DSC291: Advanced Statistical Natural Language Processing

Advanced Topics: Generative Adversarial Learning

Zhiting Hu

Lecture 19, May 31, 2022

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

Outline

- GANs (for text)

- 3 Paper presentations (15 x 3 mins)

Recap: Generative Adversarial Nets (GANs)

- Generative model $\boldsymbol{x} = G_\theta(\boldsymbol{z}), \ \ \boldsymbol{z} \sim p(\boldsymbol{z})$
  - Maps noise variable $\boldsymbol{z}$ to data space $\boldsymbol{x}$
  - Defines an implicit distribution over $\boldsymbol{x}$: $p_{g_\theta}(\boldsymbol{x})$
- Discriminator $D_\phi(\boldsymbol{x})$
  - Output the probability that $\boldsymbol{x}$ came from the data rather than the generator

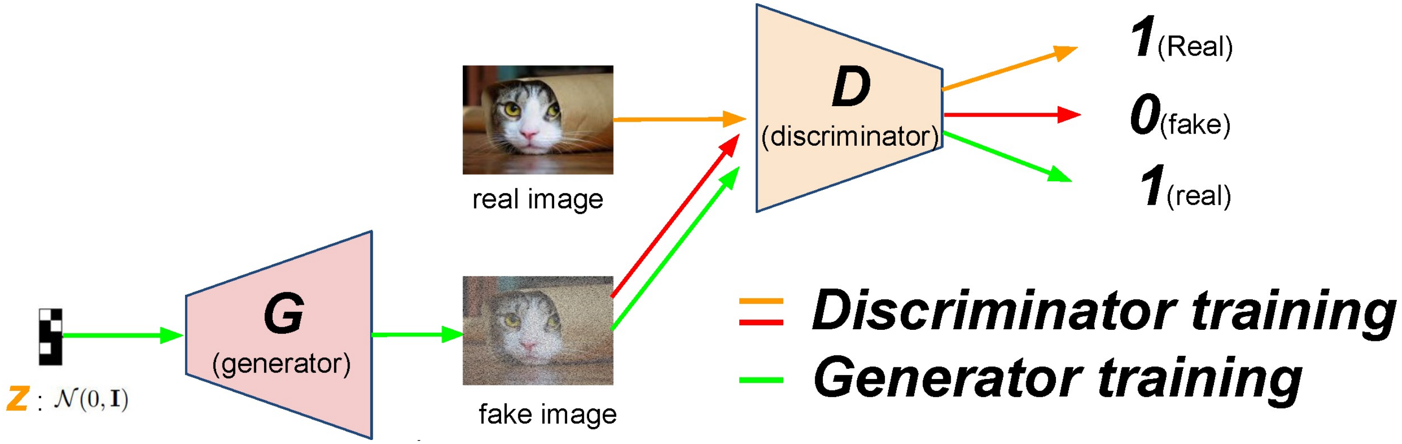


Figure courtesy: Kim

Recap: Generative Adversarial Nets (GANs)

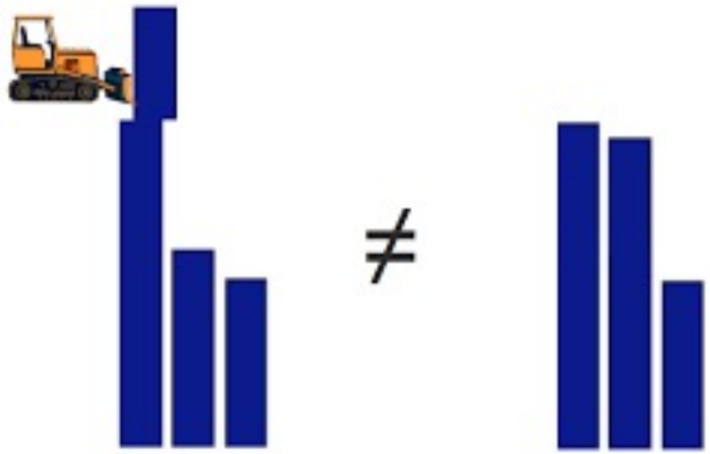
- Learning
  - A minimax game between the generator and the discriminator
  - Train D to maximize the probability of assigning the correct label to both training examples and generated samples
  - Train G to fool the discriminator

$$\max_D \mathcal{L}_D = \mathbb{E}_{\boldsymbol{x}\sim p_{data}(\boldsymbol{x})}\left[\log D(\boldsymbol{x})\right] + \mathbb{E}_{\boldsymbol{x}\sim G(\boldsymbol{z}),\boldsymbol{z}\sim p(\boldsymbol{z})}\left[\log(1 - D(\boldsymbol{x}))\right]$$
$$\min_G \mathcal{L}_G = \mathbb{E}_{\boldsymbol{x}\sim G(\boldsymbol{z}),\boldsymbol{z}\sim p(\boldsymbol{z})}\left[\log(1 - D(\boldsymbol{x}))\right].$$

