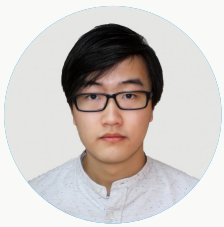
BOREALIS AI UNIVERSITY OF TORONTO

Adversarial Contrastive Estimation

ACL 2018

*AVISHEK (JOEY) BOSE, *HUAN LING, *YANSHUAI CAO


Contrastive Estimation

Many Machine Learning models learn by trying to separate positive examples from negative examples.

- Positive Examples are taken from observed real data distribution (training set)
- Negative Examples are any other configurations that are not observed
- Data is in the form of tuples or triplets (x^+, y^+) and (x^+, y^-) are positive and negative data points respectively.

Easy Negative Examples with NCE

Noise Constrastive Estimation samples negatives by taking $p(y^-|x^+)$ to be some unconditional $p_{nce}(y)$. **What's wrong with this?**

- Negative y^- in (x, y^-) is not tailored toward x
- Difficult to choose hard negatives as training progresses
- Model doesn't learn discriminating features between positive and hard negative examples

NCE negatives are easy !!!

Hard Negative Examples

Informal Definition: Hard negative examples are data points that are extremely difficult for the training model to distinguish from positive examples.

- Hard Negatives result to higher losses and thus more more informative gradients
- Not necessarily closest to a positive datapoint in embedding space

Technical Contributions

- Adversarial Contrastive Estimation: A general technique for hard negative mining using a Conditional GAN like setup.
- A novel entropy regularizer that prevents generator mode collapse and has good empirical benefits
- A strategy for handling false negative examples that allows training to progress
- Empirical validation across 3 different embedding tasks with state of the art results on some metrics

Adversarial Contrastive Estimation

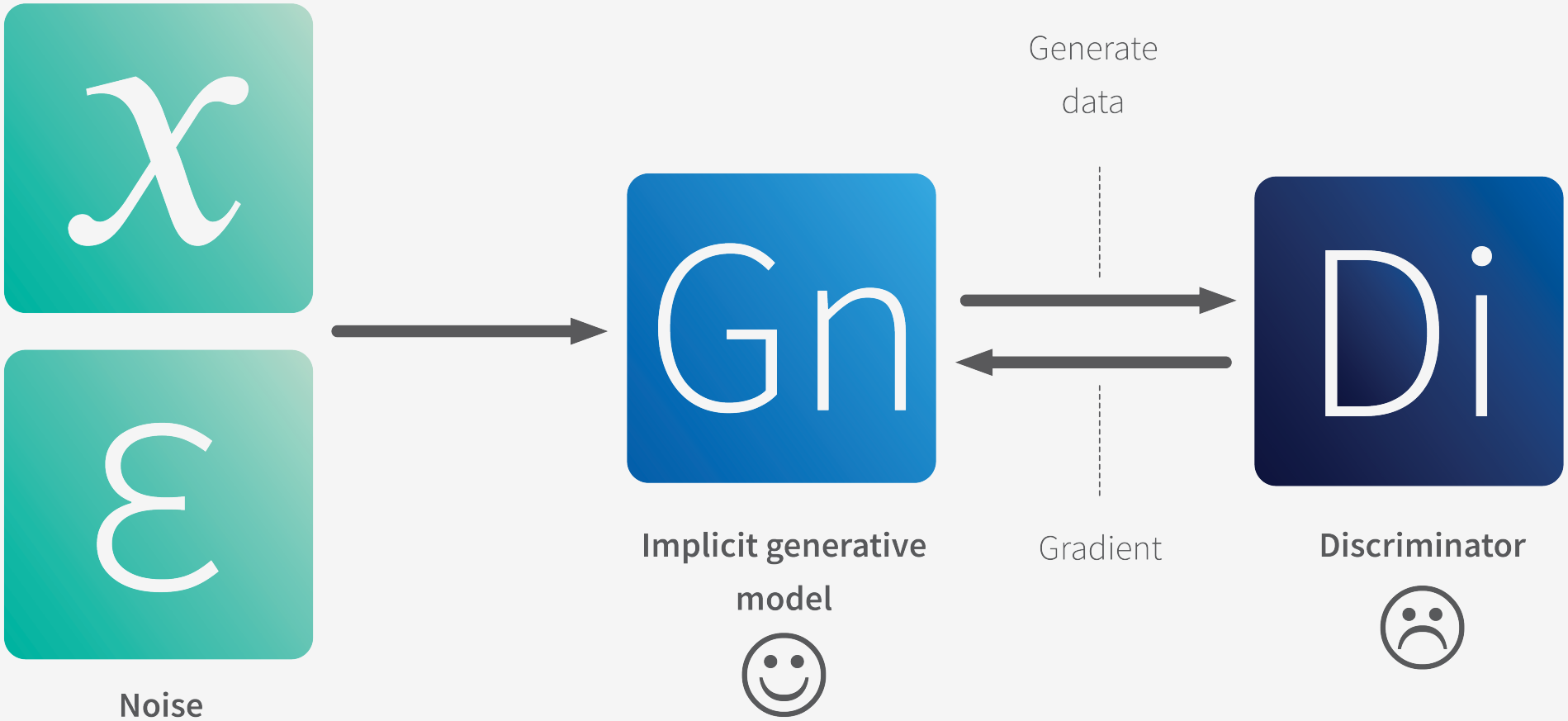
Problem:

