Text Generation Survey 2017

Ni Lao

2017

Text generation

- Non-linguistic input (logical forms, database entries, etc.) or text ⇒ Text
  - Text to text: translation, summarization
  - Structure to text: caption, question answering
  - Image to text: caption
  - A vector z to text: unconstrained generation

Blog [Karpathy+ 2016]
NIPS [Goodfellow 2016]

Why study generative models?

- Realistic generation tasks
- Semi-supervised learning
- Science

"What I cannot create,
I do not understand."

—Richard Feynman

Human: *what is moral ?*
Machine: *what empowered humanity , what intellectual the essence is .*
Human: *what is immoral ?*
Machine: *the fact that you have a child .*
Human: *what is morality ?*
Machine: *what is altruism ?*
Human: *what is the definition of altruism ?*
Machine: *if you don 't believe in god , then you don 't know .*
Human: *ok ... so what is the definition of morality ?*
Machine: *well , the truth is , you 're not a believer in god almighty .*
Human: *tell me the definition of morality , i am quite upset now !*
Machine: *i 'm not ashamed of being a philosopher !*

[Vinyals, Le, 2015]

Plan

- Unconditional image/text generation models
  - Seq2Seq
  - Variational Auto-Encoder (VAE)
  - Generative Adversarial Net (GAN)

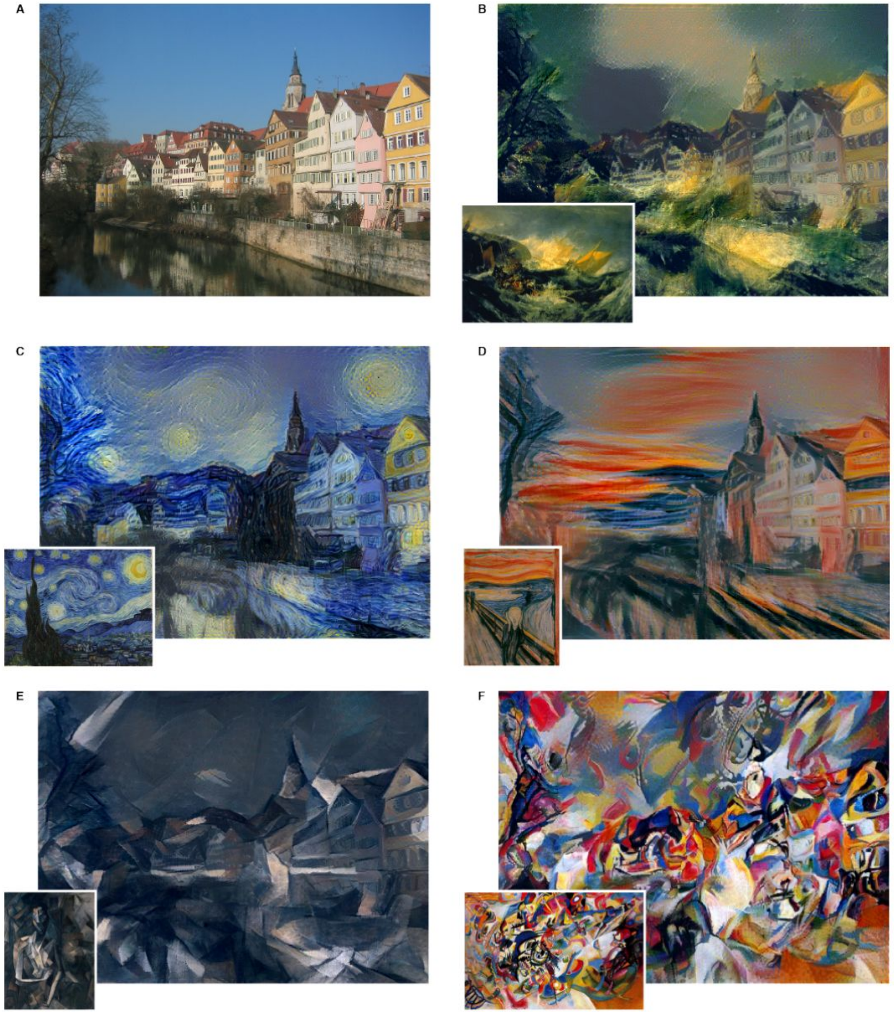
- Improved text generation models
  - Conditioned generation
  - Reinforcement Learning

Impressive results for image

Figure 2: Images that combine the content of a photograph with the style of several well-known artworks.

The images were created by finding an image that simultaneously matches the content representation of the photograph and the style representation of the artwork (see Methods).

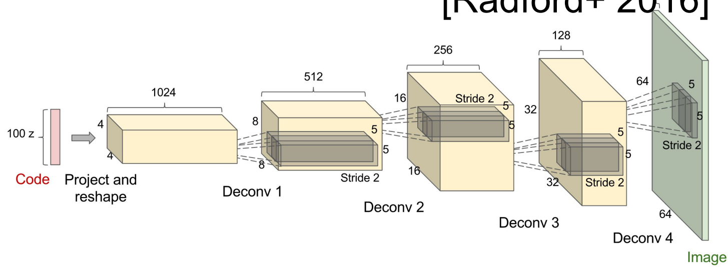
[Gatys, Ecker, Bethge, 2015]



[Radford+ 2016]

Impressive results for image

- Sampled from Deep Convolutional (DC) GAN

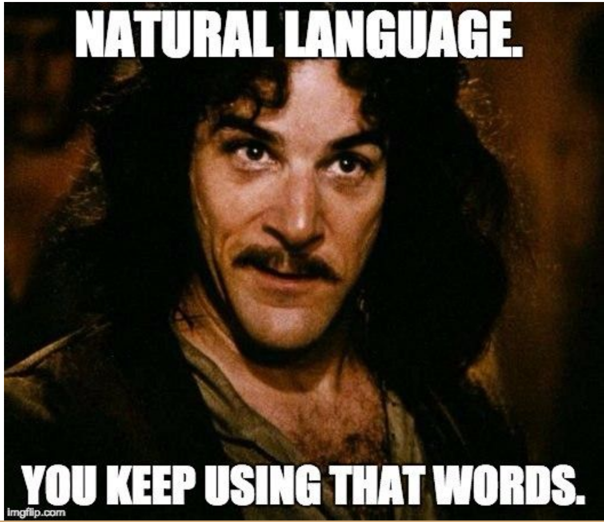


[Rajeswar+ 2017]

Immediately criticised when applied to text

- "I have a lot of respect for language. Deep-learning people seem not to"
- "They include such impressive natural language sentences as:"
  - * what everything they take everything away from
  - * how is the antoher headache
  - * will you have two moment ?
  - * This is undergoing operation a year .
- "These are not even grammatical!"
- The DNN bubble consists of models, which show great promises but not yet practical at this point

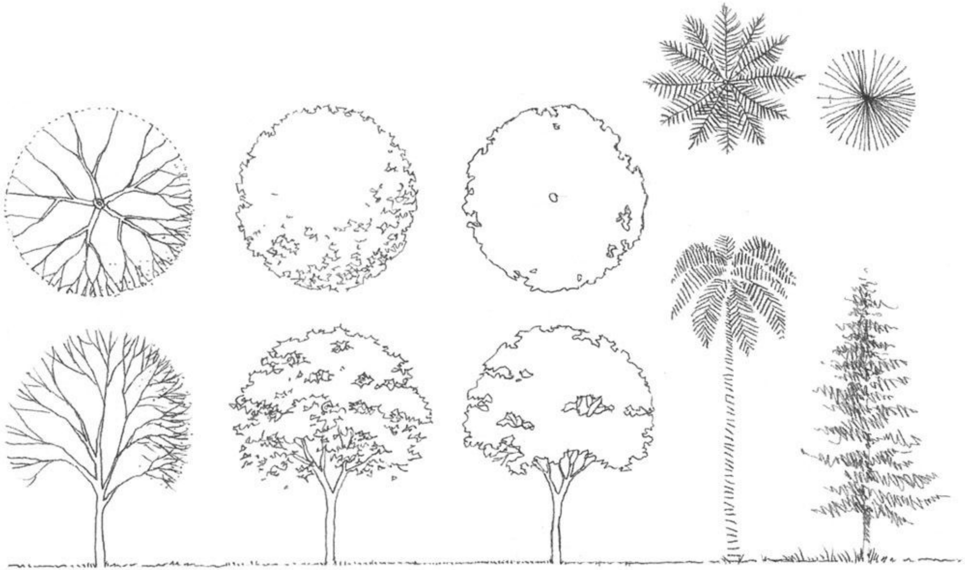
Blog [Goldberg 2017]



Seq2Seq [Sutskever, Vinyals, Le 2014]
VAE [Kingma & Welling 2014]
GAN [Goodfellow+ 2014]
ACL [Goldberg 2015]
ACL [Mooney 2015]

