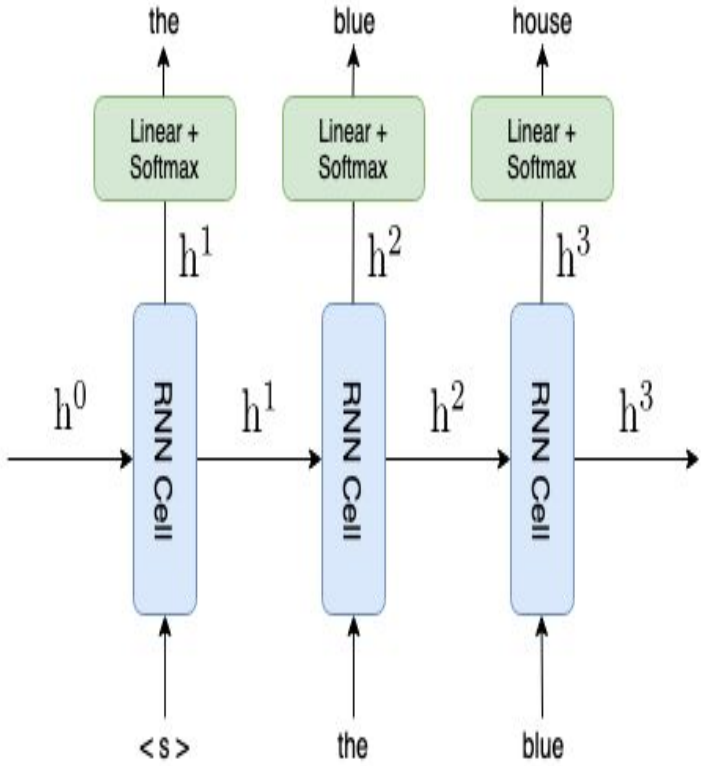
Arizona State University

Text Generation Methods

RNN based text generation -

- Train a RNN based language model and use it to generate the next token based on all previous tokens.

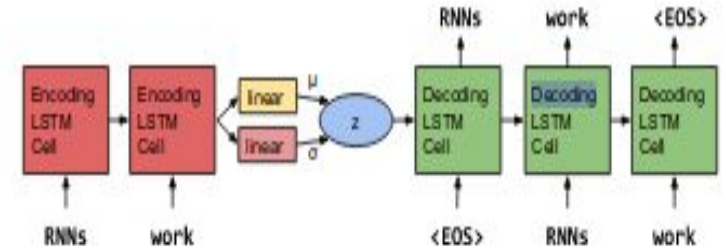
- Can be character level or word level language model



Text Generation Methods (Continued)

VAE based text generation -

- Incorporates distributed latent representations of entire sentences.



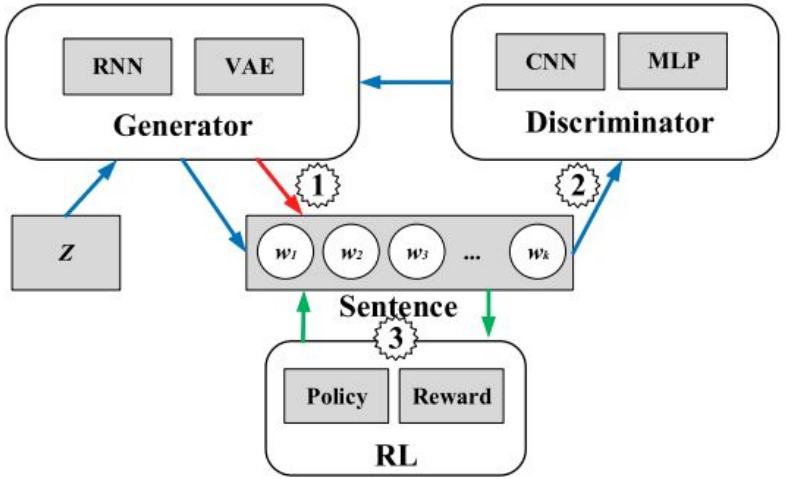
- Explicitly model holistic properties of sentences.

