CS11-747 Neural Networks for NLP

Adversarial Methods

Graham Neubig

Carnegie Mellon University
Language Technologies Institute

Site
https://phontron.com/class/nn4nlp2017/

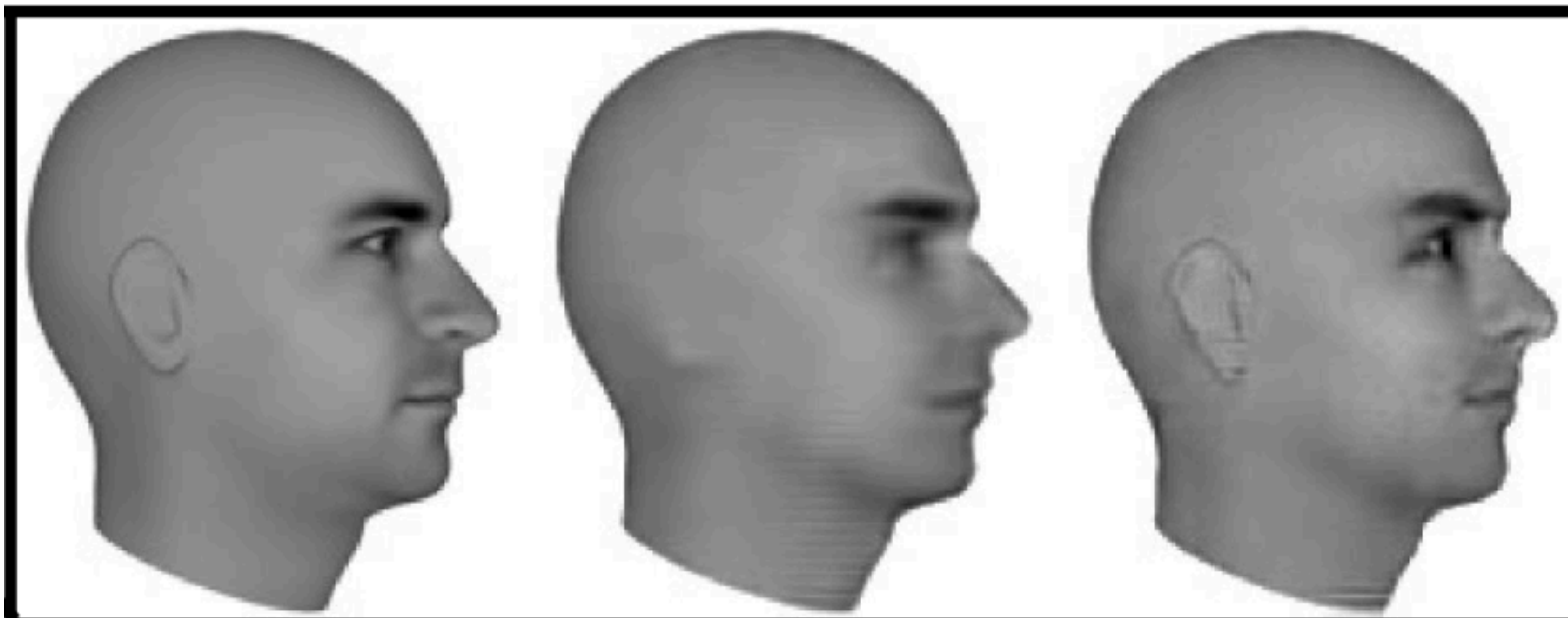
Generative Models

- Generate a sentence randomly from distribution $P(X)$
- Generate a sentence conditioned on some other information using distribution $P(X|Y)$

Problems with Generation

- Over-emphasis of common outputs, fuzziness

Real MLE Adversarial



- Note: this is probably a good idea if you are doing maximum likelihood!

Image Credit: Lotter et al. 2015

Adversarial Training

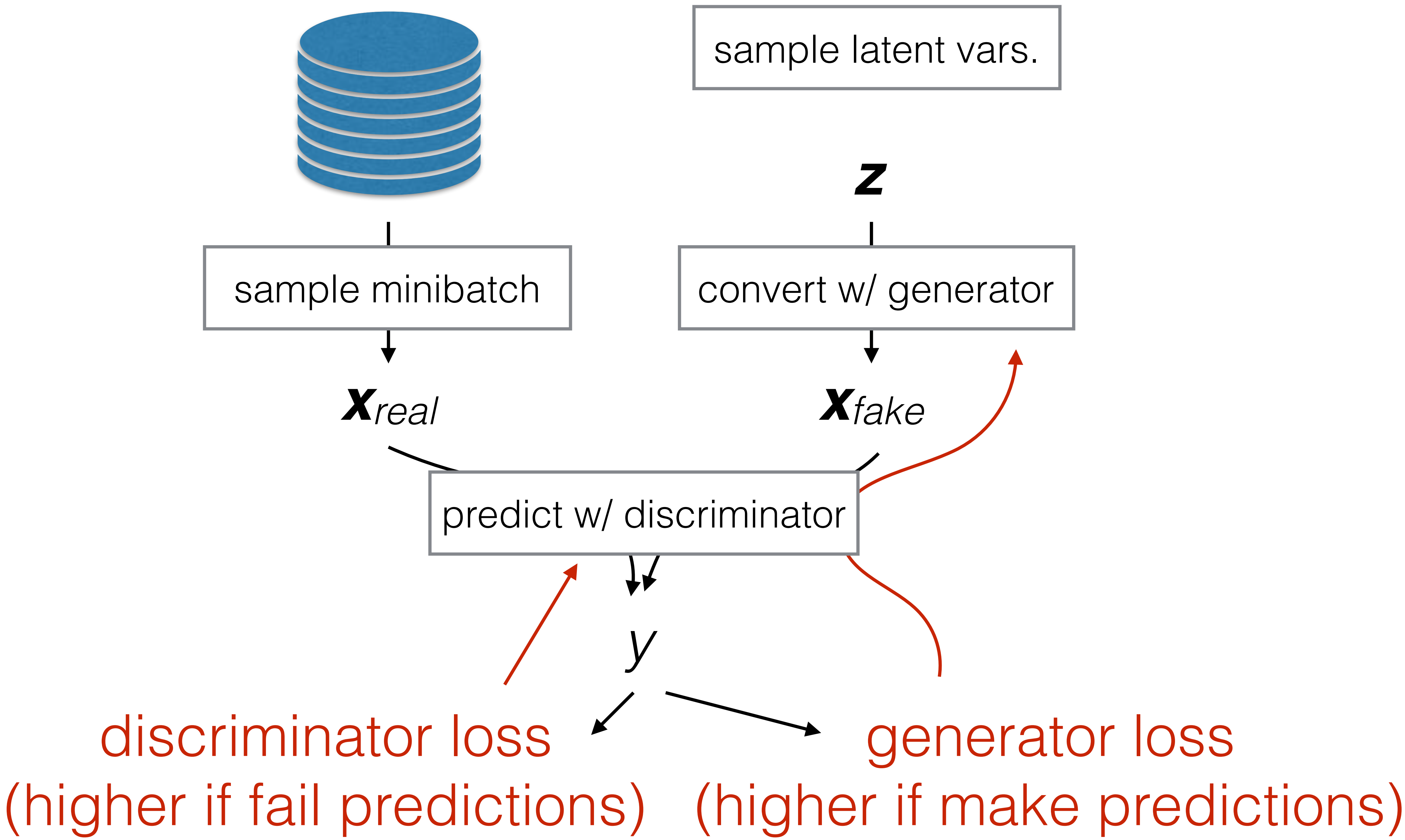
- Basic idea: create a “discriminator” that criticizes some aspect of the generated output
- **Generative adversarial networks:** criticize the generated output
- **Adversarial feature learning:** criticize the generated features to find some trait