Figure courtesy: Kim

Recap: Wasserstein GAN (WGAN)

- If our data are on a low-dimensional manifold of a high dimensional space, the model's manifold and the true data manifold can have a negligible intersection in practice
- The loss function and gradients may not be continuous and well behaved
- The Wasserstein Distance is well defined
  - Earth Mover's Distance
  - Minimum transportation cost for making one pile of dirt in the shape of one probability distribution to the shape of the other distribution



[Arjovsky et al., 2017] Slide adapted from bhiksha

Wasserstein GAN (WGAN)

- Objective

$$W\left(p_{data}, p_g\right) = \frac{1}{K} \sup_{||D||_L \leq K} \mathrm{E}_{x \sim p_{data}} [D(x)] - \mathrm{E}_{x \sim p_g} [D(x)]$$

- $||D||_L \leq K$: K- Lipschitz continuous
- Use gradient-clipping to ensure D has the Lipschitz continuity

Progressive GAN

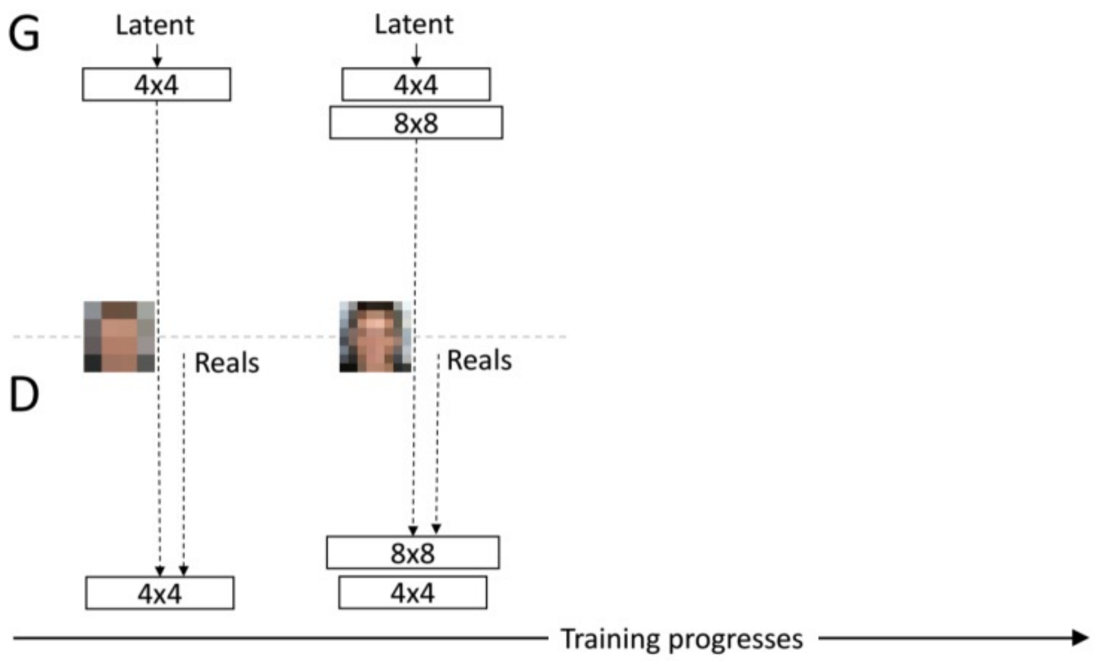
Low resolution images

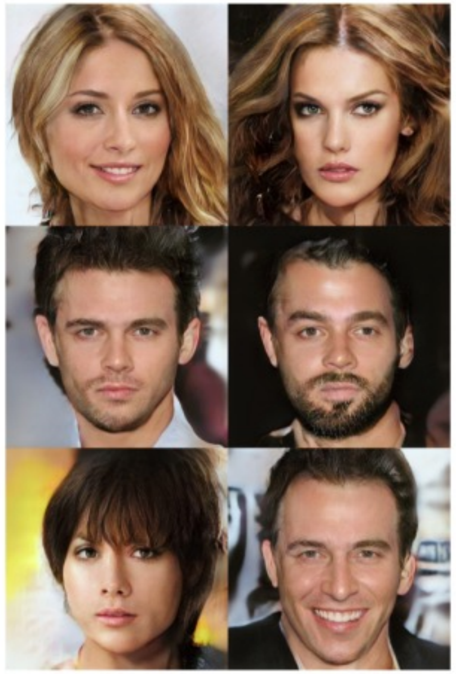
[Karras et al., 2018]