Text Generation Methods (Continued)

GAN based text generation -

- Adversarial training to improve generated text.

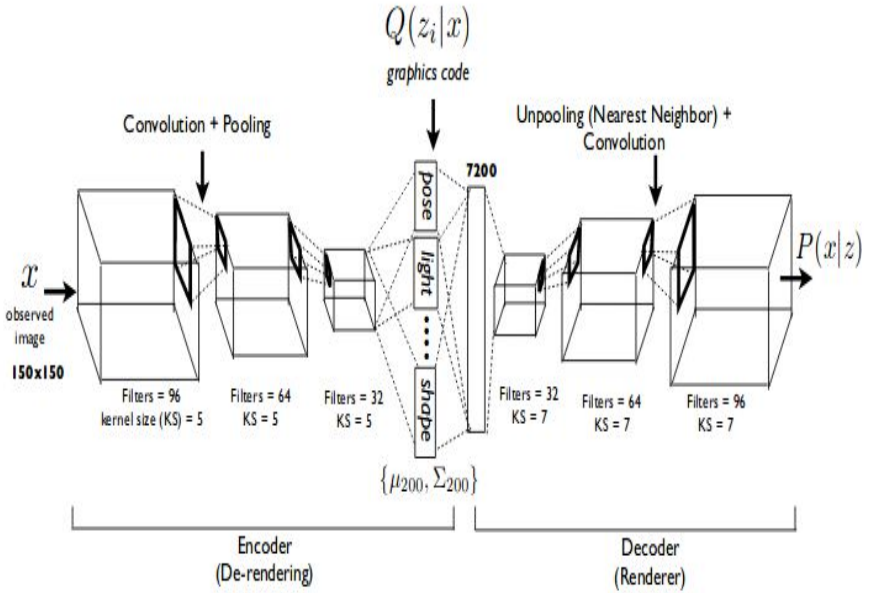
- Can be used as a training method on top of previous models.



Disentangled Representations

- Neurons in the neural network are somehow learning complete concepts alone.

- We want the latent representation of the texts to be disentangled, so that we can control the generation effectively.