Generative Adversarial Networks

Basic Paradigm

- Two players: generator and discriminator
  - **Discriminator:** given an image, try to tell whether it is real or not
  - **Generator:** try to generate an image that fools the discriminator into answering “real”

Training Method



In Equations

- Discriminator loss function:

$$\ell_D(\theta_D, \theta_G) = -\frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim P_{data}} \log D(\boldsymbol{x}) - \frac{1}{2}\mathbb{E}_{\boldsymbol{z}} \log(1 - D(G(\boldsymbol{z})))$$

High prob for real data High prob for fake data

- Generator loss function:
    - Zero sum loss:

$$\ell_G(\theta_D, \theta_G) = -\ell_D(\theta_D, \theta_G)$$

    - Heuristic non-saturating game loss:

$$\ell_G(\theta_D, \theta_G) = -\frac{1}{2}\mathbb{E}_{\boldsymbol{z}} \log D(G(\boldsymbol{z}))$$

    - Latter gives better gradients when discriminator accurate

Problems w/ Training: Mode Collapse

- GANs are great, but training is notoriously difficult
- e.g. mode collapse: generator learns to map all $\boldsymbol{z}$ to a single $\boldsymbol{x}$ in the training data
- One solution: use other examples in the minibatch as side information, making it easier to push similar examples apart (Salimans et al. 2016)

Problems w/ Training: Over-confident Discriminator

- At the beginning of training it is easy to learn the discriminator, causing it to be over-confident

- One way to fix this: label smoothing to reduce the confidence of the target

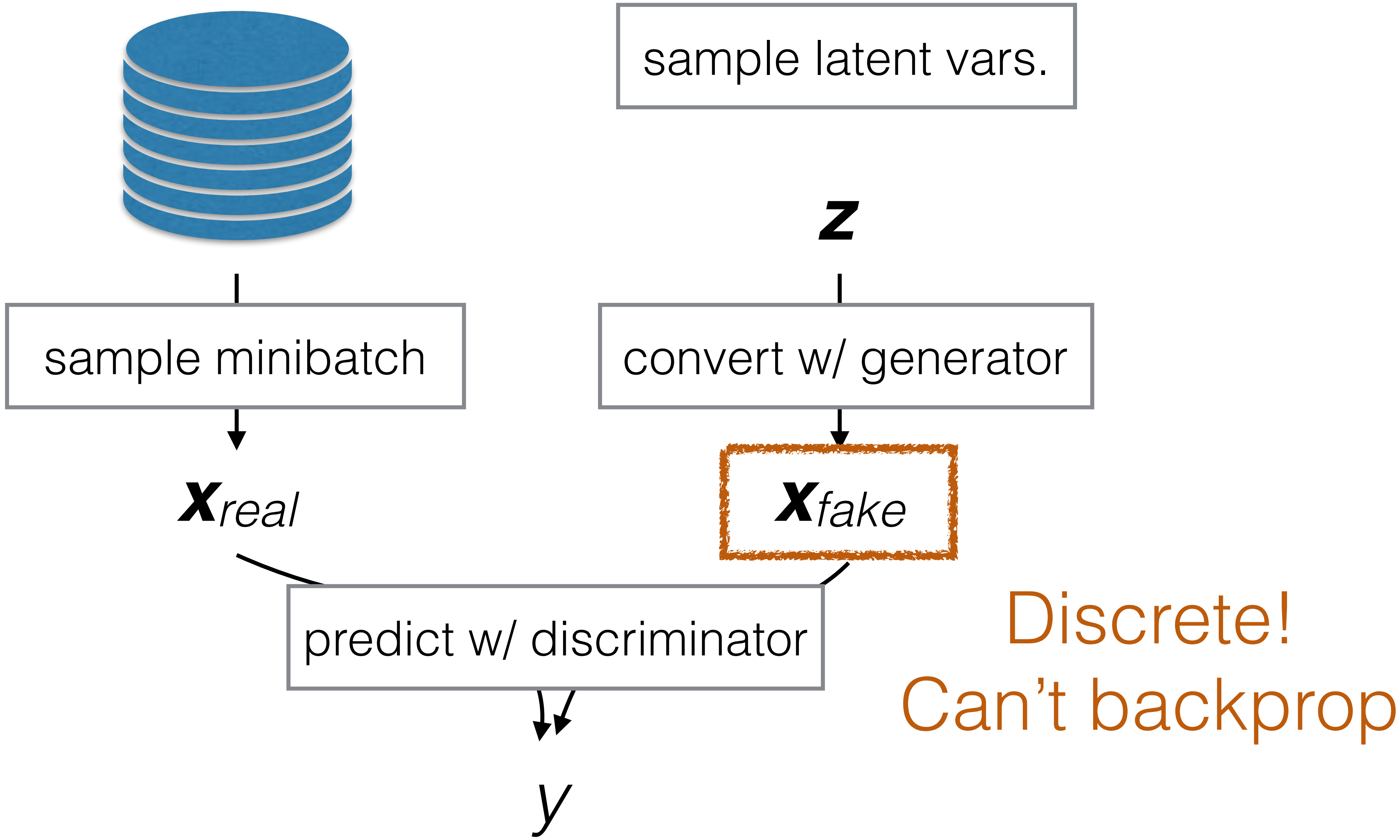
- Salimans et al. (2016) suggest one-sided label smoothing, which only smooths predictions over