Progressive GAN

Low resolution images

add in additional layers





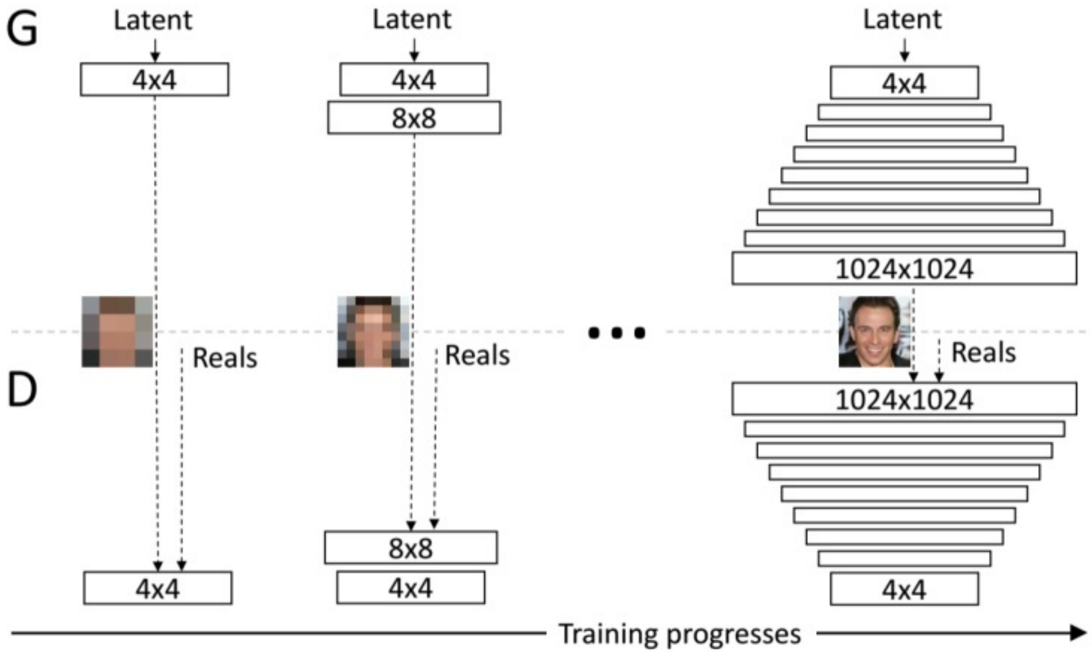
[Karras et al., 2018]

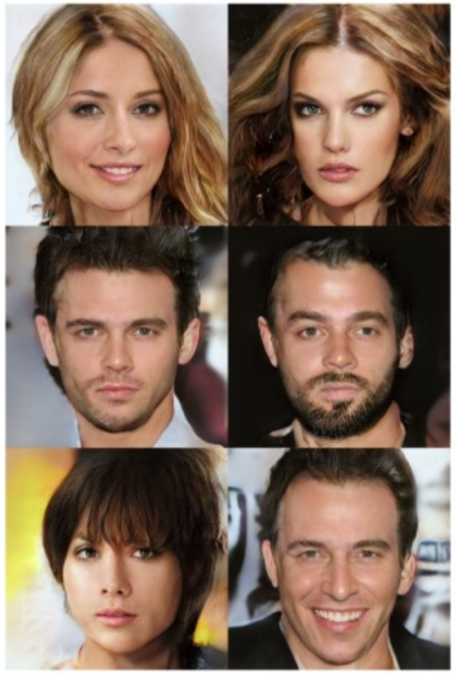
Progressive GAN

Low resolution images

add in additional layers

High resolution images





[Karras et al., 2018]

BigGAN

[Brock et al., 2018]

BigGAN

- GANs benefit dramatically from scaling

[Brock et al., 2018]

BigGAN

- GANs benefit dramatically from scaling
- 2x – 4x more parameters
- 8x larger batch size
- Simple architecture changes that improve scalability

[Brock et al., 2018]

BigGAN

- GANs benefit dramatically from scaling
- 2x – 4x more parameters
- 8x larger batch size
- Simple architecture changes that improve scalability

[Brock et al., 2018]