furious

statistician's view v.s. linguist's view

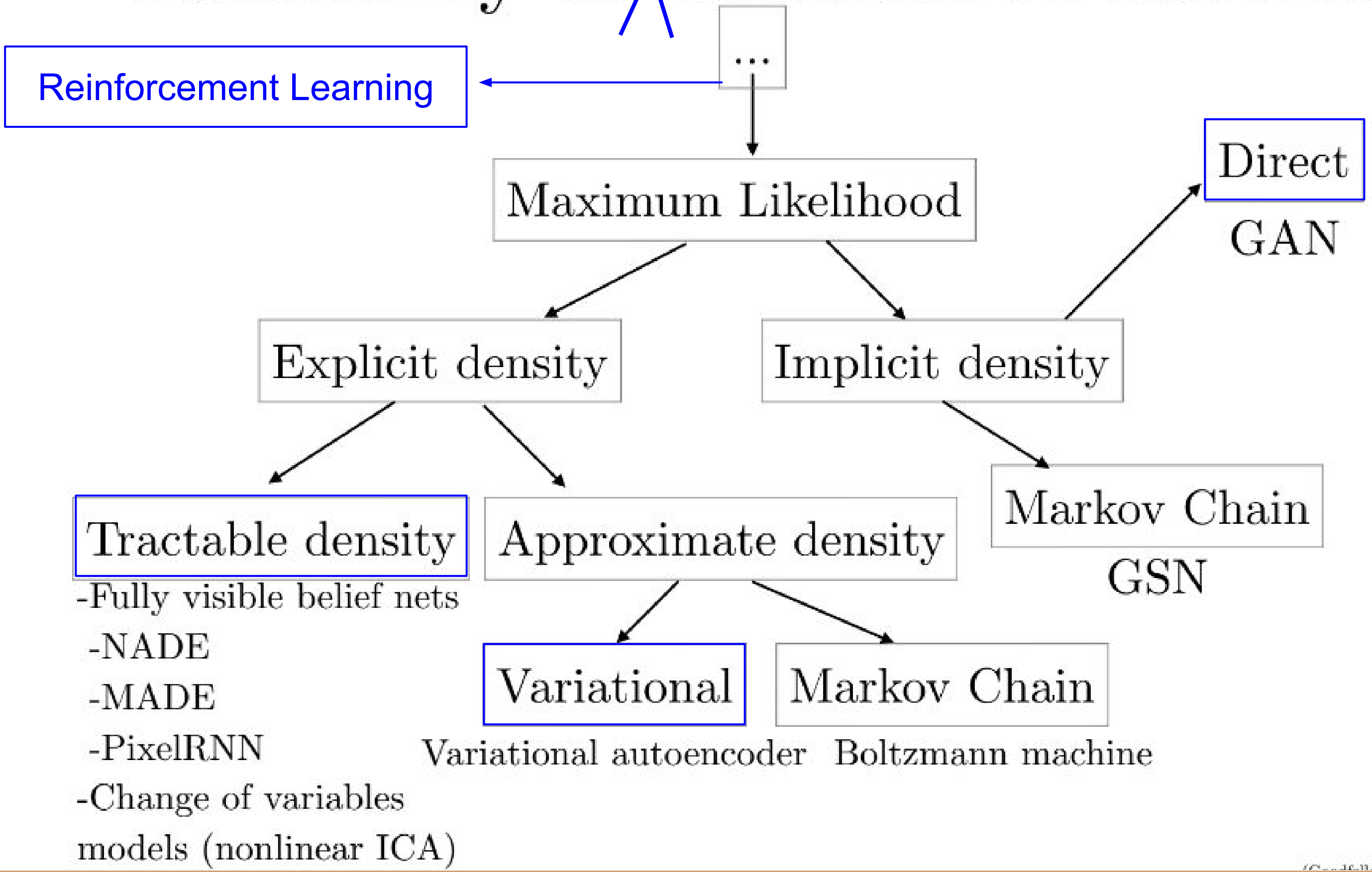


Everything can be mapped to a unit Gaussian ball (given the power of DNNs)

The world has real structures, which need to be represented by real structures

[Goodfellow 2016]

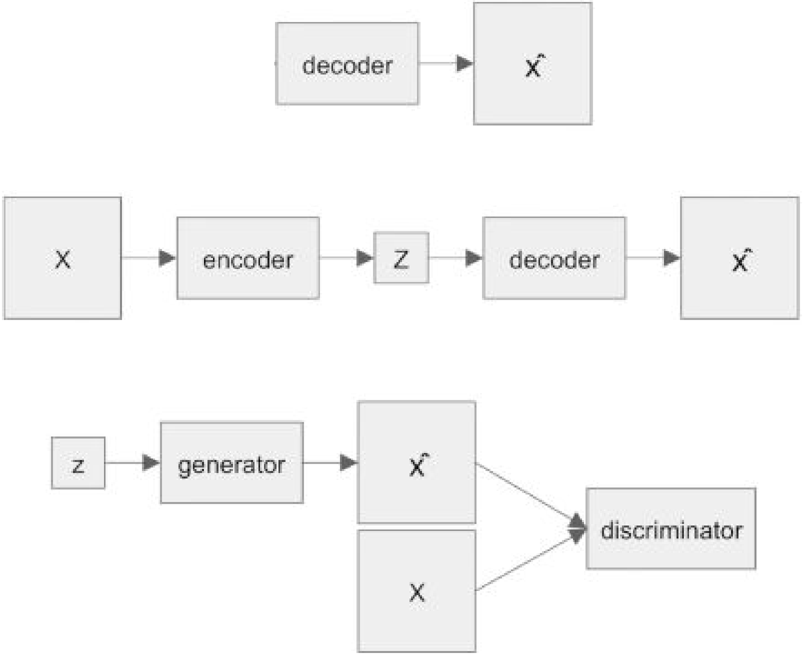
Taxonomy of Deep Generative Models



[Hochreiter & Schmidhuber 1997][Graves 2013][Sutskever, Vinyals, Le 2014][Cho+ 2014]
[Kingma, Welling 2014]
[Goodfellow+ 2014]

Three approaches to generative models

- Autoregression (e.g., LM), VAE, GAN



Autoregressive models (e.g. LM)
[Hochreiter & Schmidhuber 1997]
Graves [1308.0850]

Variational
Autoencoders (VAE)
Kingma and Welling
[1312.6114]

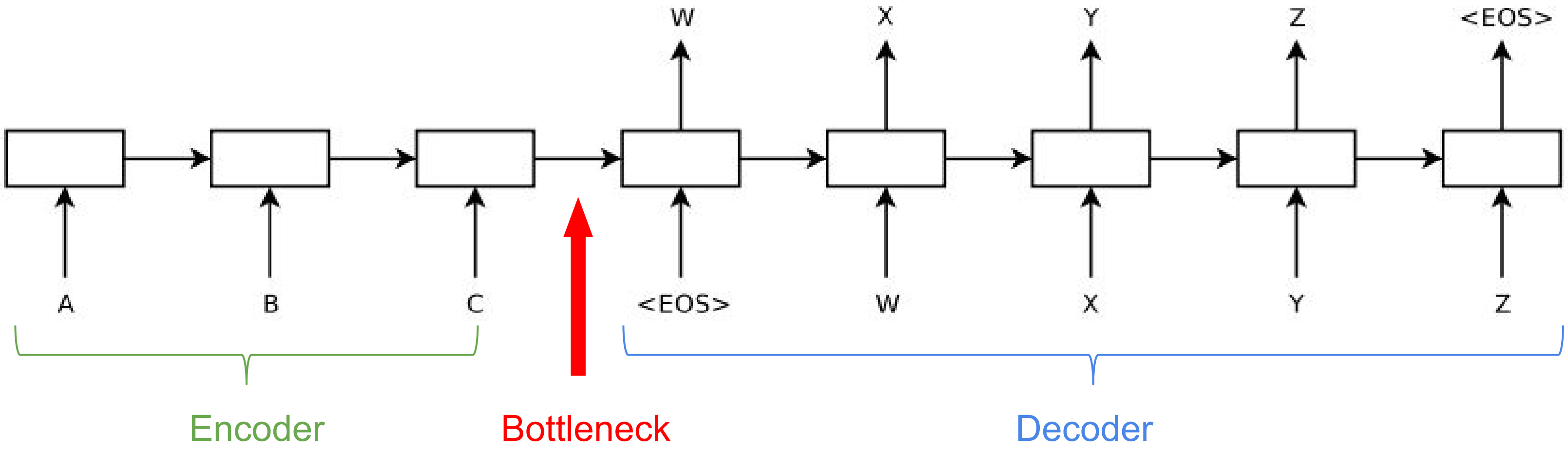
Generative Adversarial
Networks (GAN)
Goodfellow et [illegible] 406.2661]

Blog [Karpathy+ 2016]

[Sutskever, Vinyals, Le 2014]

Sequence to Sequence Models

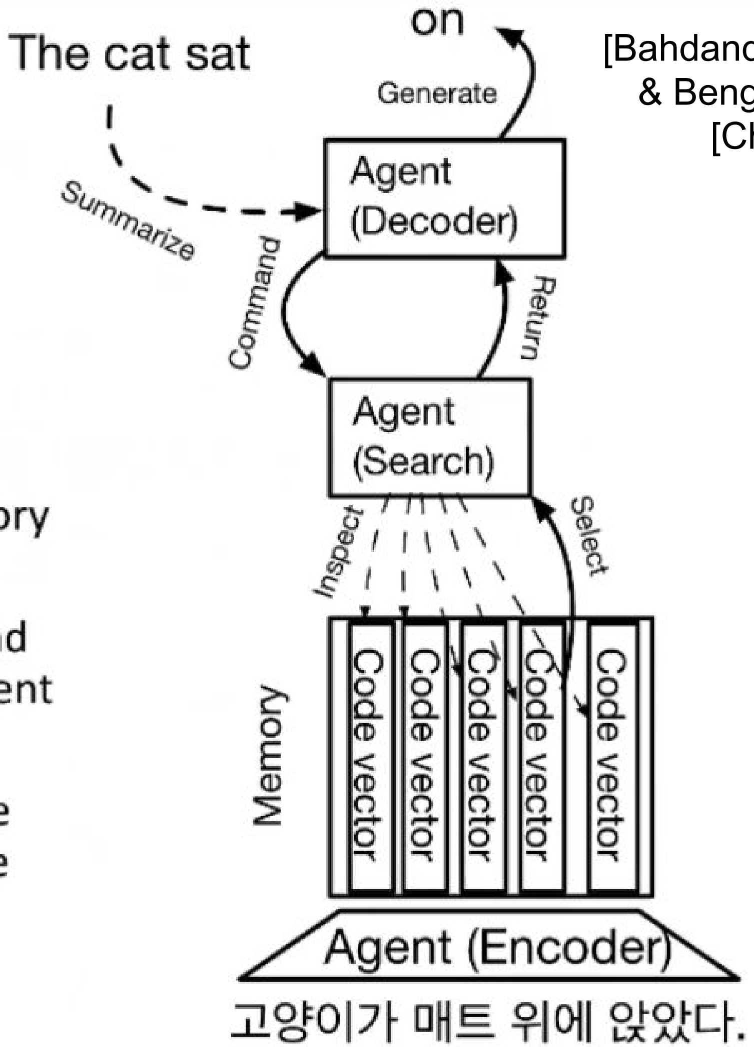
- Separate a sequence model in to encoder and decoder
- Improves a phrase-based SMT system by re-ranking top candidates
- Cannot perform well by itself due to the information bottleneck



Re-thinking sequence-to-sequence learning

- Cooperation among three agents

1. **Agent 1** (Encoder): transforms the source sentence into a set of code vectors in a memory
2. **Agent 2** (Search): searches for relevant code vectors in the memory based on the command from the Agent 3 and returns them to the Agent 3.
3. **Agent 3** (Decoder): observes the current state (previously decoded symbols), commands the Agent 2 to find relevant code vectors and generates the next symbol based on them.