Applying GANs to 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

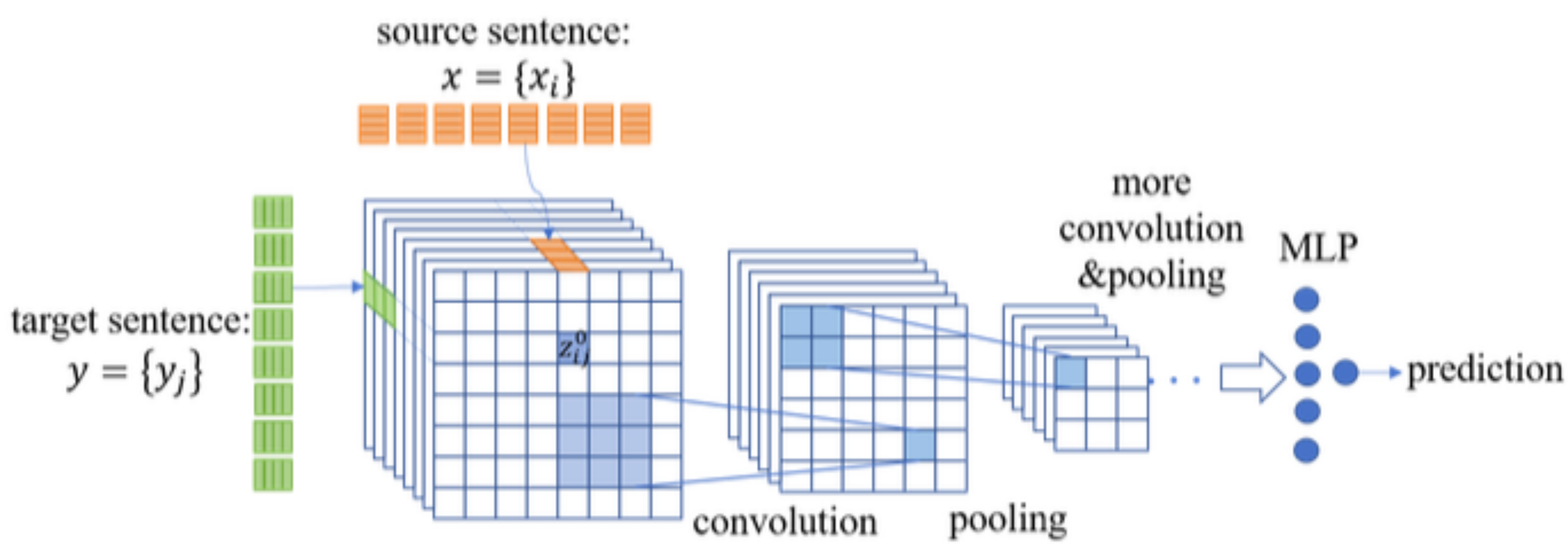


Solution: Use Learning Methods for Latent Variables

- 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

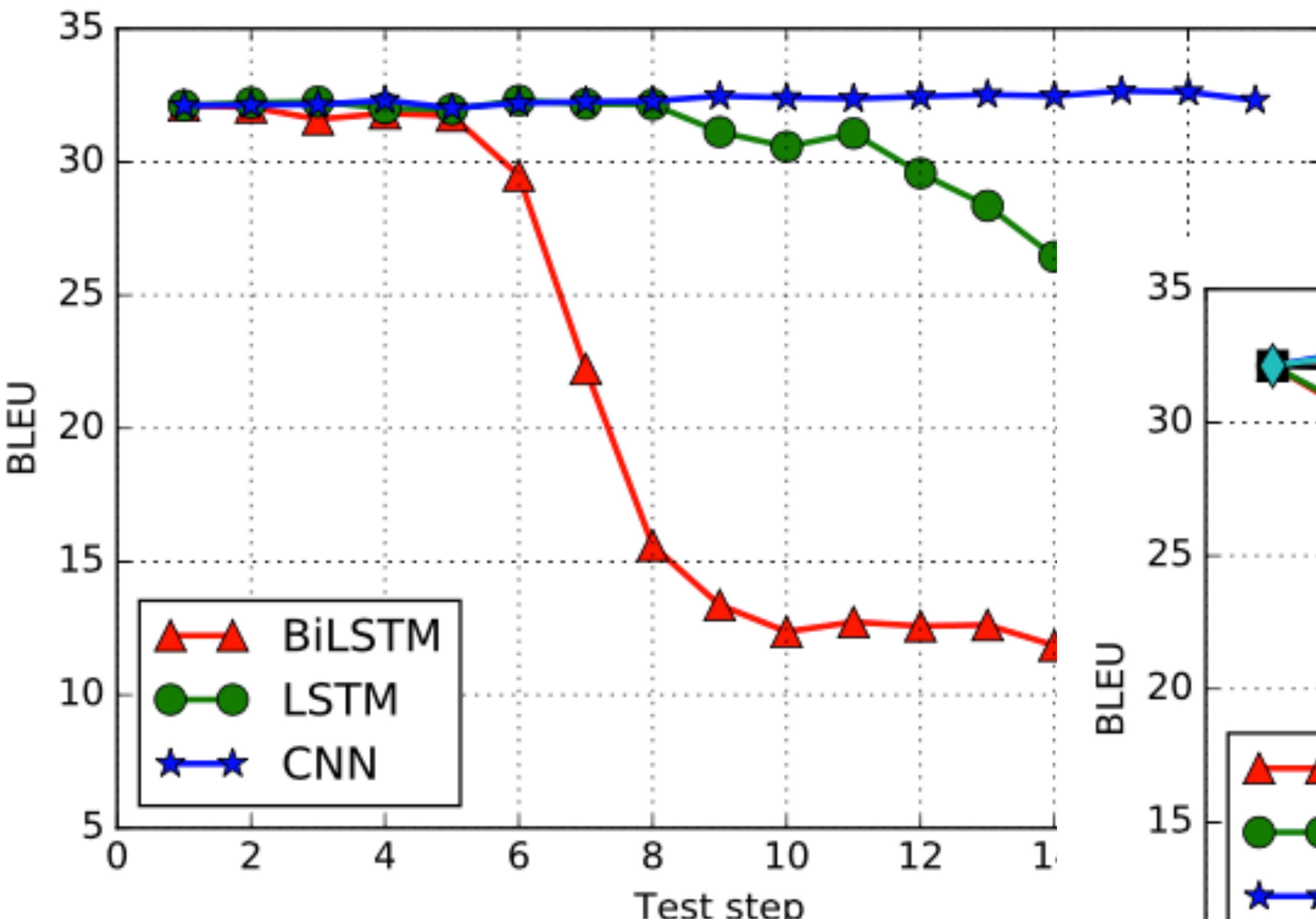