BigGAN

- GANs benefit dramatically from scaling
- 2x 4x more parameters
- 8x
- Sir

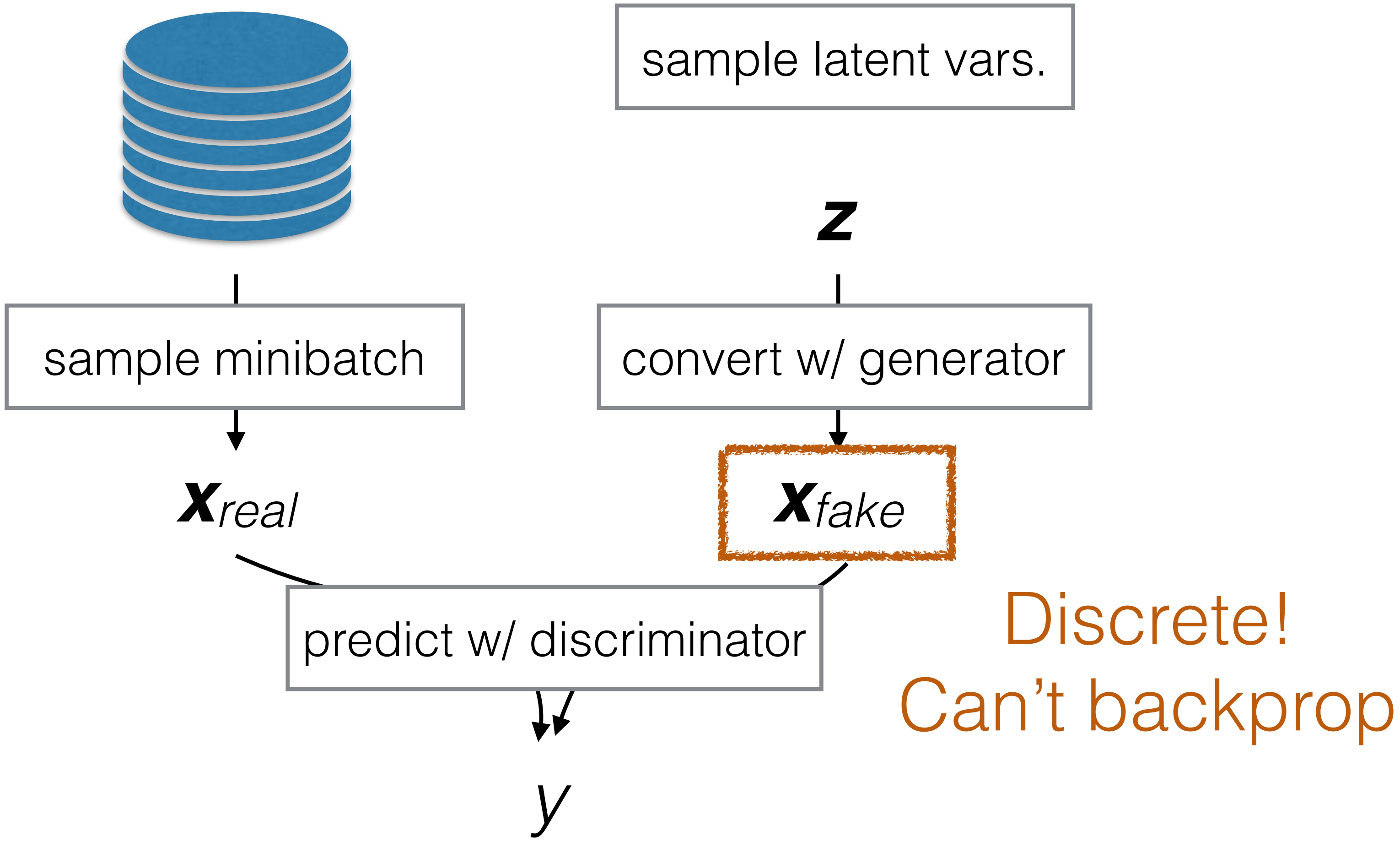
[Brock et al., 2018]

GANs for Text

Applications of GAN Objectives to Language

- GANs for Language Generation (Yu et al. 2017)
- GANs for MT (Yang et al. 2017, Wu et al. 2017, Gu et al., 2017)
- GANs for Dialogue Generation (Li et al. 2016)