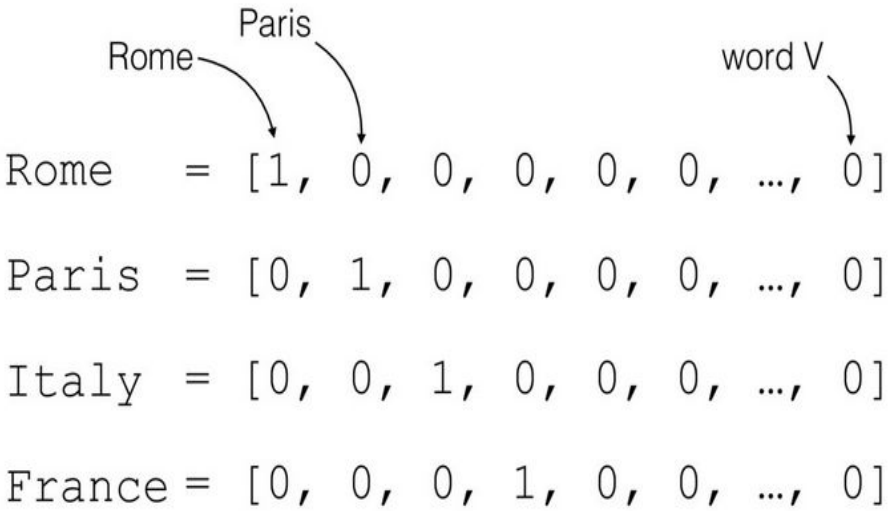


Challenges for Controllable Text Generation

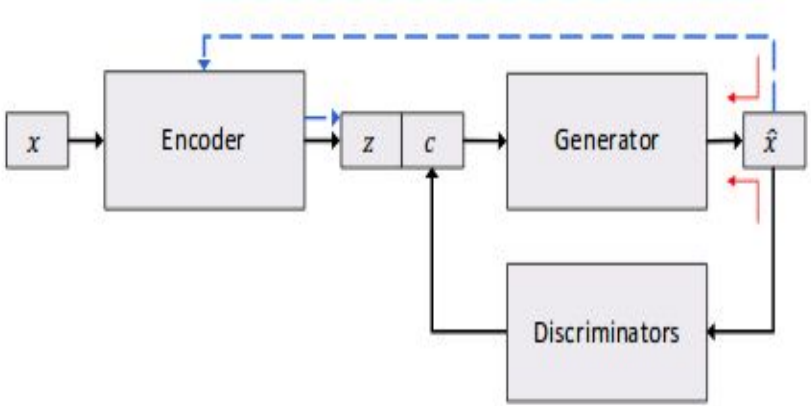
1> **Discrete Nature of Text Samples** - Difficult to do backpropagation and train the generator in adversarial based methods.

2> **Learning Disentangled representations** - Lacks explicit enforcement of independence property on full latent representation.

3> Controlled Generation of Text - Control the generation of text based on some properties generically. Byproduct of point 2.



Model



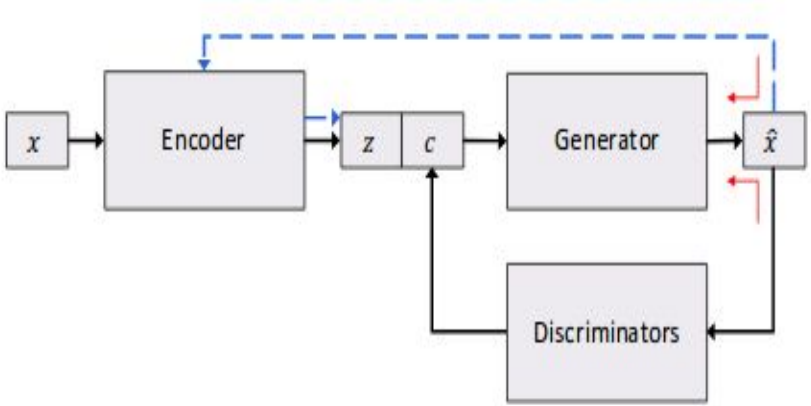
Discrete Nature(Solution)

- Continuous Approximation based on softmax with a decreasing temperature(T).

$$x = softmax(o \mid T)$$

- Low variance and fast convergence as opposed to Policy Gradient based method.

Model



Controlled Generation Solution

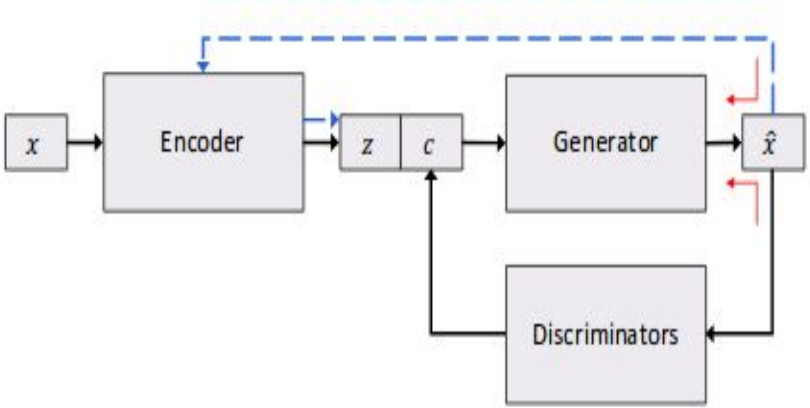
- Add a structured code to the latent representation for every attribute that needs to be controlled.

  z - Gaussian prior

  c - Continous/Discrete

- Different discriminator for each attribute

Model



Generator Training(Part1)

- Train parameters of Encoder($Theta_e$) and Generator($Theta_g$) based on the reconstruction Error of real sentences.