- Decide whether a particular generated output is true or not
- Commonly use CNNs as discriminators, either on sentences (e.g. Yu et al. 2017), or pairs of sentences (e.g. Wu et al. 2017)



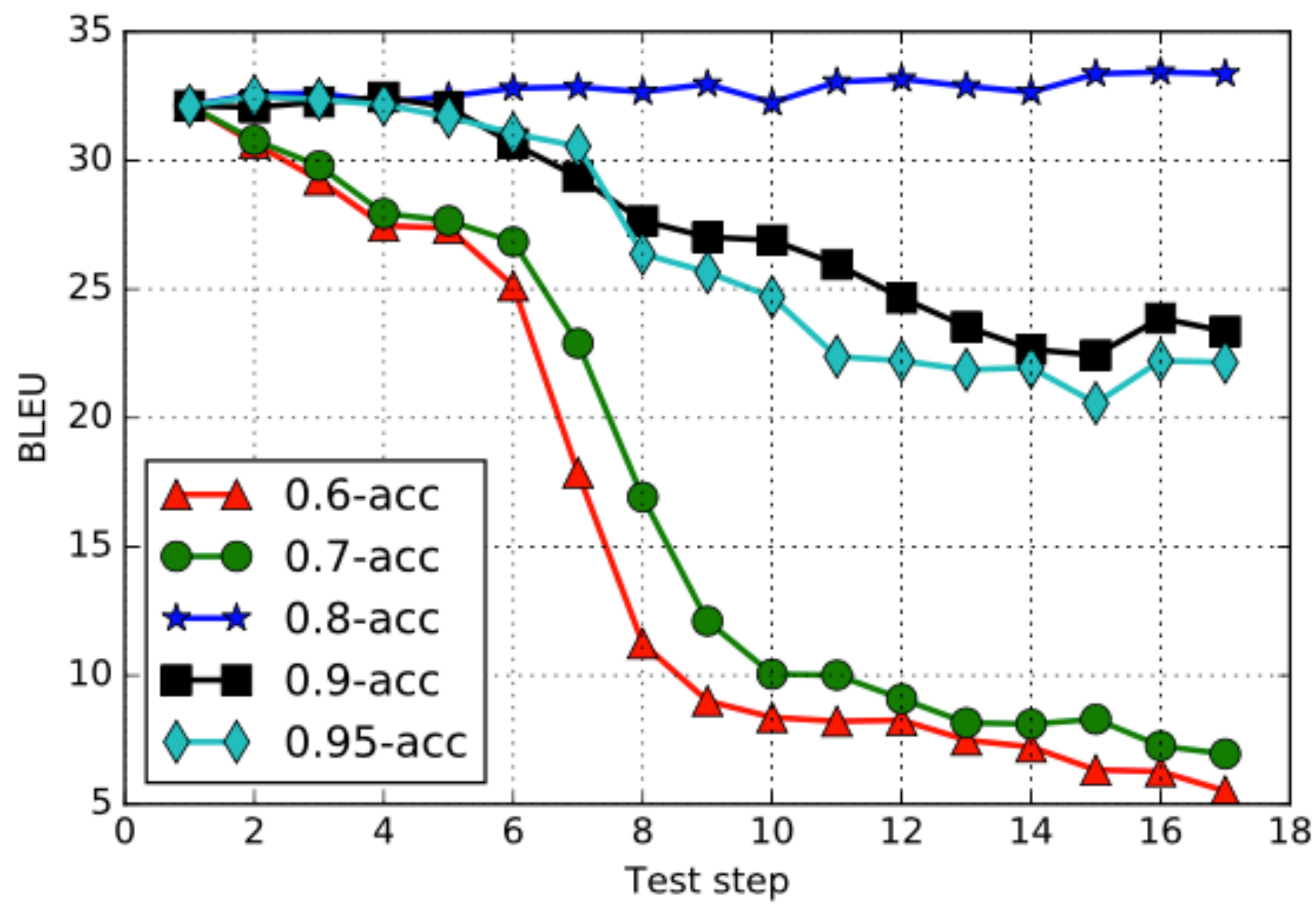
GANs for Text are Hard!