Problem! Can't Backprop through Sampling



Solutions

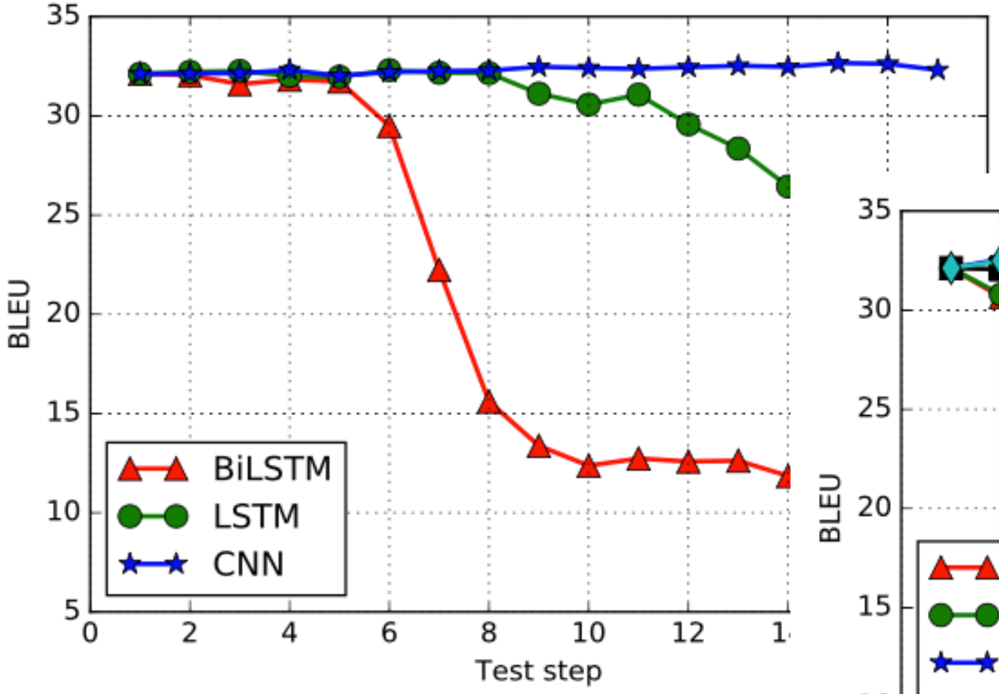
- Policy gradient reinforcement learning methods (e.g. Yu et al. 2016)
- Reparameterization trick for latent variables using Gumbel softmax (Gu et al. 2017)

Discriminators for Sequences

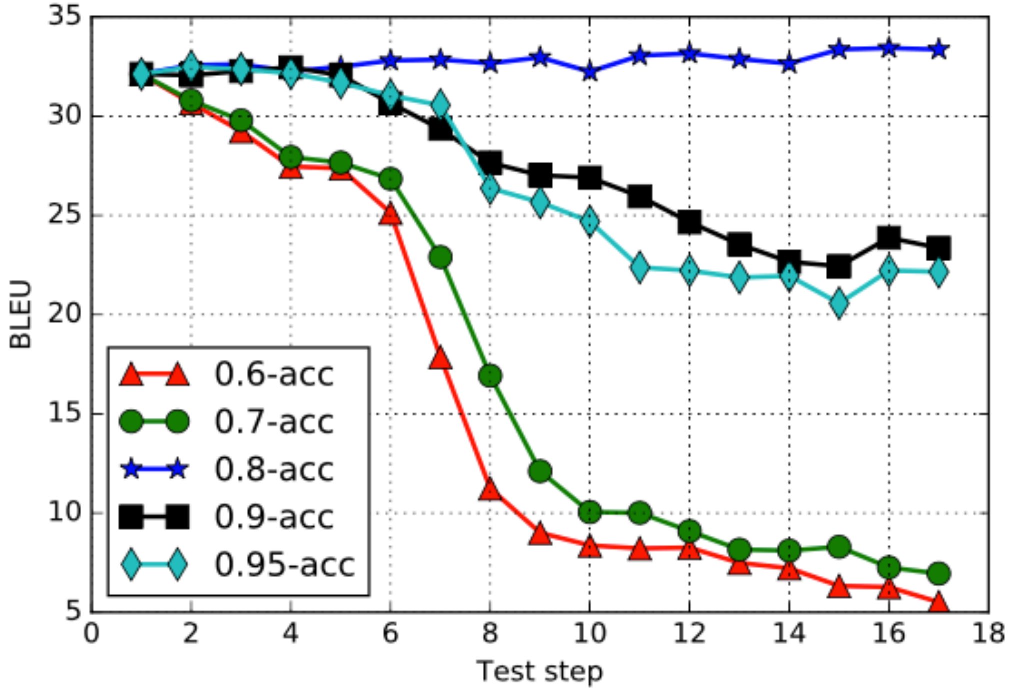
- Decide whether a particular generated output is true or not
- Classifier on sentences (e.g., Yu et al. 2017) or pairs of sentences (e.g. Wu et al. 2017)