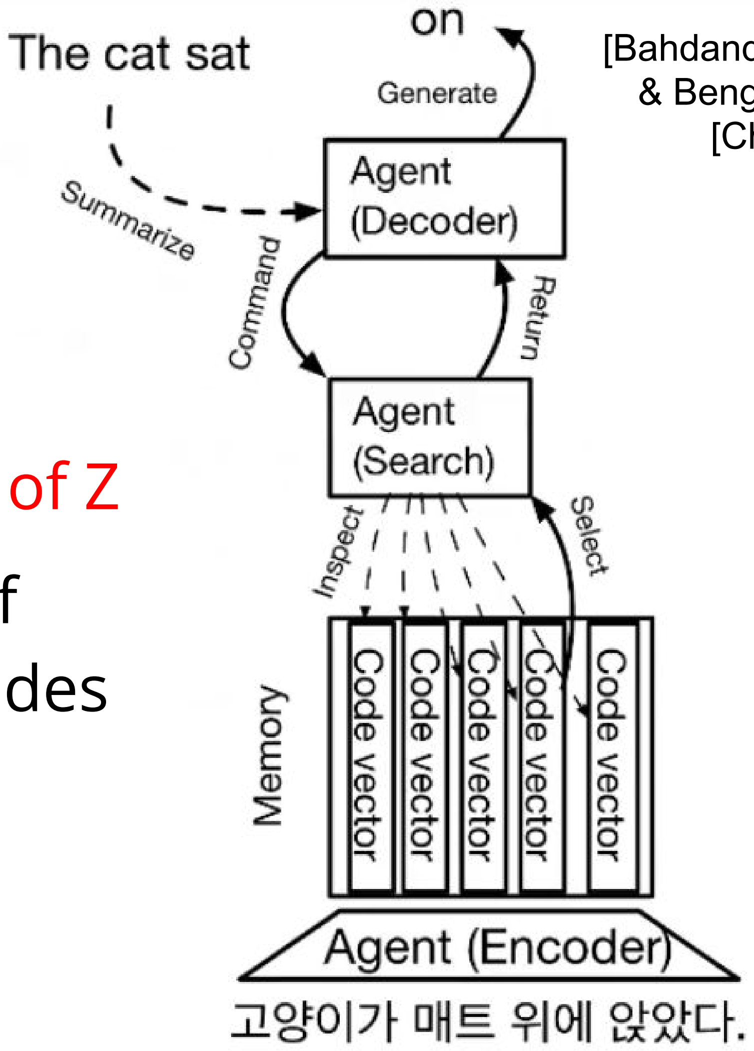


[Bahdandau, Cho & Bengio 2015]
[Cho 2017]

Re-thinking sequence-to-sequence learning

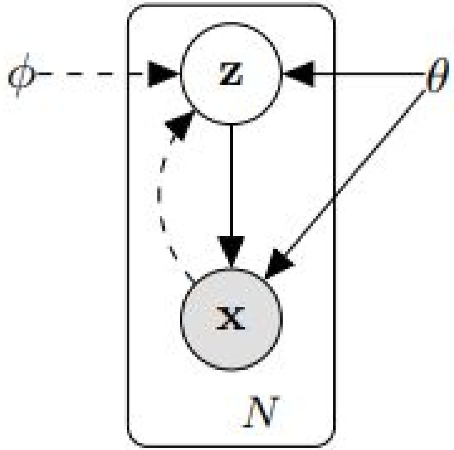
[Bahdandau, Cho & Bengio 2015]
[Cho 2017]

1. Don't generate from the ball of Z
2. Generate from a sequence of source token ids, which encodes the semantics of the target sentence



[Kingma & Welling 2014]

Variational Auto Encoder



- How can we perform efficient inference and learning in directed probabilistic models, in the presence of continuous latent variables with intractable posterior distributions, and large datasets?
  - R1: can train with standard stochastic gradient methods
  - R2: can inference efficiently with a lower bound estimator

$$E_{\mathbf{x}\sim p_d(\mathbf{x})}[-\log p(\mathbf{x})] < E_{\mathbf{x}}[\mathrm{E}_{q(\mathbf{z}|\mathbf{x})}[-\log(p(\mathrm{x}|\mathrm{z})]] + E_{\mathbf{x}}[\mathrm{KL}(q(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))]$$

- Compared to EM?

[Kingma & Welling 2014]

VAE = EM ?

- VAE

$$l(\theta) = E_{\mathbf{x}\sim p_d(\mathbf{x})}[-\log p(\mathbf{x})] < E_{\mathbf{x}}[\mathrm{E}_{q(\mathbf{z}|\mathbf{x})}[-\log(p(\mathbf{x}|\mathbf{z})]] + E_{\mathbf{x}}[\mathrm{KL}(q(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))] = l^{VAE}(\theta,q)$$

- EM
  - Variational methods approximate the probability P by adding extra parameters Q

$$l(\theta) = \log P(\mathrm{x}\,|\,\theta) = \log \sum_{\mathrm{z}} Q(\mathrm{z}) \frac{P(\mathrm{x},\mathrm{z}\,|\,\theta)}{Q(\mathrm{z})} \geq \sum_{\mathrm{z}} Q(\mathrm{z}) \log \frac{P(\mathrm{x},\mathrm{z}\,|\,\theta)}{Q(\mathrm{z})} = l^{EM}(\theta, Q)$$

    – Jensen's inequality: $\log \sum_{\mathbf{z}} P(\mathbf{z})\, f(\mathbf{z}) \geq \sum_{\mathbf{z}} P(\mathbf{z}) \log f(\mathbf{z})$

  - $l^{EM}(x)$ is an lower bound of $l(x)$, and the gap is a KL divergence.

[Kingma & Welling 2014]

VAE = EM

- VAE

$$\sum_z Q(z)\log P(x|z,\theta) \qquad\qquad \sum_z Q(z)\log\frac{P(z)}{Q(z)}$$

$$l(\theta) = E_{\mathbf{x}\sim p_d(\mathbf{x})}[-\log p(\mathbf{x})] < E_{\mathbf{x}}\boxed{[\mathrm{E}_{q(\mathbf{z}|\mathbf{x})}[-\log(p(\mathbf{x}|\mathbf{z})]]} + E_{\mathbf{x}}\boxed{[\mathrm{KL}(q(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))]} = l^{VAE}(\theta,q)$$

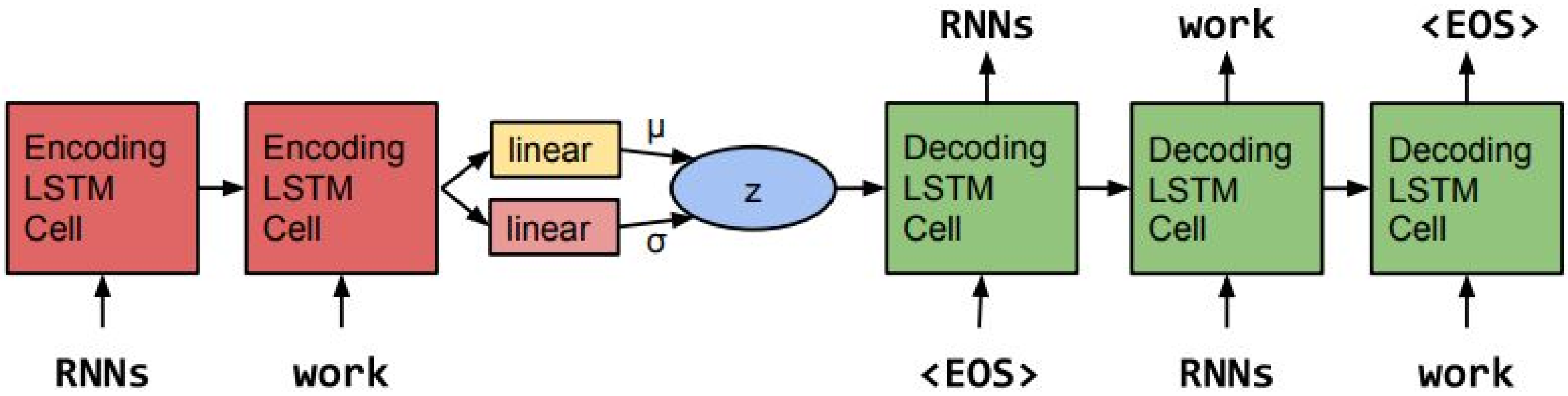
- EM

$$l(\theta) = \log P(\mathrm{x}\mid\theta) = \log\sum_z Q(\mathrm{z})\frac{P(\mathrm{x},\mathrm{z}\mid\theta)}{Q(\mathrm{z})} \geq \sum_z Q(\mathrm{z})\log\frac{P(\mathrm{x},\mathrm{z}\mid\theta)}{Q(\mathrm{z})} = l^{EM}(\theta,Q)$$

[Bowman+ 2016]

VAE with text

- Modeling P(x|z) -- without the conditioning inputs c



[Bowman+ 2016]