We want to **generate** negatives that ... **"fool"** a **discriminative** model into misclassifying.

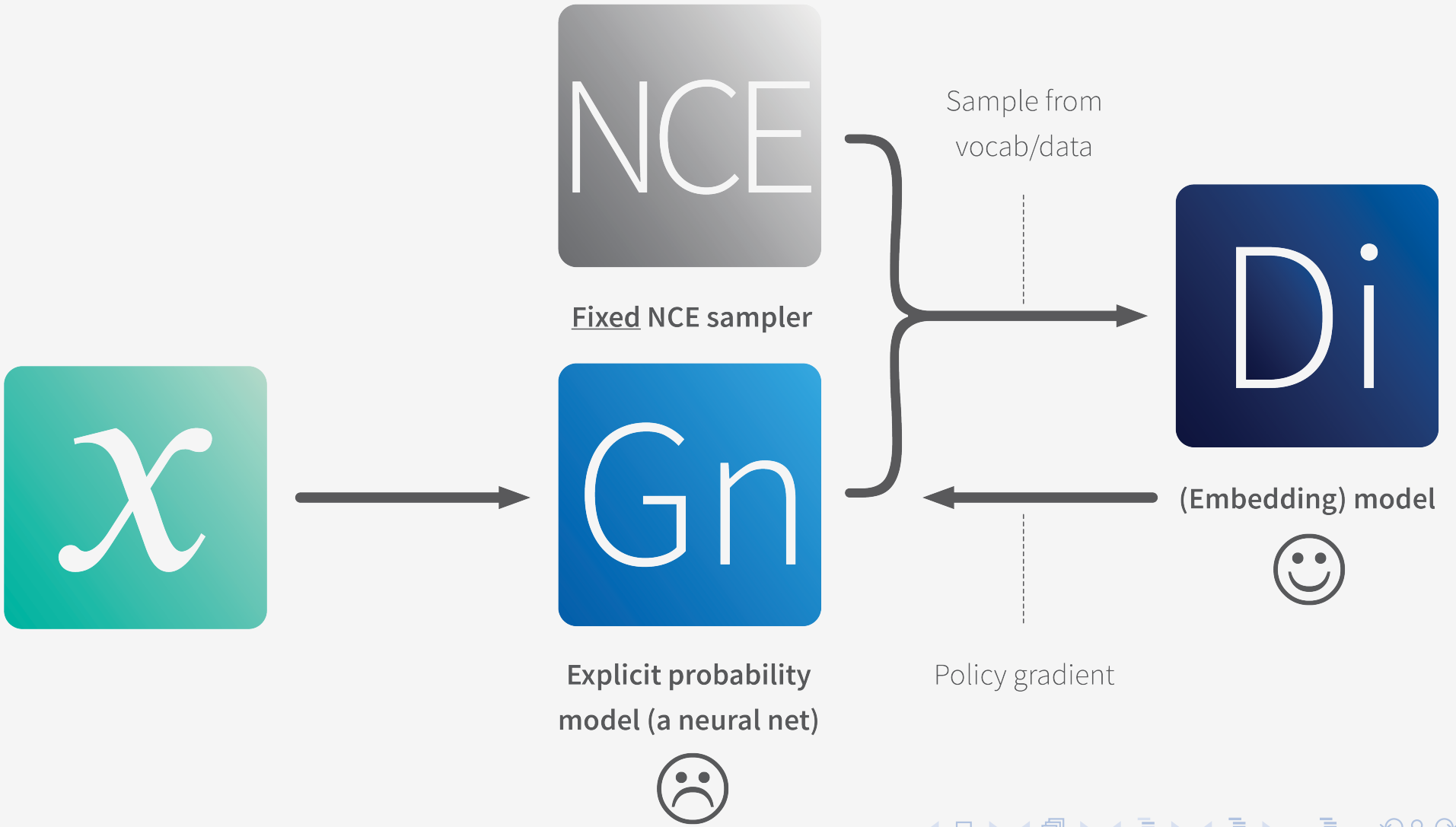
Solution:

Use a Conditional GAN to sample hard negatives given x^+. We can augment NCE with an adversarial sampler, $\lambda p_{nce}(y) + (1 - \lambda) g_\theta(y|x)$.

Conditional GAN



Adversarial Contrastive Estimation



The ACE Generator

- The ACE generator defines a categorical distribution over all possible y^- values
- Picking a negative example is a discrete choice and not differentiable
- Simplest way to train via Policy Gradients is the REINFORCE gradient estimator
- Learning is done via a GAN style min-max game

$$\min_{\omega} \max_{\theta} V(\omega, \theta) = \min_{\omega} \max_{\theta} \mathbb{E}_{p^+(x)} L(\omega, \theta; x) \tag{1}$$

Technical Contributions for effective training

Problem:

GAN training can suffer from mode collapse? What happens if the generator collapses on its favorite few negative examples?

Solution:

Add a entropy regularizer term to the generators loss:

$$R_{ent}(x) = \max(0, c - H(g_\theta(y|x))) \tag{2}$$

- $H(g_\theta(y|x))$ is the entropy of the categorical distribution
- $c = \log(k)$ is the entropy of a uniform distribution over k choices

Technical Contributions for effective training

Problem:

The Generator can sample false negatives $\rightarrow$ gradient cancellation

Solution:

Apply an additional two-step technique, whenever computationally feasible.

1. **Maintain an in memory hash map of the training data and Discriminator filters out false negatives**
2. Generator receives a penalty for producing the false negative
3. Entropy Regularizer spreads out the probability mass

Technical Contributions for effective training

Problem:

The Generator can sample false negatives $\rightarrow$ gradient cancellation

Solution:

Apply an additional two-step technique, whenever computationally feasible.

1. Maintain an in memory hash map of the training data and Discriminator filters out false negatives
2. **Generator receives a penalty for producing the false negative**
3. Entropy Regularizer spreads out the probability mass

Technical Contributions for effective training

Problem:

The Generator can sample false negatives $\rightarrow$ gradient cancellation

Solution:

Apply an additional two-step technique, whenever computationally feasible.

1. Maintain an in memory hash map of the training data and Discriminator filters out false negatives
2. Generator receives a penalty for producing the false negative
3. **Entropy Regularizer spreads out the probability mass**

Technical Contributions for effective training

Problem:

REINFORCE is known to have extremely high variance.

Solution:

Reduce Variance using the self-critical baseline. Other baselines and gradient estimators are also good options.

Technical Contributions for effective training

Problem:

The generator is not learning from the NCE samples.