(Yang et al. 2017)

Type of Discriminator



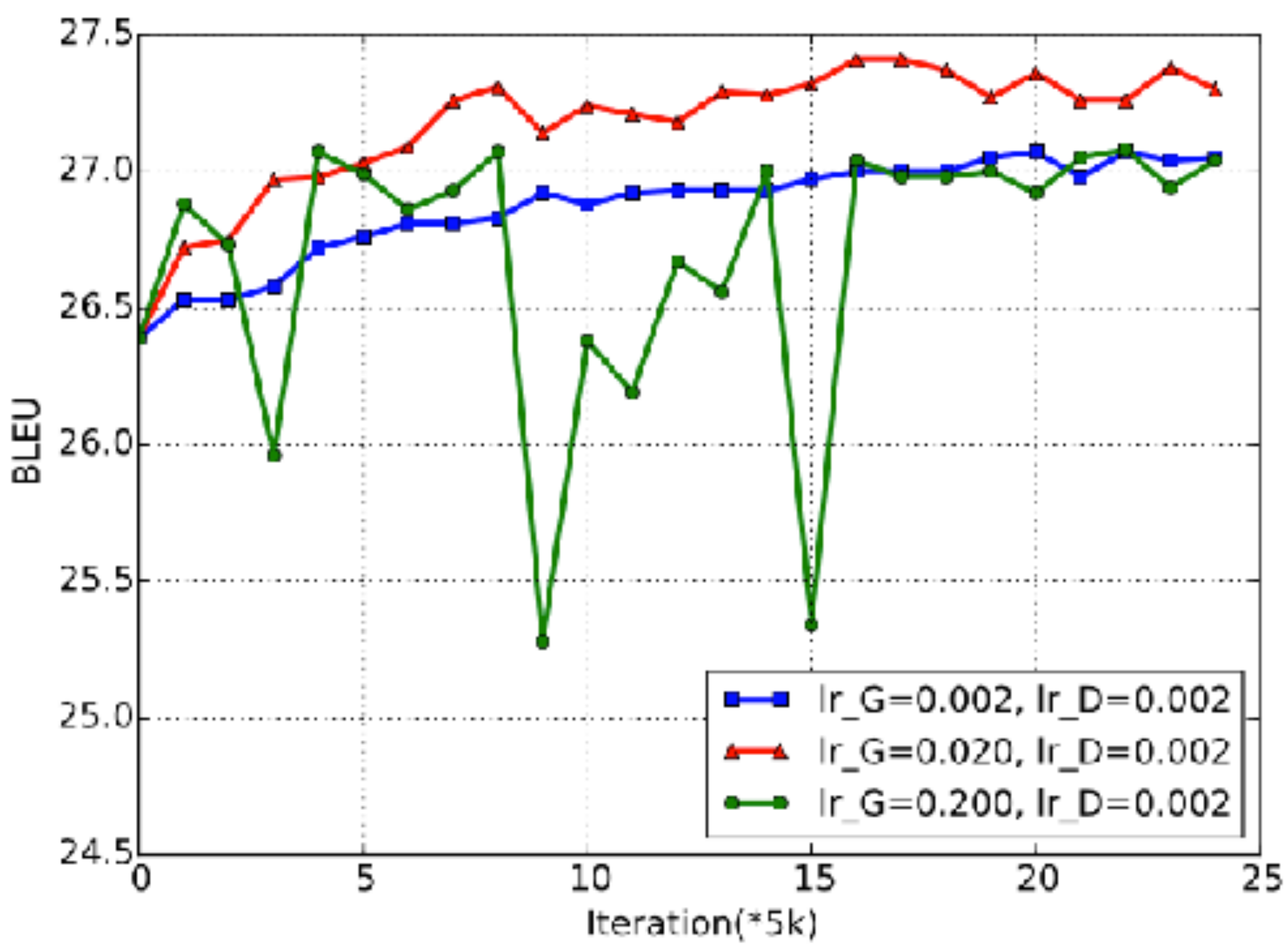
Strength of Discriminator



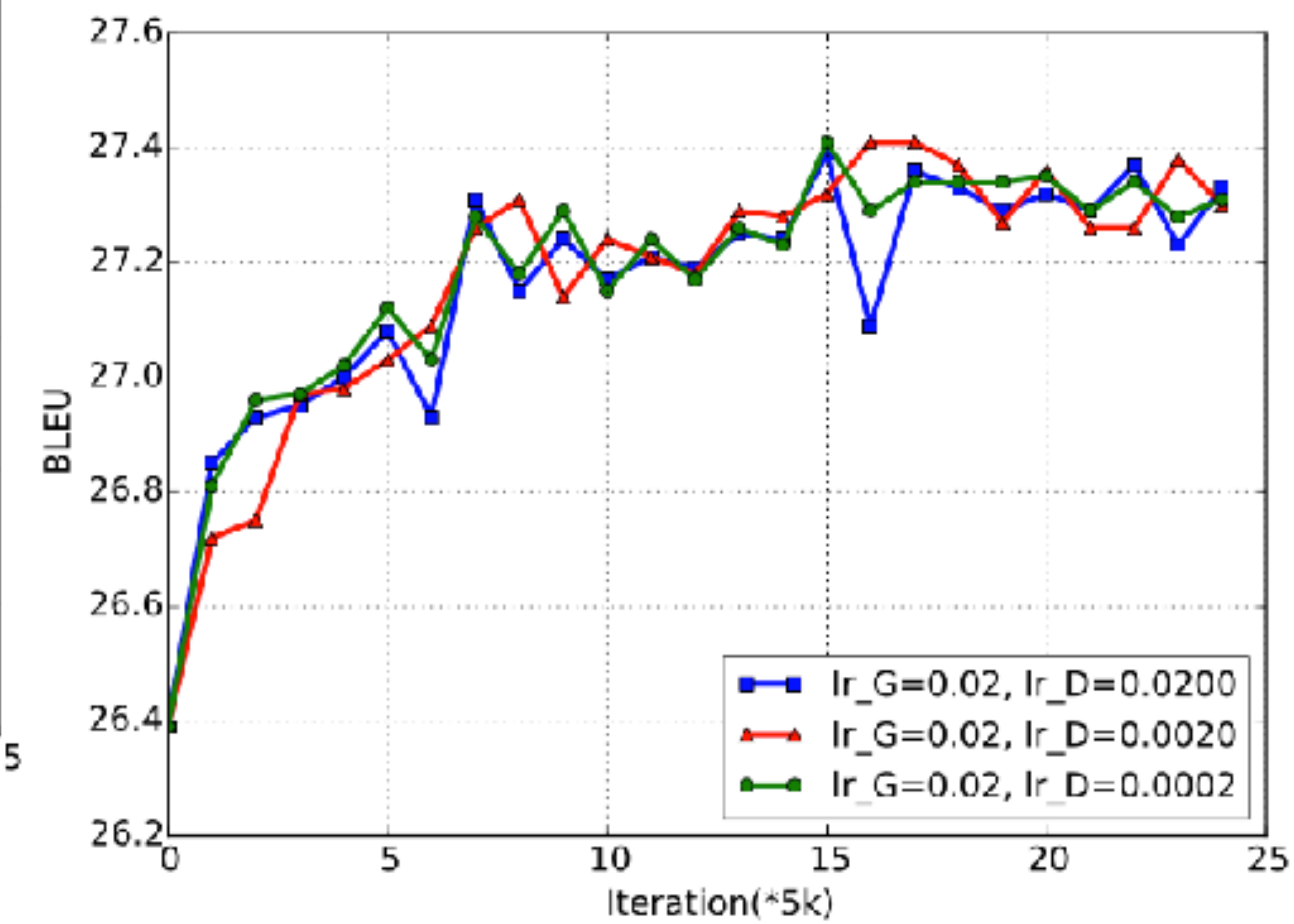