- Also, train the encoder so that the latent representation is close to a Gaussian prior.

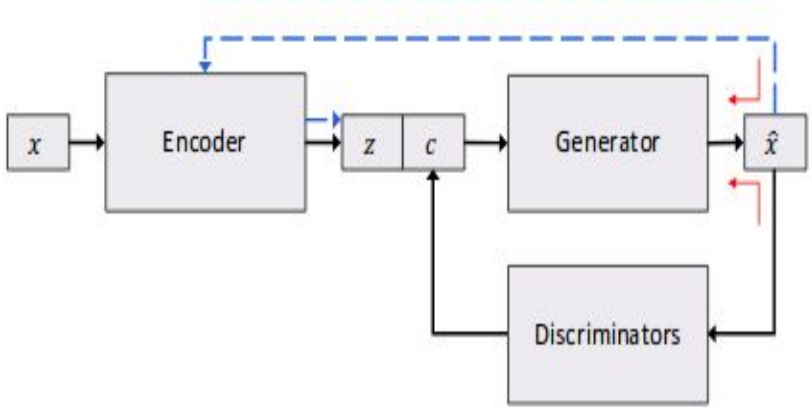
Mathematically:-

$$L_{VAE}(\theta_G,\theta_E;x) = KL(q_E(z|x)\|p(z)) - \mathbb{E}_{qE(z|x)\mathbf{qD(c|x)}}[\log p_G(x|z,c)],$$

Loss function

Forcing latent code towards Gaussian prior

Model



Generator Training(Part 2)

- Discriminator produces extra learning signals enforcing the generator to produce text with certain attributes conditioned on code.
- It is still possible that other attributes not explicitly modelled may also entangle with the code.

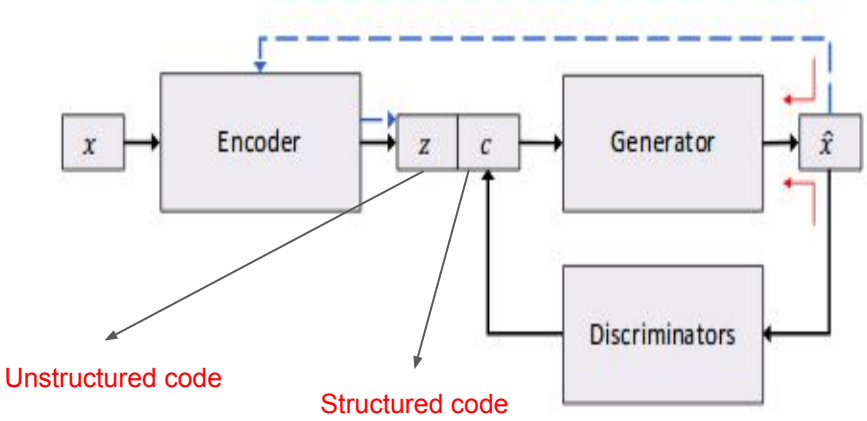
Mathematically:-

$$L_{Attr,c}(\theta_G) = -\mathbb{E}_{p(z)p(c)}[\log_{qD}(c|\, G_T(z,c))]$$

Loss function

Train the generator based on discriminator O/P of generated text

Model



Generator Training(Part 3)

- We reuse the Encoder as a discriminator, and match it with the unstructured code(z)