VAE with text

- (Again) Fundamentally limited to generate all sentences from a ball in R^n

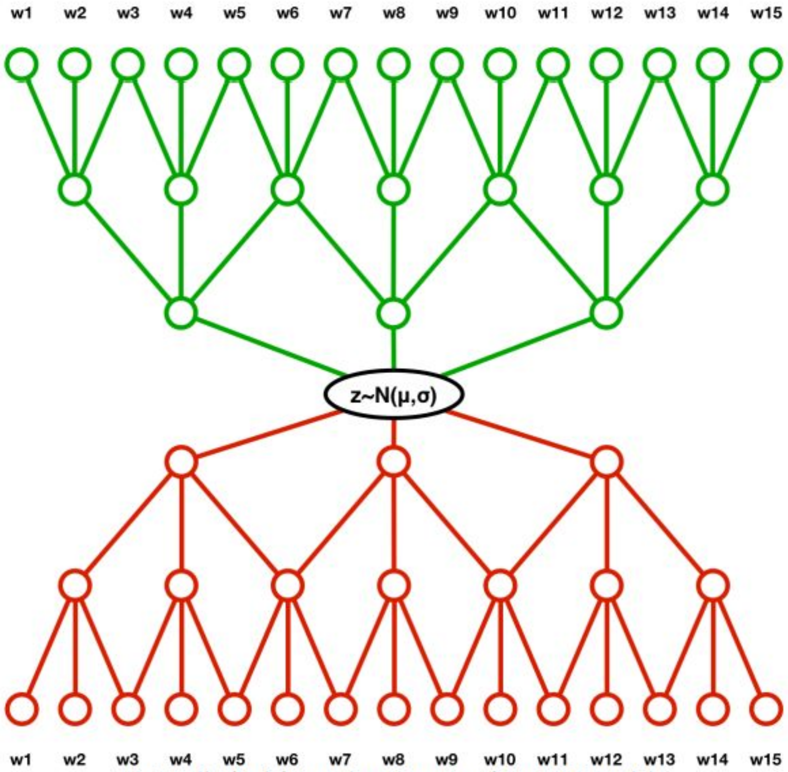
- Implementation
  - Need to add a KL term to prevent q(z|x) from overfitting
  - Need more tricks to prevent the KL term from collapsing
    - or else it falls back to a language model p(x|z)

$$
\begin{aligned}
\mathcal{L}(\theta; x) &= -\mathrm{KL}(q_\theta(\vec{z}|x)||p(\vec{z})) \\
&\quad + \mathbb{E}_{q_\theta(\vec{z}|x)}[\log p_\theta(x|\vec{z})] \\
&\leq \log p(x) \quad .
\end{aligned}
$$

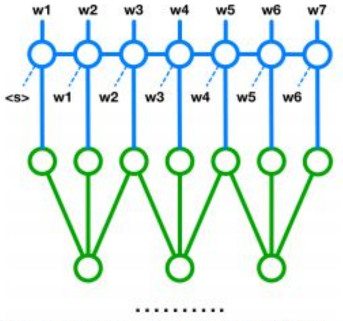
[Semeniuta+ 2016]

Convolutional VAE with text

- Adding conv layers reduces the KLD collapsing problem
- Also trains/runs faster



(a) Fully feed-forward component of our VAE model



(b) Hybrid model with LSTM decoder

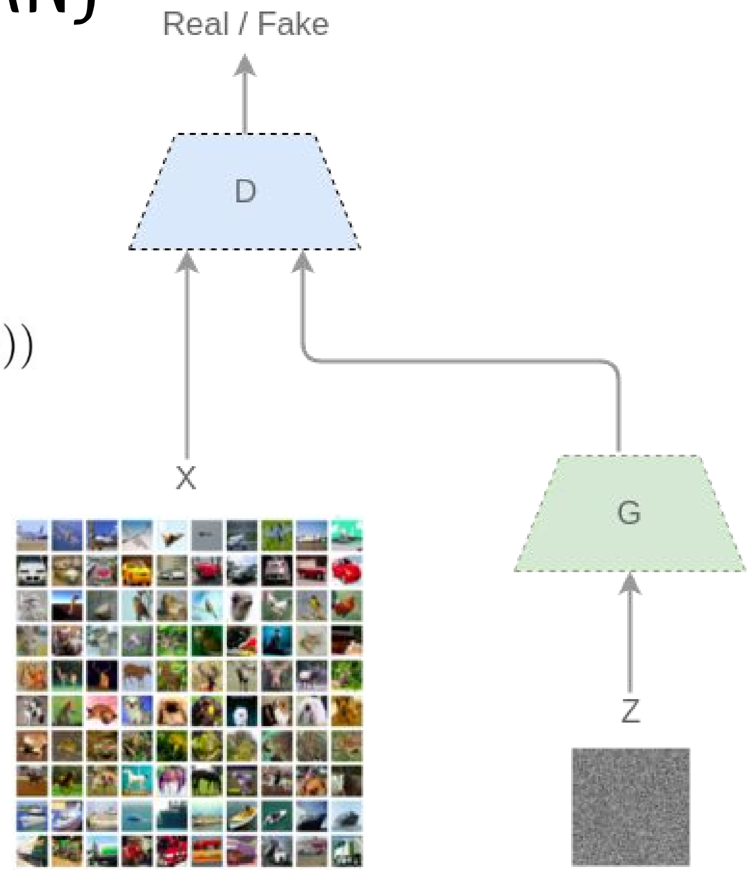
(c) Hybrid model with ByteNet decoder

[Goodfellow+ 2014]

Generative Adversarial Nets (GAN)

- Alternate between optimizing two models:

$$J^{(D)} = -\frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}} \log D(\boldsymbol{x}) - \frac{1}{2}\mathbb{E}_{\boldsymbol{z}} \log\left(1 - D\left(G(\boldsymbol{z})\right)\right)$$
$$J^{(G)} = -J^{(D)}$$

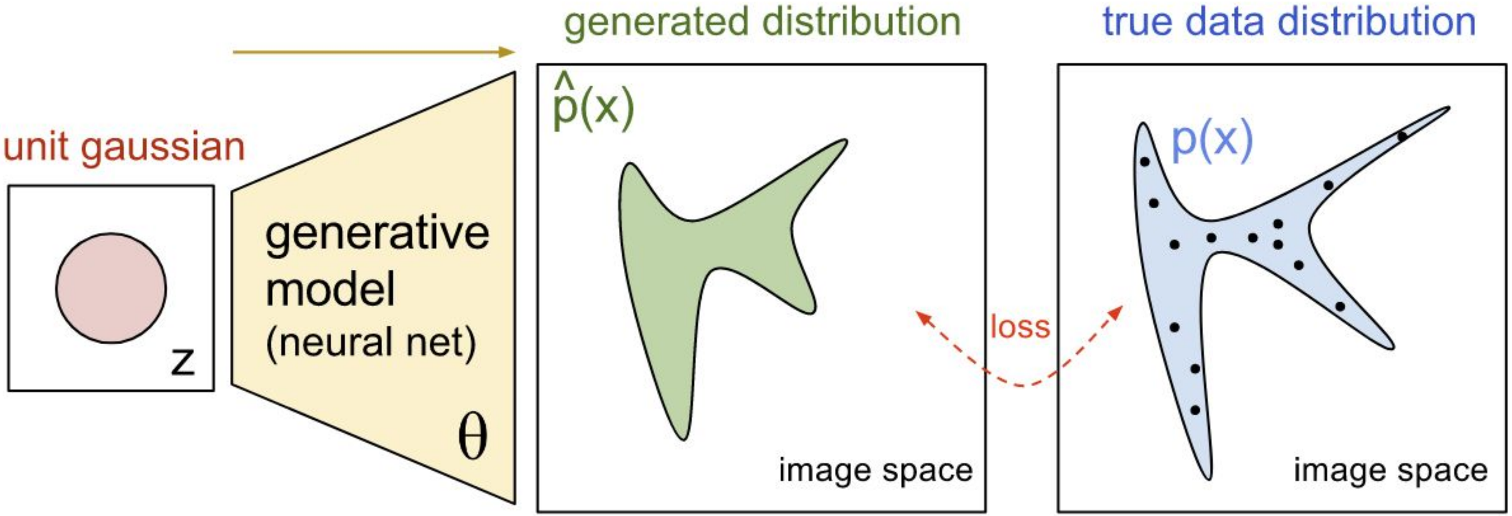


"the biggest breakthrough in Machine Learning in the last 1-2 decades."
-- Yann Lecun

[Goodfellow+ 2014]

GAN Intuition

- the green distribution starting out random and then the training process iteratively changing the parameters θ to stretch and squeeze it to better match the blue distribution.



[Goodfellow+ 2014]

GAN loss function

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

- The minimax game

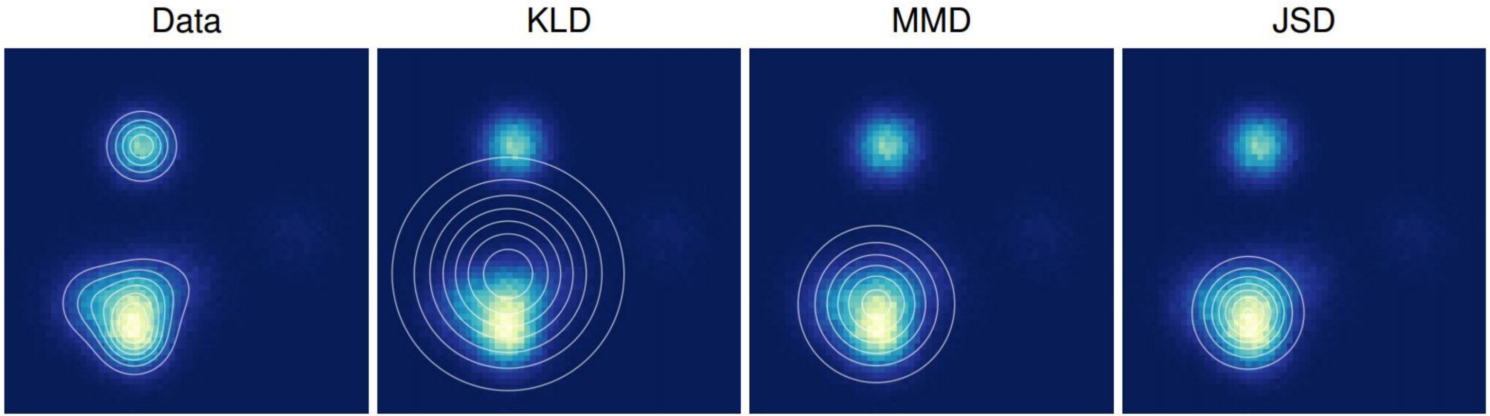
$$\min_G \max_D V(D, G) = \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z} \sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))].$$