GANs for Text are Hard! (Yang et al. 2017)

Type of Discriminator

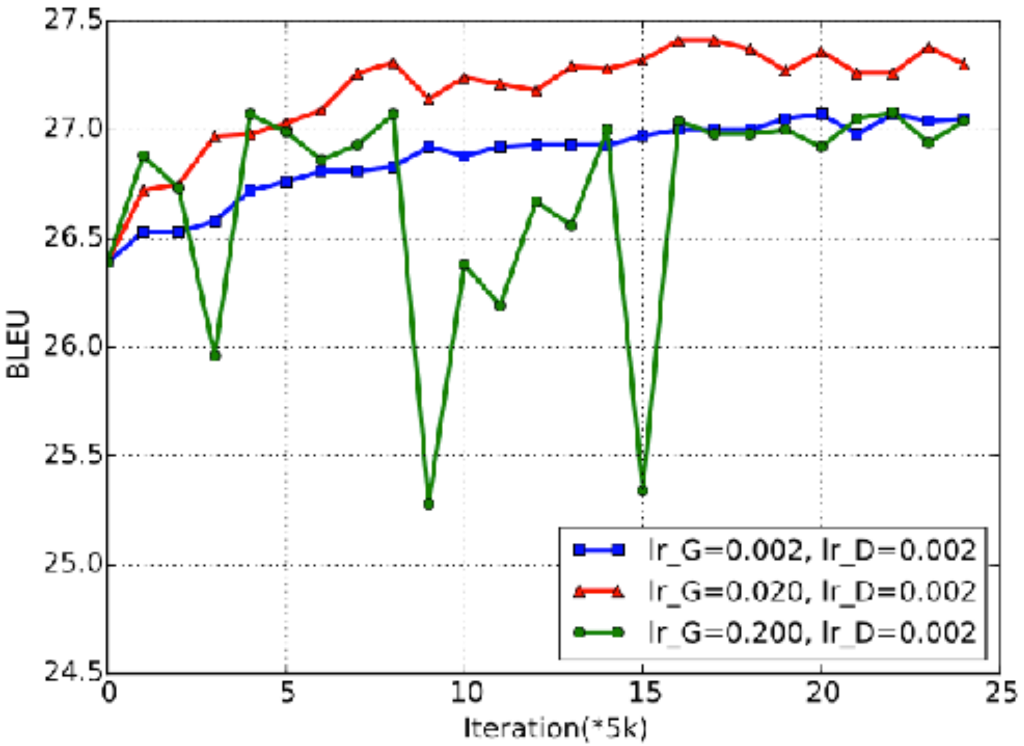


Strength of Discriminator

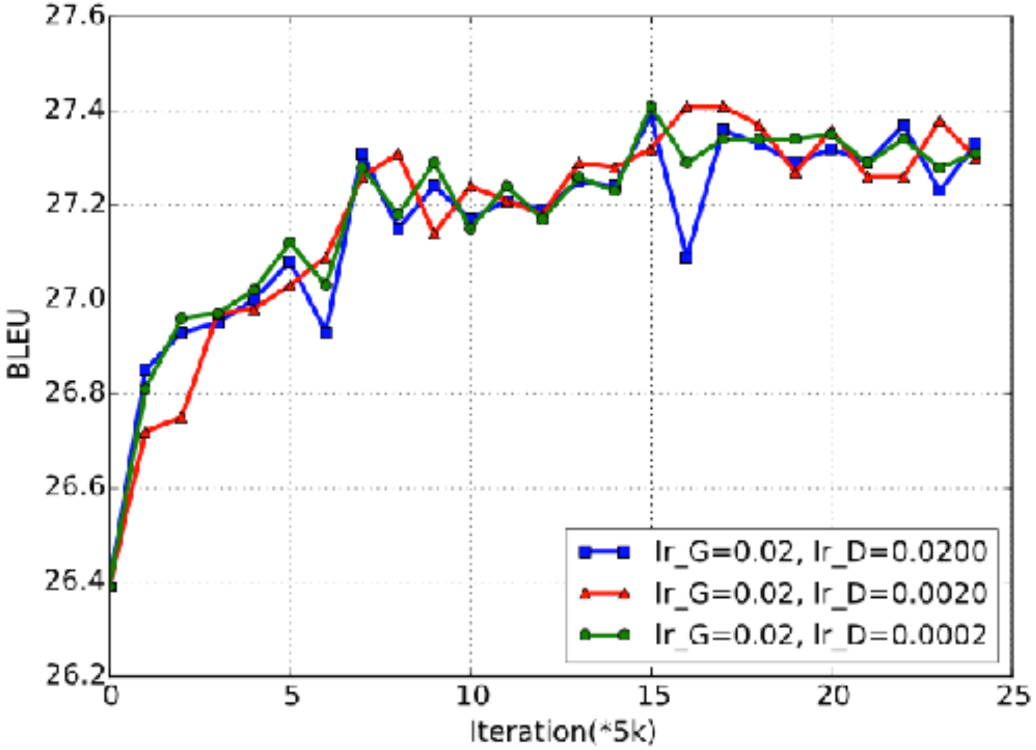


GANs for Text are Hard! (Yang et al. 2017)