O/P is z instead of c

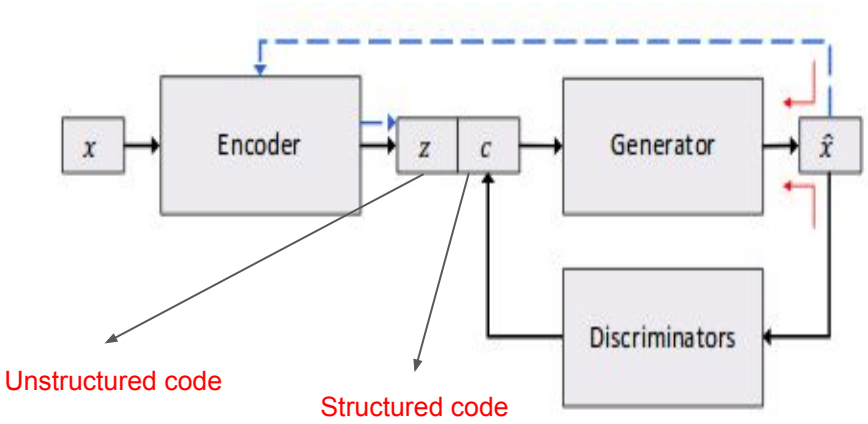
Mathematically:-

$$L_{Attr,z}(\theta_G) = -\mathbb{E}_{p(z)p(c)}[\log_{qE}(z|\, G_T(z,c))]$$

Loss function

Train the generator based on discriminator O/P of generated text

Model



Generator Training(Overall Loss)

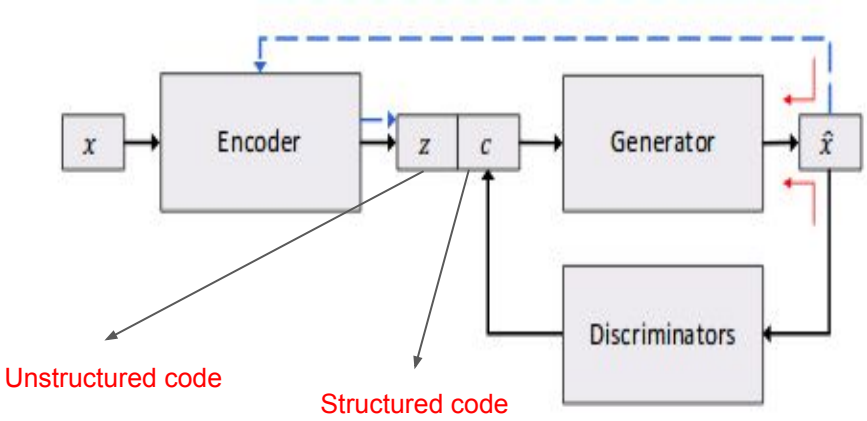
- Overall Loss is the sum of all the losses.
- λ_c and λ_z are balancing parameters for the loss

Overall Loss function:-

$$\min_{\theta G} L_G = L_{VAE} + \lambda_c L_{Attr,c} + \lambda_z L_{Attr,z'}$$

Loss function

Model



Discriminator Training

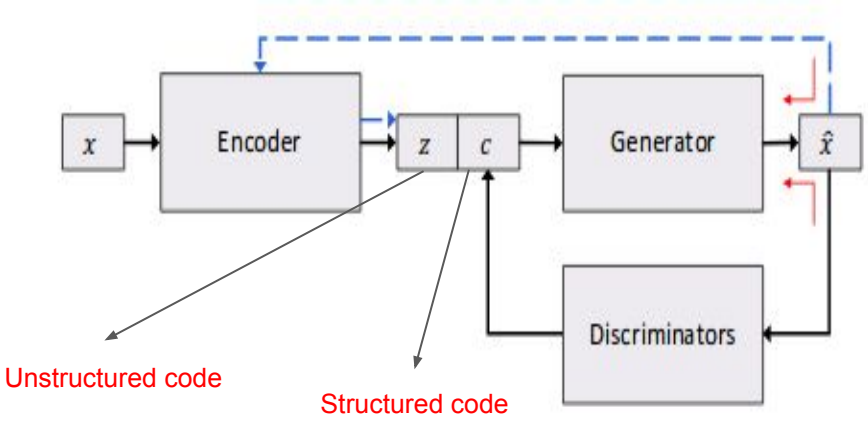
- Can be formulated as a sentence classifier for categorical code, or a probabilistic regressor for continuous code.
- Use labelled examples to train the discriminator, to embed the text characteristics into the code.