- Is equivalent to minimizing JSD, if D is optimized more frequently than G

$$\begin{aligned} C(G) &= \max_D V(G, D) \\ &= \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}}[\log D_G^*(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z} \sim p_{\boldsymbol{z}}}[\log(1 - D_G^*(G(\boldsymbol{z})))] \\ &= -\log(4) + KL\left(p_{\text{data}} \middle\| \frac{p_{\text{data}} + p_g}{2}\right) + KL\left(p_g \middle\| \frac{p_{\text{data}} + p_g}{2}\right) \\ &= -\log(4) + 2 \cdot JSD\left(p_{\text{data}} \| p_g\right) \end{aligned}$$

JSD vs KLD

- Let's consider their behavior given an under capacity unimodal Gaussian model



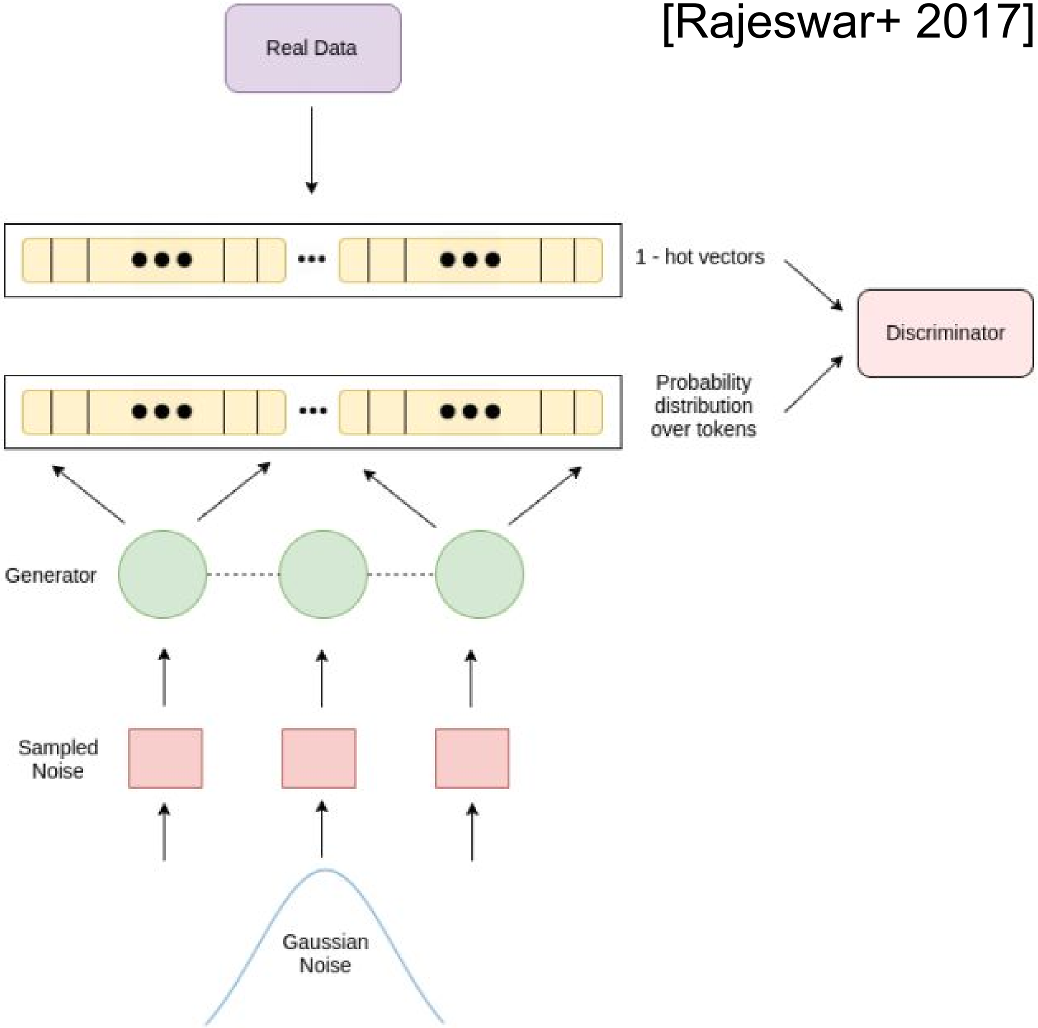
[1] L. Theis, A van den Oord, M. Bethge. A note on the evaluation of generative models. ICLR 2016.

- KLD makes sure that all the modes in data are covered
- JSD drops modes in data to avoid the penalty from covering "no data land"

[Rajeswar+ 2017]

(Naive) GAN on text

- To backprop through the discrete outputs simply forces the discriminator to operate on continuous valued output distributions



[Rajeswar+ 2017]

(Naive) GAN on text

Level	Model	PTB	CMU-SE
Word	LSTM	what everything they take everything away from .	<s>will you have two moment ? </s>
		may tea bill is the best chocolate from emergency .	<s>i need to understand deposit length . </s>
		can you show show if any fish left inside .	<s>how is the another headache ? </s>
		room service , have my dinner please .	<s>how there , is the restaurant popular this cheese ? </s>
	CNN	meanwhile henderson said that it has to bounce for.	<s>i 'd like to fax a newspaper . </s>
		I'm at the missouri burning the indexing manufacturing and through .	<s>cruise pay the next in my replacement . </s>
			<s>what 's in the friday food ? ? </s>

Table 4: Word level generations on the Penn Treebank and CMU-SE datasets

[Rajeswar+ 2017]

(Naive) GAN on text

POSITIVE	NEGATIVE
best and top notch newtonmom . good buy homeostasis money well spent kickass cosamin of time and fun . great britani ! I lovethis.	usuall the review omnium nothing non-functionable extreme crap-not working and eeeeeew a horrible poor imposing se400
QUESTION	STATEMENT
<s>when 's the friday convention on ? </s> <s>how many snatched crew you have ? </s> <s>how can you open this hall ? </s>	<s>i report my run on one mineral . </s> <s>we have to record this now . </s> <s>i think i deeply take your passenger .</s>

Table 5: Coditional generation of text. Top row shows generated samples conditionally trained on amazon review polarity dataset with two attributes 'positive' and 'negative'. Bottom row has samples conditioned on the 'question' attribute

Plan

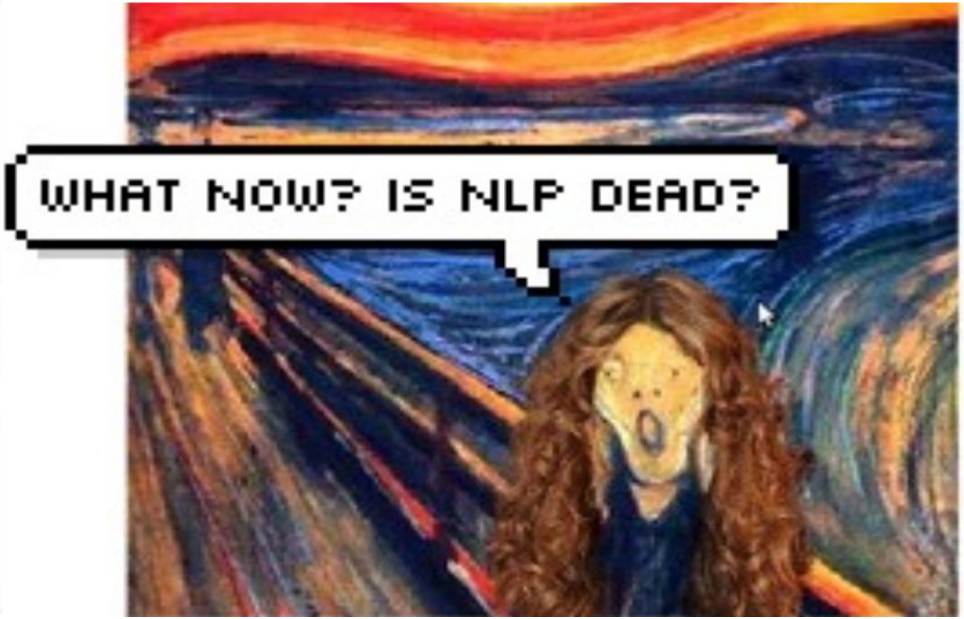
- Unconditional image/text generation models
  - Seq2Seq
  - Variational Auto-Encoder (VAE)
  - Generative Adversarial Net (GAN)

- Improved text generation models
  - Conditioned generation
  - Reinforcement Learning

[Lapata ACL 2017]

LSTM & Lapata’s scream

- LSTM has been applied to all kinds of NLP tasks, and has greatly simplified system designs



An LSTM is not enough

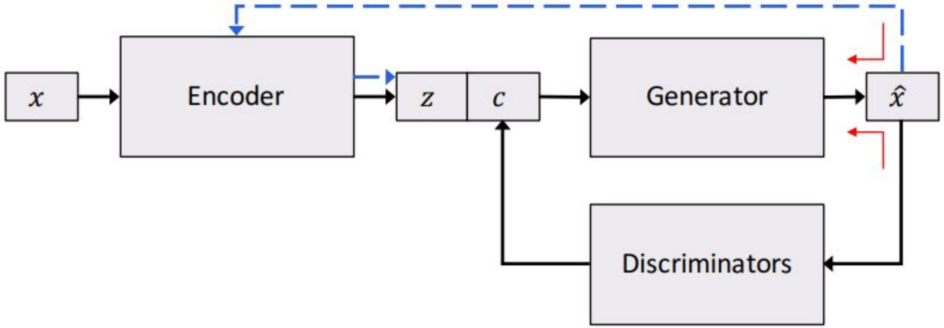
- To generate diverse and human-like sequences while respecting the conditional attributes
  - Benefit from unsupervised training
    - e.g., LM, AE, GAN
  - Respect additional conditioning inputs
    - Context which contain the semantics of the output (e.g., image or other structure)
    - e.g., writing style, sentiment, questions in an answer summarization task, etc.
  - Task-specific loss function
    - e.g., beyond MLE training (Reinforcement Learning)
    - e.g., evaluation beyond N-Grams

[Hu+ 2017]

Conditional VAE for style

- additional attributes c (such as tense, sentiment, style, etc.) can be injected into decoder

- VAE is extended with GAN that predict c to ensure that the generator respects them



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 $(\boldsymbol{z}, \boldsymbol{c})$

[Hu+ 2017]

Conditional VAE for style

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

Table 4. Samples by varying the unstructured code z given sentiment ("positive"/"negative") and tense ("past"/"present"/"future") code.