GANs for Text are Hard!

(Wu 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)

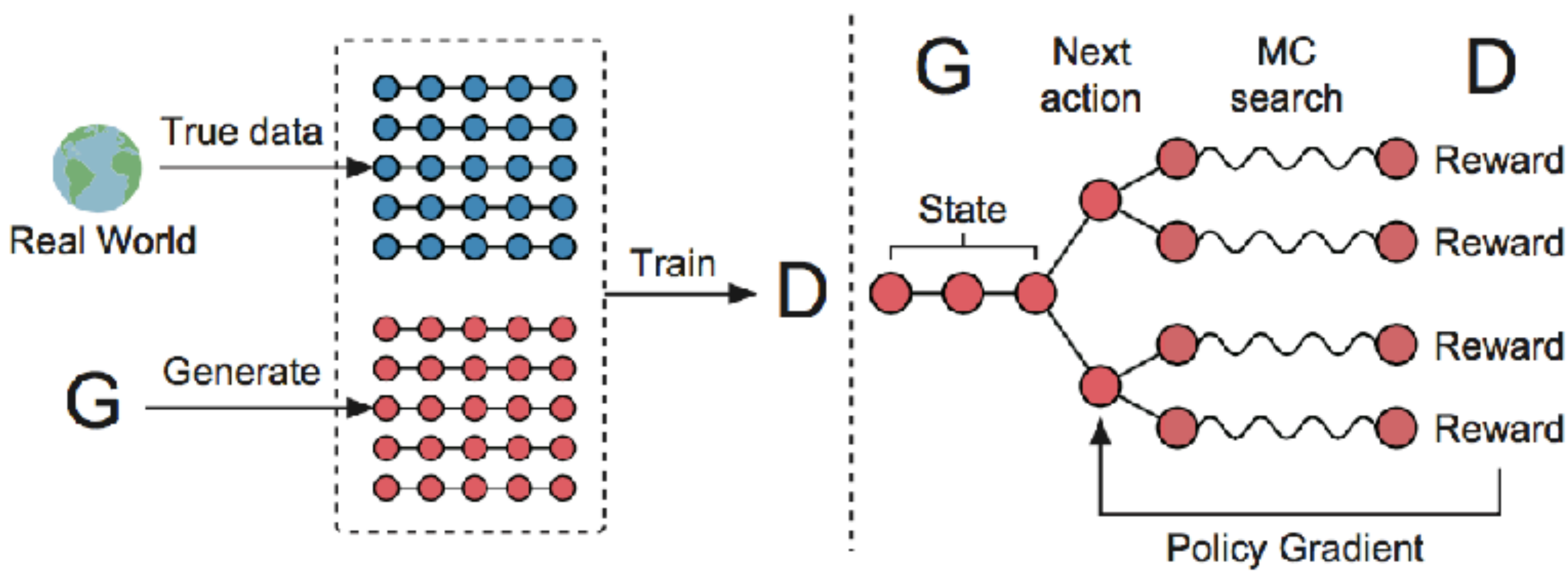
D(this is a fake)