Mathematically:-

$$L_S(\theta_D) = -E_{X_L}[\log q_D(c_L|x_L)]$$

Training on labelled data

Model



Discriminator Training

- Use samples generated from generator as a augmented data for training the discriminator
- Use labelled examples to train the discriminator, to embed the text characteristics into the code.

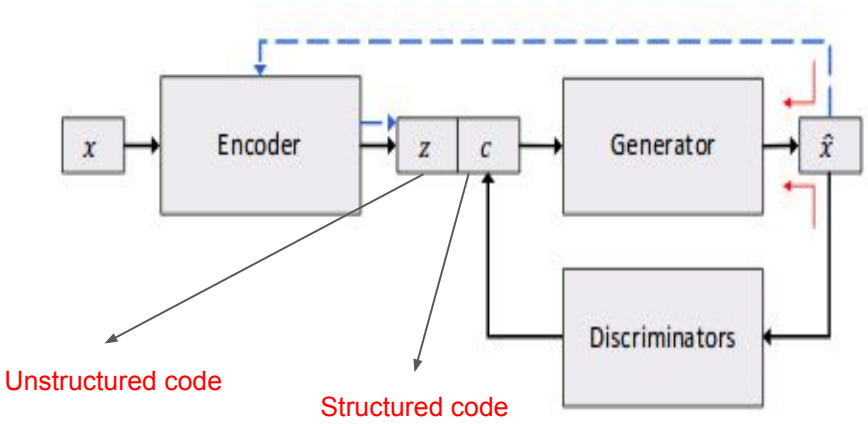
Mathematically:-

$$L_u(\theta_D) = - \mathbb{E}_{p_G(\hat{}x|z,c)p(z)p(c)} [\log q_D(c|x) + \beta H(q_D(c'|\hat{}x))],$$

Training discriminator on generator data

Shannon entropy on generated data to compensate for noisy data from generator

Model



Discriminator Training(Overall Loss)

- Use samples generated from generator as a augmented data for training the discriminator
- Use labelled examples to train the discriminator, to embed the text characteristics into the code.

Overall loss function :-

$$\min_{\theta_D} L_D = L_s + \lambda_u L_u$$

Training

Algorithm 1 Controlled Generation of Text

Input: A large corpus of unlabeled sentences $\mathcal{X} = \{\boldsymbol{x}\}$
A few sentence attribute labels $\mathcal{X}_L = \{(\boldsymbol{x}_L, \boldsymbol{c}_L)\}$
Parameters: $\lambda_c, \lambda_z, \lambda_u, \beta$ – balancing parameters

1: Initialize the base VAE by minimizing Eq.(4) on $\mathcal{X}$ with $\boldsymbol{c}$ sampled from prior $p(\boldsymbol{c})$
2: **repeat**
3: Train the discriminator D by Eq.(11)
4: Train the generator G and the encoder E by Eq.(8) and minimizing Eq.(4), respectively.
5: **until** convergence

Output: Sentence generator G conditioned on disentangled representation $(\boldsymbol{z}, \boldsymbol{c})$

Experiments

Generate short sentences(length <= 15)

Sentence corpus - IMDB text corpus(1.4m)

Sentiment -

- SST dataset - 6920/872/1821
- IMDB review datasets - 5K/1K/10K
- SST small - 250 labelled samples
- Lexicon - word level sentiment labels

Tense -

- Dataset of labelled words
- **5250 words** and phrases labeled with one of **{“past”, “present”, “future”}**

Results (Disentangled Representation)

w/ independency constraint	w/o independency constraint
the film is strictly routine !	the acting is bad .
the film is full of imagination .	the movie is so much fun .
after watching this movie , i felt that disappointed .	none of this is very original .
after seeing this film , i 'm a fan .	highly recommended viewing for its courage , and ideas .
the acting is uniformly bad either .	too bland
the performances are uniformly good .	highly watchable
this is just awful .	i can analyze this movie without more than three words .
this is pure genius .	i highly recommend this film to anyone who appreciates music .