[Jiwei Li+ 2017]

The problems with MLE

- "(MLE) objective. Despite its success, this oversimplified training objective leads to problems: responses are dull, generic (Sordoni et al. , 2015 ; Serban et al. , 2016a ; Li et al. , 2016a), repetitive, and short-sighted (Li et al. , 2016d)"

- Exposure bias
    - Model is not exposed to its own errors during training

- Loss mismatch
    - Log-likelihood of gold sequences vs the task specific eval metric

[Zhao+ 2017]

Conditional VAE for diversity

- Task -- Switchboard (Godfrey and Holliman, 1997)
    - two-sided telephone conversations with manually transcribed speech and alignment

- Problem
    - Generic 'safe' response

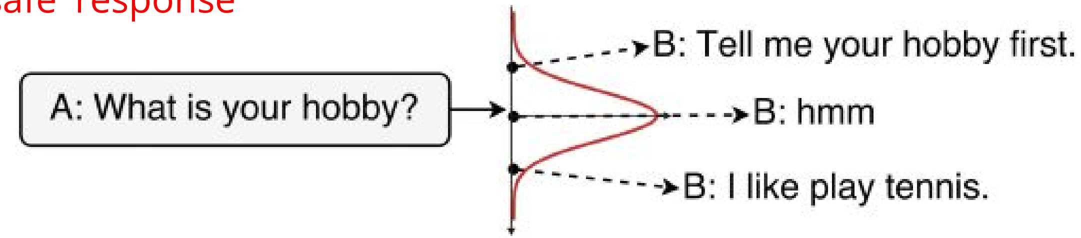
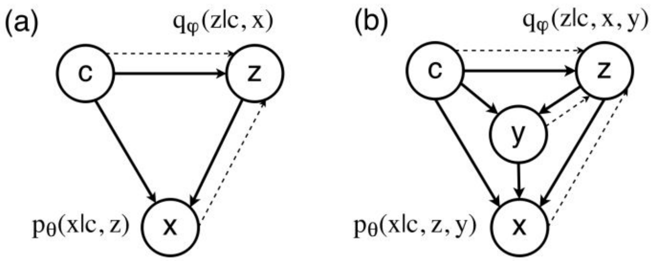


Figure 1: Given A's question, there exists many valid responses from B for different assumptions of the latent variables, e.g., B's hobby.

[Zhao+ 2017]

Conditional VAE for diversity

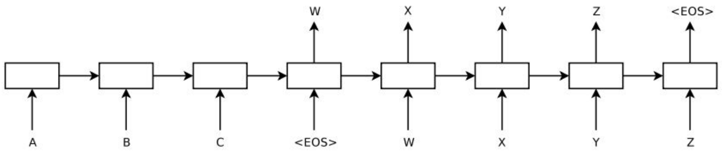
- Improve diversity by sampling z
    - Optionally add linguistic features y (e.g. dialog acts) to further constrain the style



1. Sample a latent variable z from the prior network $p_\theta(z|c)$.
2. Generate x through the response decoder $p_\theta(x|z, c)$.

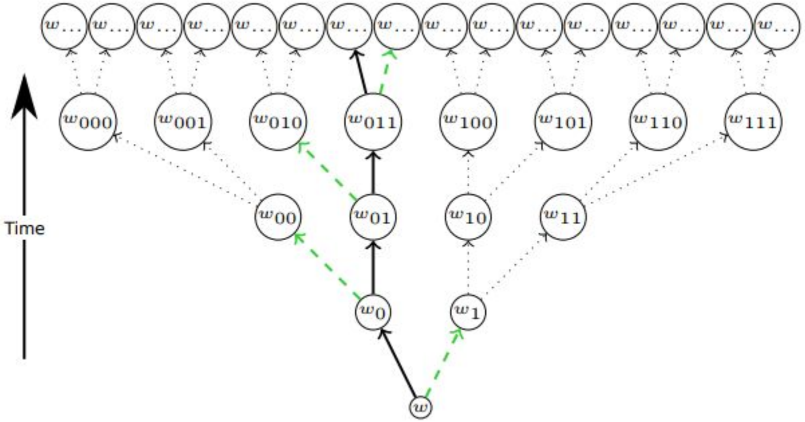
Example 1-Topic: Recycling **Context**: A: are they doing a lot of recycling out in Georgia? **Target-B** (statement): well at my workplace we have palaces for aluminium cans and we have a separate trash can for recyclable paper	
Baseline+Sampling	**kgCVAE+Greedy**
1. well I'm a graduate student and have two kids	1. (non-understand) pardon
2. well I was in last year and so we've had lots of recycling	2. (statement) oh you're not going to have a curbside pick up here
3. I'm not sure	3. (statement) okay I am sure about a recycling center
4. well I don't know I just moved here in new york	4. (yes-answer) yeah so

Exposure bias

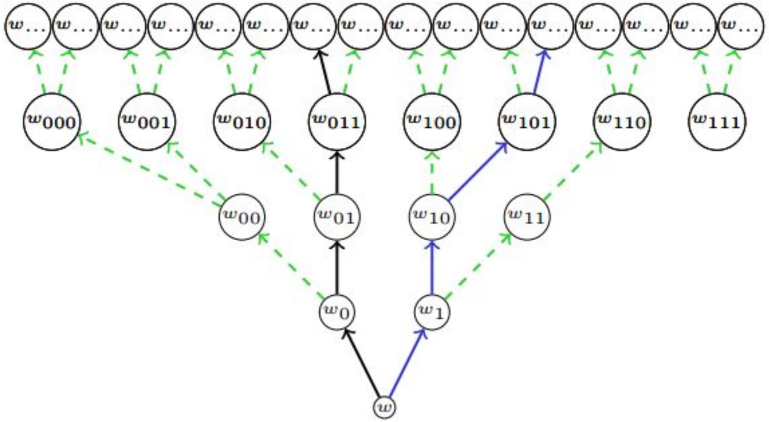


[Ranzato+ 2016]
[Lamb+ 2016]

- The model is exposed only to the ground truth but not its own predictions during MLE training
- Not an issue for reinforcement learning (RL), since the training sequences are generated by the model itself



Training with exposure bias



Training in expectation (Reinforce)

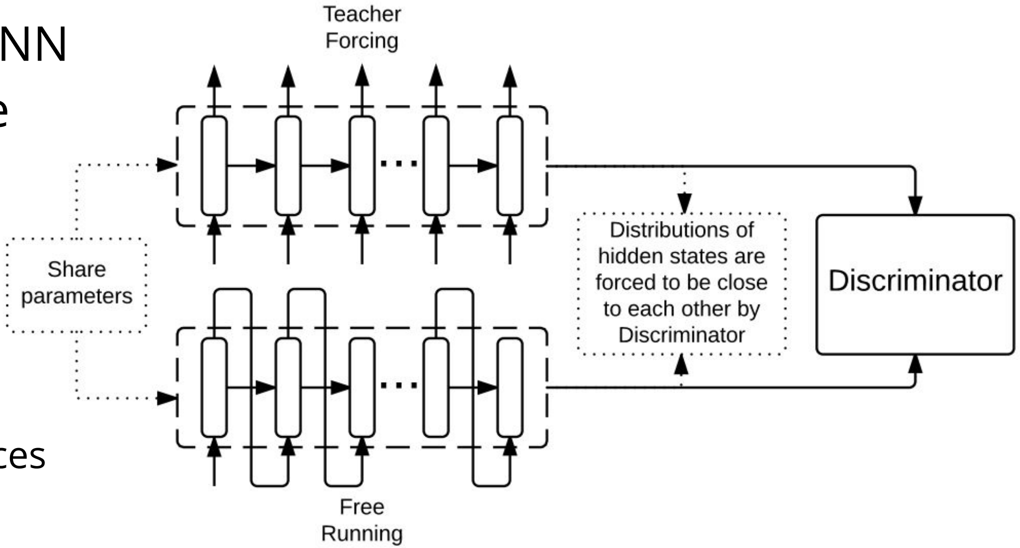
[Lamb+ 2016]

Professor forcing

- Adversarial learning to make RNN hidden states indistinguishable during training and inference

- No improvement for
  - word-level LM
  - speech synthesis of short sequences

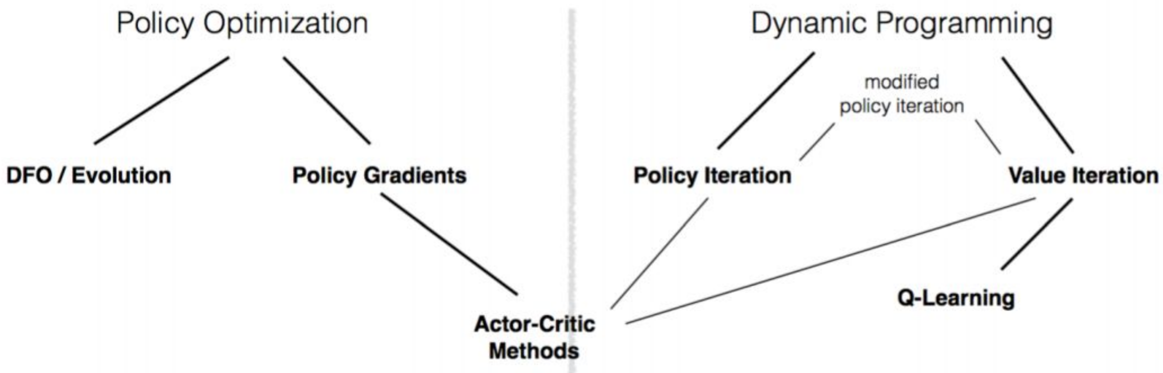
- Main disadvantages
  - overhead of training the discriminator