Learning Rate for Generator



Learning Rate for Discriminator



Stabilization Trick: Assigning Reward to Specific Actions

- Getting a reward at the end of the sentence gives a credit assignment problem
- Solution: assign reward for partial sequences (Yu et al. 2016, Li et al. 2017)

D(this)

D(this is)

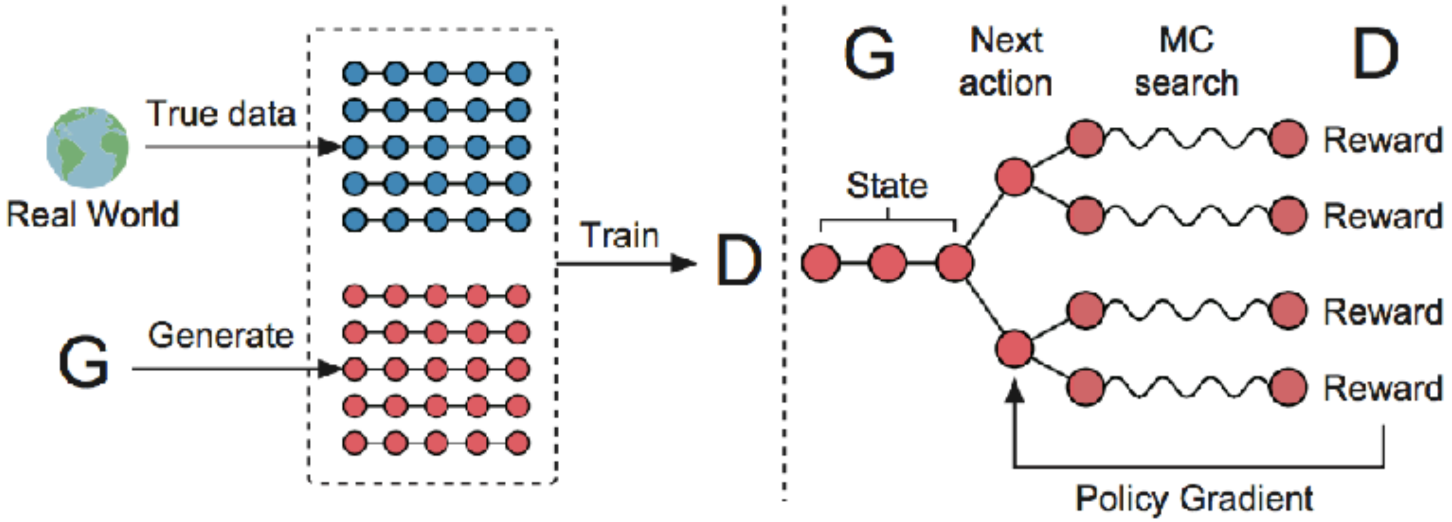
D(this is a)

D(this is a fake)

D(this is a fake sentence)

Stabilization Tricks: Performing Multiple Rollouts

- Like other methods using discrete samples, instability is a problem
- This can be helped somewhat by doing multiple rollouts (Yu et al. 2016)



Interesting Application: GAN for Data Cleaning (Yang et al. 2017)

- The discriminator tries to find “fake data”
- What about the real data it marks as fake? This might be noisy data!
- Selecting data in order of discriminator score does better than selecting data randomly.

Adversarial Feature Learning

- Adversaries over Features vs. Over Outputs

- Generative adversarial networks

$$\boldsymbol{x} \longrightarrow \boldsymbol{h} \longrightarrow \boldsymbol{y}$$

Adversary!