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

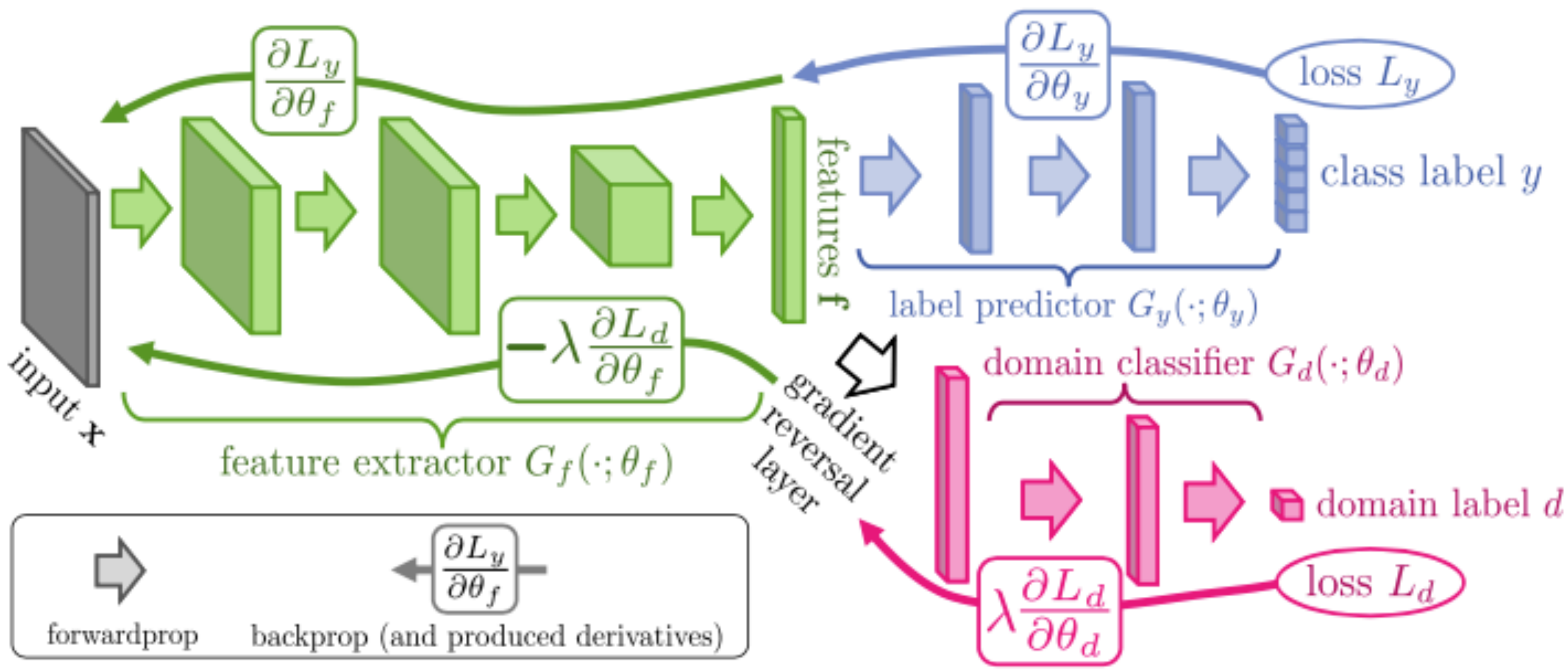
$$\boldsymbol{x} \longrightarrow \boldsymbol{h} \longrightarrow \boldsymbol{y}$$

Adversary!

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

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

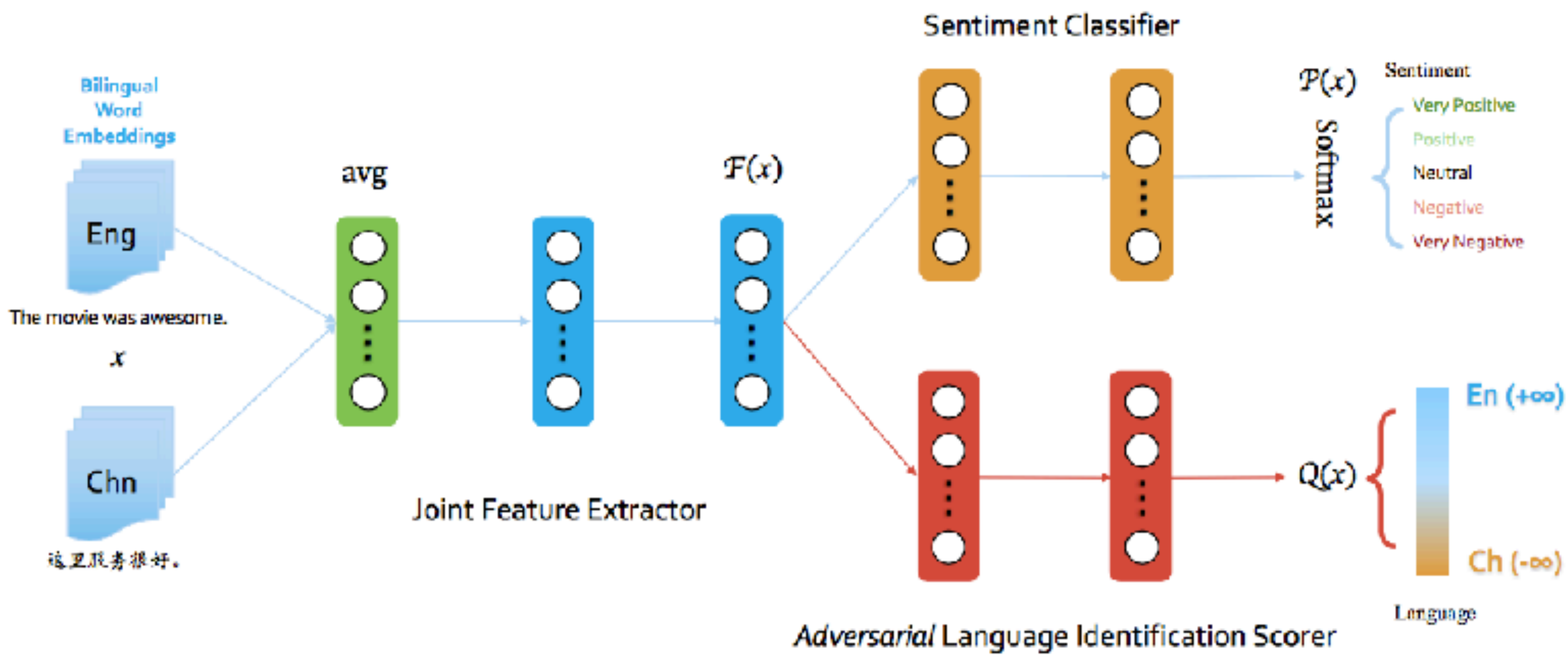
- Learn features that cannot be distinguished by domain