The RL landscape

Book [Sutton & Barto 1998]
NIPS [Abbeel & Schulman 2016]

- Policy gradient methods
  - Simple, flexible
- Bootstrapping methods
  - Data efficient

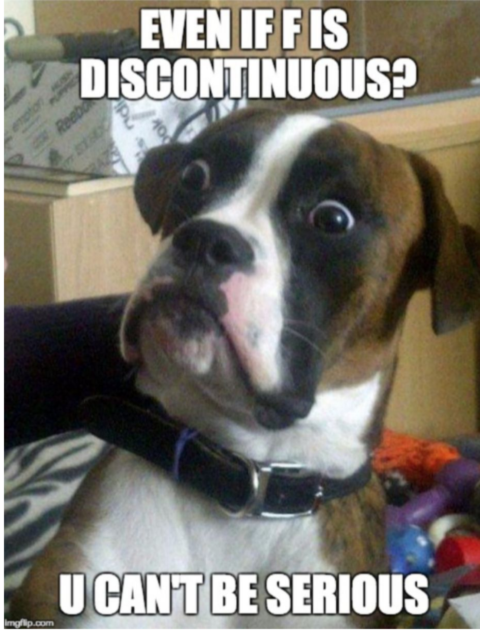


TRPO

PPO

- Trust region
  - Stable training
- Experience replay
  - Data efficient
  - Stable training

Policy Gradient (REINFORCE)



- MLE optimizes log likelihood of approximate gold programs

$$J^{ML}(\theta) = \sum_q \log P(a_{0:T}^{best}(q)|q, \theta)$$

- RL optimizes the expected reward under a stochastic policy P

$$J^{RL}(\theta) = \sum_q \mathbb{E}_{P(a_{0:T}|q,\theta)}[R(q, a_{0:T})]$$

- The gradient is almost the same as that for MLE except for a weight P(R-B)

$$\nabla_\theta J^{RL}(\theta) = \sum_q \sum_{a_{0:T}} \boxed{P(a_{0:T}|q, \theta)[R(q, a_{0:T}) - B(q)]} \nabla_\theta \log P(a_{0:T}|q, \theta)$$

- The baseline does not change the solution but improves convergences, e.g.,

$$B(q) = \sum_{a_{0:T}} P(a_{0:T}|q, \theta) R(q, a_{0:T})$$

[Williams 1992]

[Ranzato+ 2016]

Policy Gradient for text generation

- Other (naive) approaches to deal with exposure bias
  - DAD: at each step randomly pick from the ground truth data or the model prediction
  - E2E: at time step t + 1 we propagate as input the top k words predicted at the previous time step instead of the ground truth word

PROPERTY	XENT (MLE)	DAD	E2E	MIXER (REINFORCE)
avoids exposure bias	No	Yes	Yes	Yes
end-to-end	No	No	Yes	Yes
sequence level	No	No	No	Yes
TASK				
summarization	13.01	12.18	12.78	**16.22**
translation	17.74	20.12	17.77	**20.73**
image captioning	27.8	28.16	26.42	**29.16**

Book [Sutton & Barto 1998]
NIPS [Abbeel & Schulman 2016]

Challenges of applying RL

- Large search space (sparse rewards)
  - ==> Supervised pretraining (MLE)
  - ==> Systematic exploration
- Credit assignment (delayed reward)
  - ==> bootstrapping (E.g., train a value function to estimate the future reward)
  - ==> rollout n-steps
- Train speed (cold start)
  - ==> experience replay
- Train stability (multi-epoch optimization)
  - ==> trust region approaches (e.g., PPO)
  - ==> experience replay

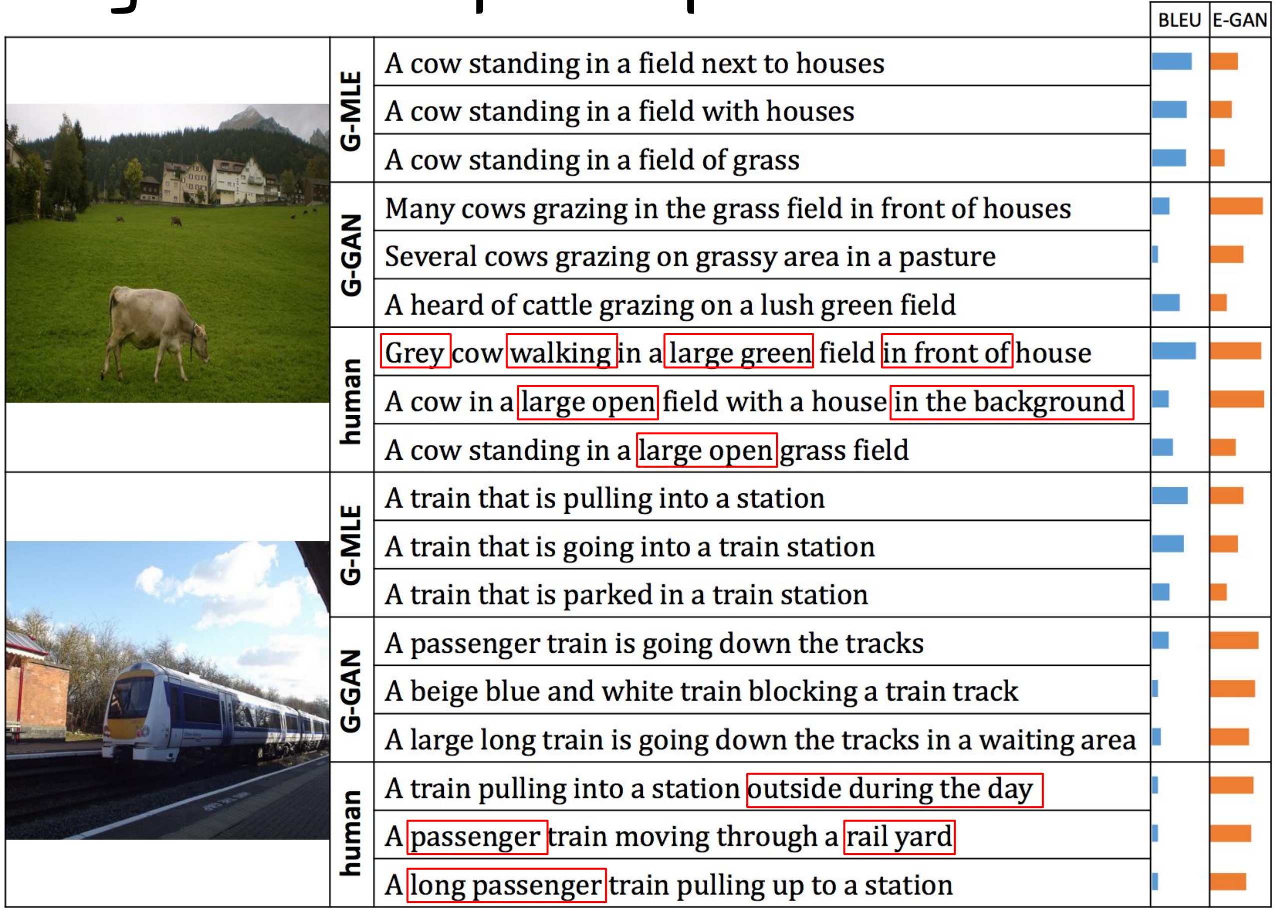
MLE training favors generic 'safe' responses

[Vinyals+ 2015]
[Dai+ 2017]

- Natural responses often contain low frequency words

- meanwhile, P(x) only decreases with longer x
  - E.g. NMT uses a coverage penalty to reduce this problem

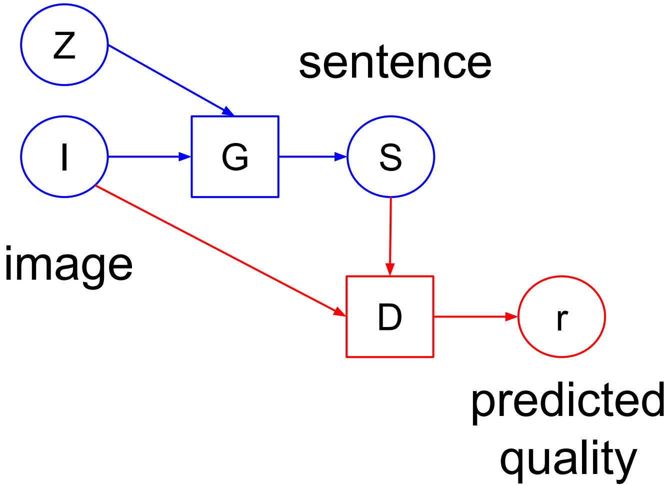
- N-gram based metrics (e.g. BLEU) favors MLE results



[Dai+ 2017]

Conditional GAN for natural responses

- Train a discriminator,
    - which provides scores that correlate much better with the human judgement than any of the automatic metrics
- Generation is conditioned on the image *I*
    - Z only controls the diversity (not semantics) of the generation
- Apply policy gradient to (correctly) pass the discriminator's reward to the generator
    - avoids the exposure biases of MLE training
- Apply Monte Carlo rollout to estimate the future reward of an action
    - improves sample efficiency
    - avoids vanishing gradients