Results (contd..)

Varying the code of tense	
i thought the movie was too bland and too much	this was one of the outstanding thrillers of the last decade
i guess the movie is too bland and too much	this is one of the outstanding thrillers of the all time
i guess the film will have been too bland	this will be one of the great thrillers of the all time

Varying the unstructured code z	
("negative", "past")	*("positive", "past")*
the acting was also kind of hit or miss .	his acting was impeccable
i wish i 'd never seen it	this was spectacular , i saw it in theaters twice
by the end i was so lost i just did n't care anymore	it was a lot of fun
("negative", "present")	*("positive", "present")*
the movie is very close to the show in plot and characters	this is one of the better dance films
the era seems impossibly distant	i 've always been a big fan of the smart dialogue .
i think by the end of the film , it has confused itself	i recommend you go see this, especially if you hurt
("negative", "future")	*("positive", "future")*
i wo n't watch the movie	i hope he 'll make more movies in the future
and that would be devastating !	i will definitely be buying this on dvd
i wo n't get into the story because there really is n't one	you will be thinking about it afterwards, i promise you

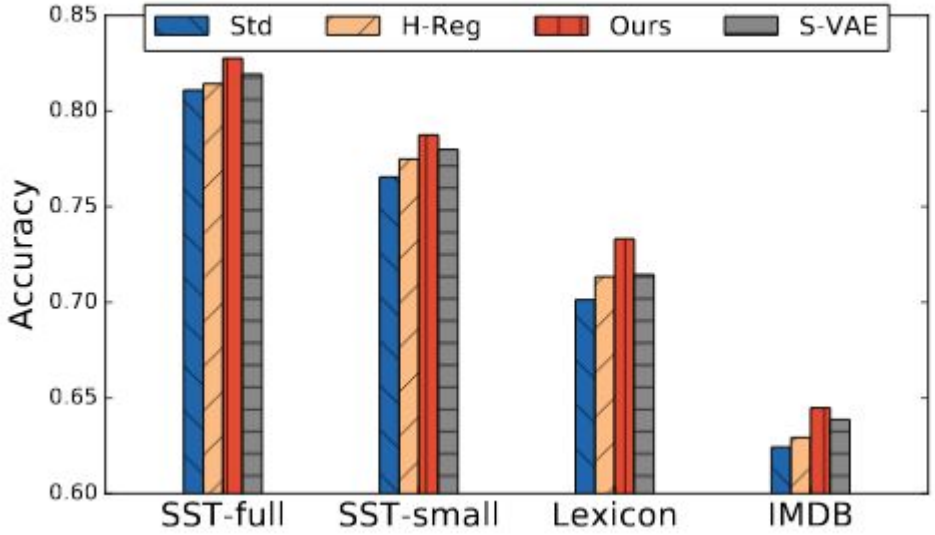
Results (contd...)

Failure cases	
the plot is not so original	it does n't get any better the other dance movies
the plot weaves us into <unk>	it does n't reach them , but the stories look
he is a horrible actor 's most part	i just think so
he 's a better actor than a standup	i just think !

Results(Sentiment Accuracy)

Model	Dataset		
	SST-full	SST-small	Lexicon
S-VAE	0.822	0.679	0.660
Ours	**0.851**	**0.707**	**0.701**

Results(Data Augmentation)



Questions