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

Learning Language-invariant Representations

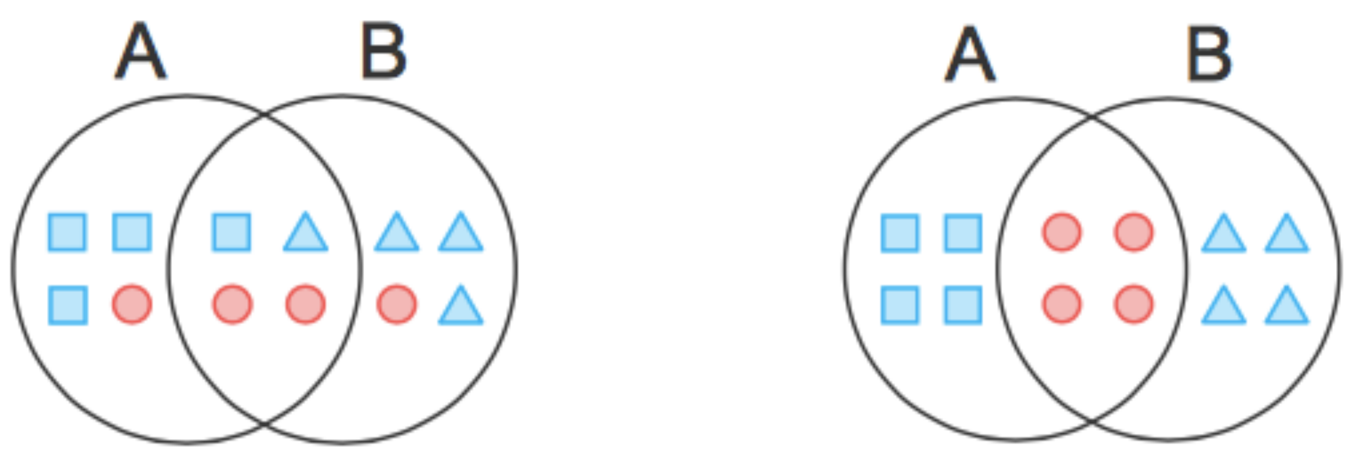
- Chen et al. (2016) learn language-invariant representations for text classification



- Also on multi-lingual machine translation (Xie et al. 2017)

Adversarial Multi-task Learning (Liu et al. 2017)

- Basic idea: want some features in a shared space across tasks, others separate



(a) Shared-Private Model (b) Adversarial Shared-Private Model

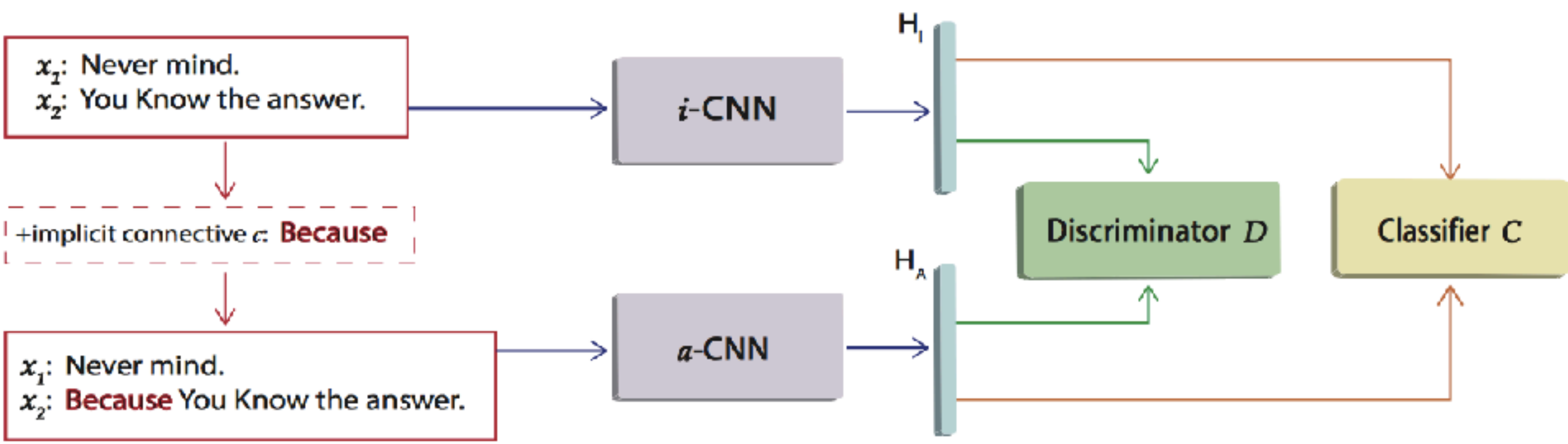
- Method: adversarial discriminator on shared features, orthogonality constraints on separate features

Implicit Discourse Connection Classification w/ Adversarial Objective

(Qin et al. 2017)

- Idea: implicit discourse relations are not explicitly marked, but would like to detect them if they are

- Text with explicit discourse connectives should be the same as text without!



Professor Forcing
(Lamb et al. 2016)

- Halfway in between a discriminator on discrete outputs and feature learning
  - Generate output sequence according to model
  - But train discriminator on hidden states

(sampled or true output sequence)

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

Adversary!

Unsupervised Style Transfer for Text (Shen et al. 2017)

- Two potential styles (e.g. positive and negative sentences)

- Use professor forcing to discriminate between true style 1, fake style 2->1, and another for vice-versa

Sentiment transfer from negative to positive
I would recommend find another place. I would recommend this place again!
Do not like it at all! All in all, it's great!
I regret not having the time to shop around. I have a great experience here.
Average Mexican food. Authentic Italian food.

Questions?