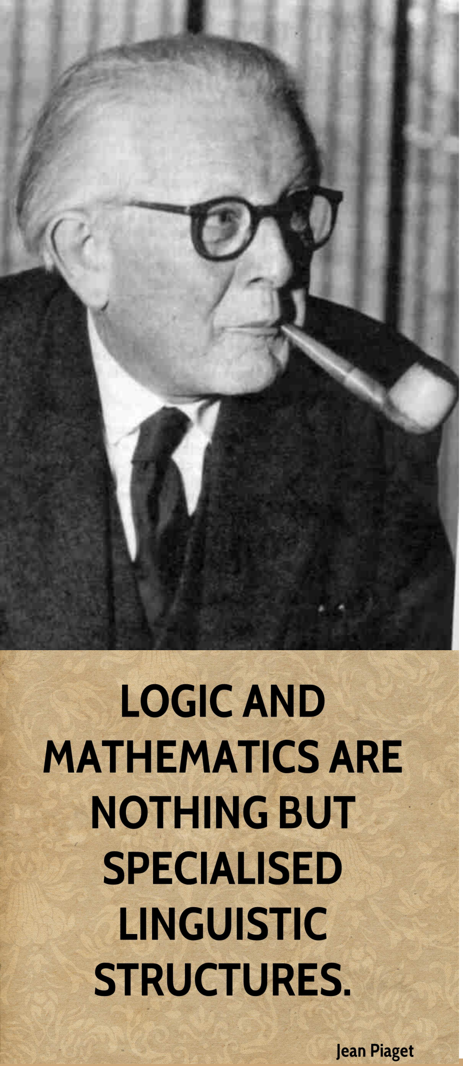


Conditional GAN for natural responses

- More details appears in the GAN samples

	human	G-GAN, z_1	G-GAN, z_2	G-MLE
	people are on motorcycles. there are green cars behind them. the signs are all brown with chinese written on it.	men are riding on a motorcycle. the man is $\boxed{\text{wearing tan boots}}$, and a $\boxed{\text{white and blue jacket}}$ with $\boxed{\text{beige stripes on}}$. the street is made of $\boxed{\text{cobblestone}}$. there are $\boxed{\text{tall bright green}}$ trees on the sidewalk.	two people are riding motorcycles. there are many trees on the sidewalk. there is a red and white painted letter on the side of the ledge. tall buildings are on the background.	a man is riding a bike. there are trees on the sidewalk. there are people walking on the sidewalk. there is a tall building in the background.
	A baseball player is swinging a bat. He is wearing a black helmet and a black and white uniform. A catcher is behind him wearing a gray uniform. The catcher has a brown glove on his hand. Two men can be seen standing behind a green fence.	a baseball player in a white and blue uniform is holding a white bat. there is a umpire behind the batter in the blue and white uniform. he is getting ready to catch the ball. there is a crowd of $\boxed{\text{people behind him watching}}$ him.	men are on a baseball field $\boxed{\text{on a sunny day.}}$ the player is wearing a black and white uniform. there is a catcher behind him. the field is green with brown dirt and white shiny lines.	a baseball player is standing on a baseball field. he is wearing a blue helmet on his head. the catcher is wearing a black and gray uniform. the court is green with white lines.

Figure 8: Examples of images with different descriptive paragraphs generated by a human, *G-GAN* with different **z**, and *G-MLE*.



ACL [Liang+ 2017]

Language, Translation & Control

1) Natural languages are programming languages to control human behavior

2) For machines and human to understand each other, they just need translation models trained with control theory

NIPS [LeCun 2016]

LeCun's Cake

- RL is needed to optimize the right objective
  - we don't really care about likelihoods
- Supervised learning is good for pretraining
  - to avoid the cold start problem in RL
  - to deal with large search spaces
- Training should be mostly unsupervized for good representation learning
  - to fill the huge capacities of DNNs

"Pure" Reinforcement Learning (cherry)
- The machine predicts a scalar reward given once in a while.
- **A few bits for some samples**

Supervised Learning (icing)
- The machine predicts a category or a few numbers for each input
- Predicting human-supplied data
- **10→10,000 bits per sample**

Unsupervised/Predictive Learning (cake)
- The machine predicts any part of its input for any observed part.
- Predicts future frames in videos
- **Millions of bits per sample**

Thanks

[Kingma & Welling 2014]

VAE = EM + H(Q)

- VAE

$$\sum_z Q(z)\log\frac{P(x|z,\theta)}{Q(z)} - H(Q) \qquad\qquad \sum_z Q(z)\log\frac{P(z)}{Q(z)}$$

$$l(\theta) = E_{\mathbf{x}\sim p_d(\mathbf{x})}[-\log p(\mathbf{x})] < E_{\mathbf{x}}\boxed{[\mathrm{E}_{q(\mathbf{z}|\mathbf{x})}[-\log(p(\mathbf{x}|\mathbf{z})]]} + E_{\mathbf{x}}\boxed{[\mathrm{KL}(q(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))]} = l^{VAE}(\theta,q)$$

- EM

$$l(\theta) = \log P(\mathrm{x}\,|\,\theta) = \log\sum_z Q(\mathrm{z})\frac{P(\mathrm{x},\mathrm{z}\,|\,\theta)}{Q(\mathrm{z})} \geq \sum_z Q(\mathrm{z})\log\frac{P(\mathrm{x},\mathrm{z}\,|\,\theta)}{Q(\mathrm{z})} = l^{EM}(\theta,Q)$$

ConvNets & Paragios' Depression

- "Well, I am not that old, but I have been involved with computer vision for almost two decades now." ... "There were always trends and dominant topics in the field"
- "this is far from being the case anymore." ... "one can question what is the 'added' scientific value."
- "there are three deep learning stages: denial, doubt, and acceptance/adoption! I guess I navigate on the ocean between the last two stages without a compass."


Computer Vision Research: The deep "depression"

Published on June 5, 2016

Nikos Paragios + Follow

Distinguished Professor @CentraleSupelec / @University Paris-Sa...

12 articles

1,198 73 226

[Zhao+ 2017]

Conditional VAE for diversity

- Improved quality v.s. LSTM baseline
- improve KL by predicting input BOW

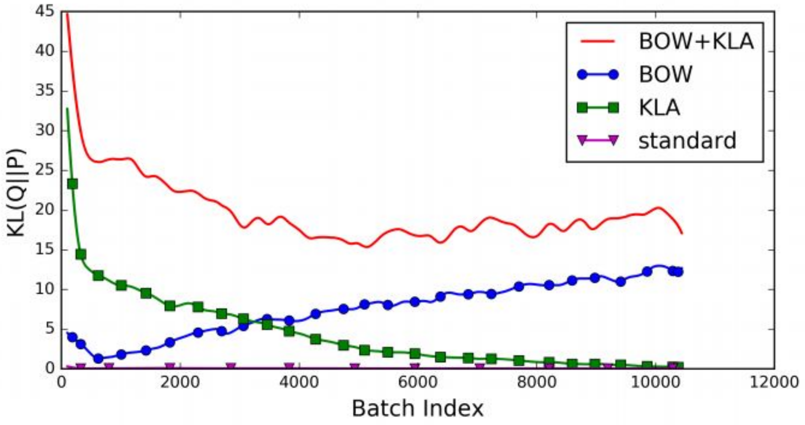


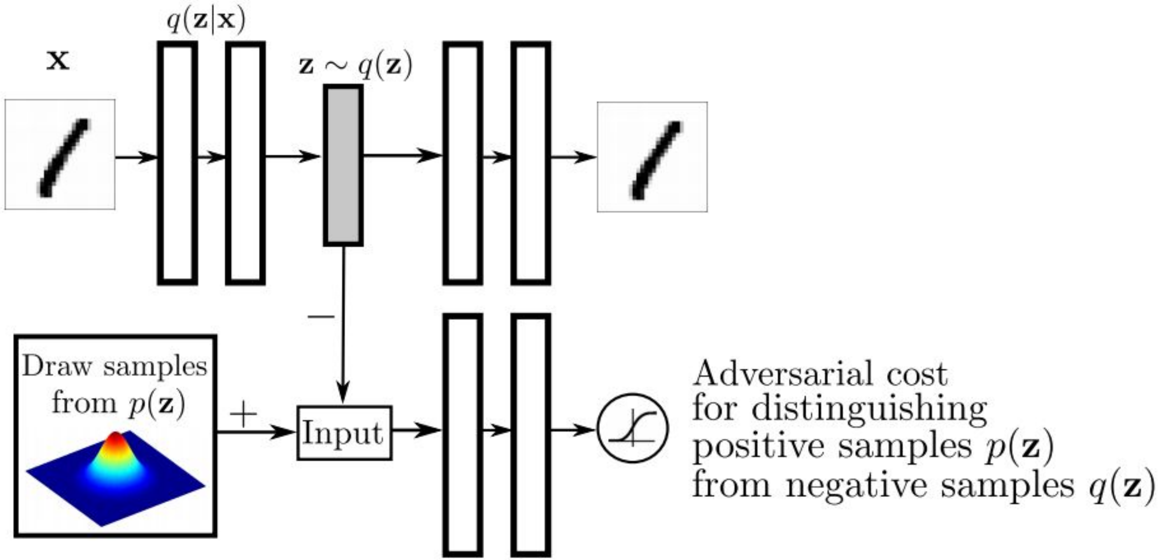
Figure 6: The value of the KL divergence during training with different setups on Penn Treebank.

Metrics	Baseline	CVAE	kgCVAE
perplexity (KL)	35.4 (n/a)	20.2 (11.36)	16.02 (13.08)
BLEU-1 prec	0.405	0.372	**0.412**
BLEU-1 recall	0.336	0.381	**0.411**
BLEU-2 prec	0.300	0.295	**0.350**
BLEU-2 recall	0.281	0.322	**0.356**
BLEU-3 prec	0.272	0.265	**0.310**
BLEU-3 recall	0.254	0.292	**0.318**
BLEU-4 prec	0.226	0.223	**0.262**
BLEU-4 recall	0.215	0.248	**0.272**
A-bow prec	0.387	**0.389**	0.373
A-bow recall	0.337	**0.361**	0.336
E-bow prec	0.701	0.705	**0.711**
E-bow recall	0.684	0.709	**0.712**
DA prec	**0.736**	0.704	0.721
DA recall	0.514	**0.604**	0.598

[Makhzani+ 2015]

Adversarial Autoencoders

- Matching the aggregated posterior to the prior ensures that generating from any part of prior space results in meaningful samples



[Goldberg 2015]

one morning,
as a parsing researcher woke
from an uneasy dream,
he realized that
he somehow became an expert
in distributional lexical semantics.

[Goldberg 2015]

to summarize

- Magic is bad. Understanding is good.
  Once you Understand you can control and improve.
- Word embeddings are just distributional semantics in disguise.
- Need to think of what you actually want to solve.
  --> focus on a specific task!
- Inputs >> fancy math.
- Look beyond just words. • Look beyond just English.