- Adversarial feature learning

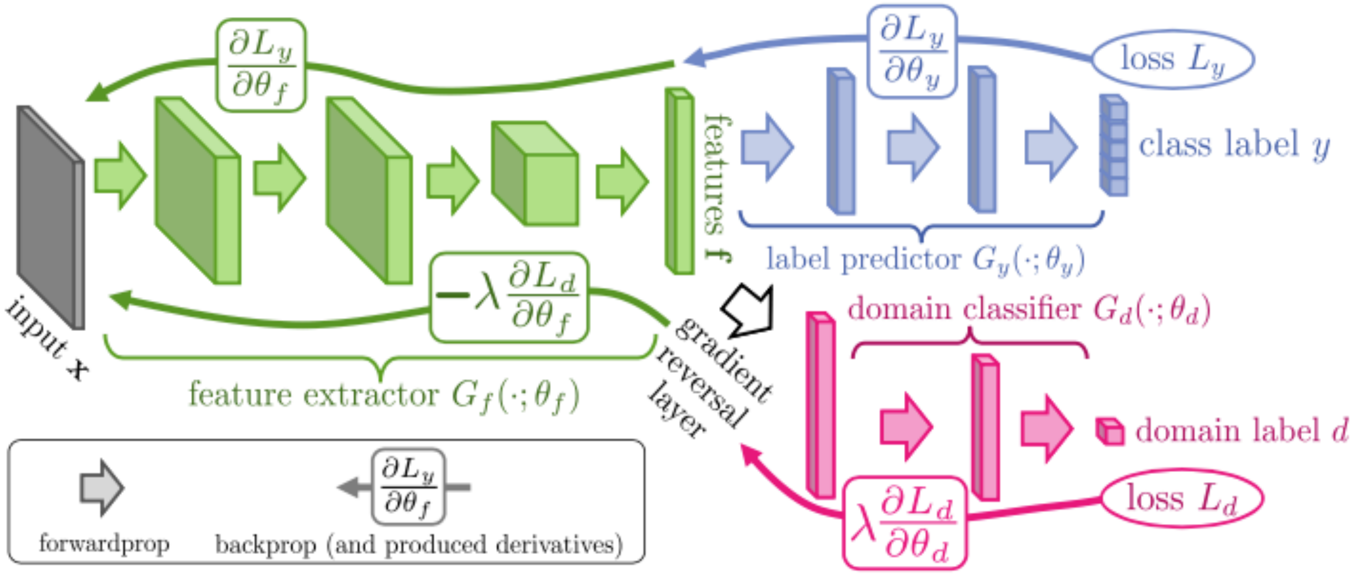
$$\boldsymbol{x} \longrightarrow \boldsymbol{h} \longrightarrow \boldsymbol{y}$$

Adversary!

- Why adversaries over features?
  - Non-generative tasks
  - Continuous features easier than discrete outputs

Learning Domain-invariant Representations (Ganin et al. 2016)

- Learn features that cannot be distinguished by domain



- Interesting application to synthetically generated or stale data (Kim et al. 2017)

Recap

Advanced Statistical Natural Language Processing

What is NLP?

Statistical machine learning (ML) methods

We'll cover only a subset of advanced, latest methods

Recap: Components of a ML solution (roughly)

- Loss
- Experience
- Optimization solver
- Model architecture

$$\min_{\theta} \mathcal{L}(\theta, \mathcal{E})$$

Optimization solver

Loss

Model architecture

Experience

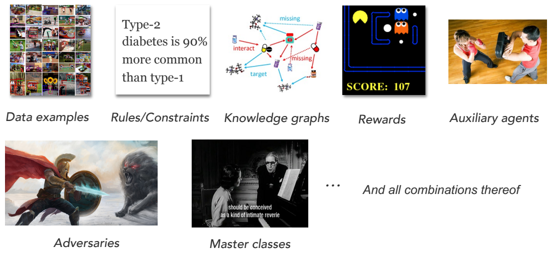
Machine learning solutions

(1) How can we make more efficient use of the data?

- Algorithms
  - **Supervised learning**: MLE, maximum entropy principle
  - **Unsupervised learning**: EM, variational inference, VAEs
  - **Self-supervised learning**: successful instances, e.g., BERT, GPT-3, contrastive learning, applications to downstream tasks
  - **Distant/weakly supervised learning**: successful instances
  - **Data augmentation**

Machine learning solutions

(2) Can we incorporate other types of experiences in learning?



- ○ Learning from rewards
  - ▪ Reinforcement learning: policy-based vs value-based, on-policy vs off-policy, extrinsic reward vs intrinsic reward, …
- ○ Learning from auxiliary models, e.g., adversarial models:
  - ▪ Generative adversarial learning (GANs and variants)

- Other ML topics not covered
  - ○ Meta learning
  - ○ Learning in dynamic environment
    - ▪ Online learning, lifelong/continual learning, …

NLP Tasks

- Language modeling
- Classification
- Sequence Labeling
- Parsing (structured prediction)
- Generation

Questions?