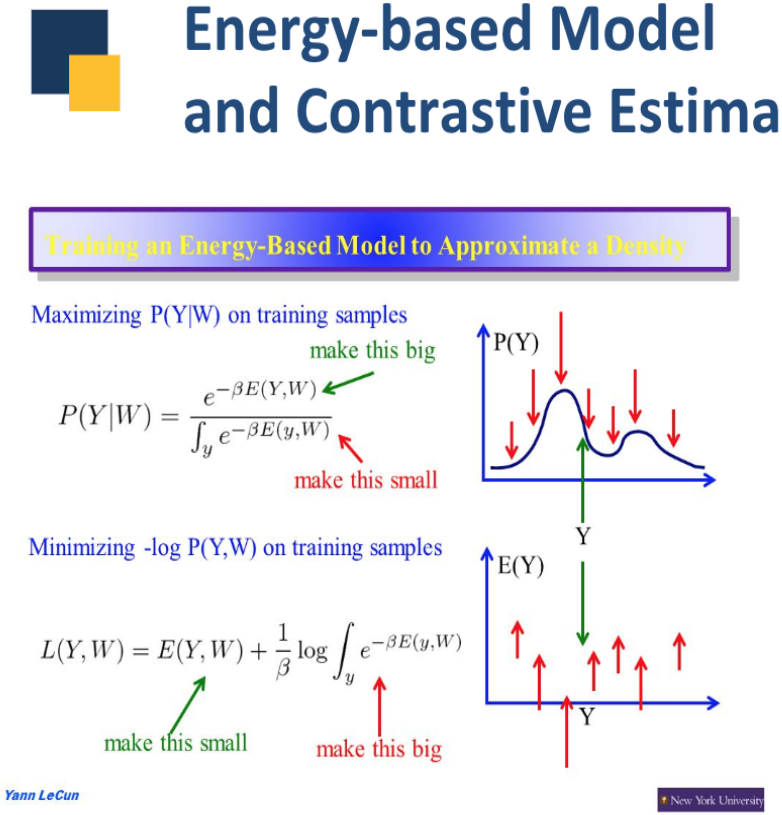
Solution:

Use Importance Sampling. Generator can leverage NCE samples for exploration in an off-policy scheme. The modified reward now looks like

$$g_\theta(y^-|x)/p_{nce}(y^-)$$

.

Related Work



LeCun et al.
2005, 2006

Method	Example models	Where to push up?
(exact) Maximum Likelihood (MLE)	Exact softmax	Everywhere
Contrastive Divergence (CD) (Hinton, 2002) (Carreira-Perpiñán and Hinton, 2005)	RBM (Hinton, 2002), (Carreira-Perpiñán and Hinton, 2005)	Neighborhood around observation (defined by 1 or more step(s) Gibbs sampling)
Noise Contrastive Estimation (NCE) (Dyer, 2014), (Mnih and Teh, 2012), (Vaswani et al, 2013), (Mnih and Kavukcuoglu, 2013)	Skip-gram or cbow word2vec, transD, Order Embeddings (Gutmann, and Hyavarinen,2012), (Mikolov et al, 2012), (Ji et al 2015), (Vendrov et al 2016)	Random places
Adversarial Contrastive Estimation (ACE)	This work	Adversarially learned hard negative locations

Contemporary Work

GANs for NLP that are close to our work

- MaskGAN Fedus et. al 2018
- Incorporating GAN for Negative Sampling in Knowledge Representation Learning Wang et. al 2018
- KBGAN Cai and Wang 2017

Example: Knowledge Graph Embeddings

Data in the form of triplets *(head entity, relation, tail entity)*. For example {United states of America, partially contained by ocean, Pacific}

Basic Idea: The embeddings for h, r, t should roughly satisfy $h + r \approx t$

Link Prediction:
Goal is to learn from observed positive entity relations and predict missing links.

ACE for Knowledge Graph Embeddings

Positive Triplet: $\xi^+ = (h^+, r^+, t^+)$

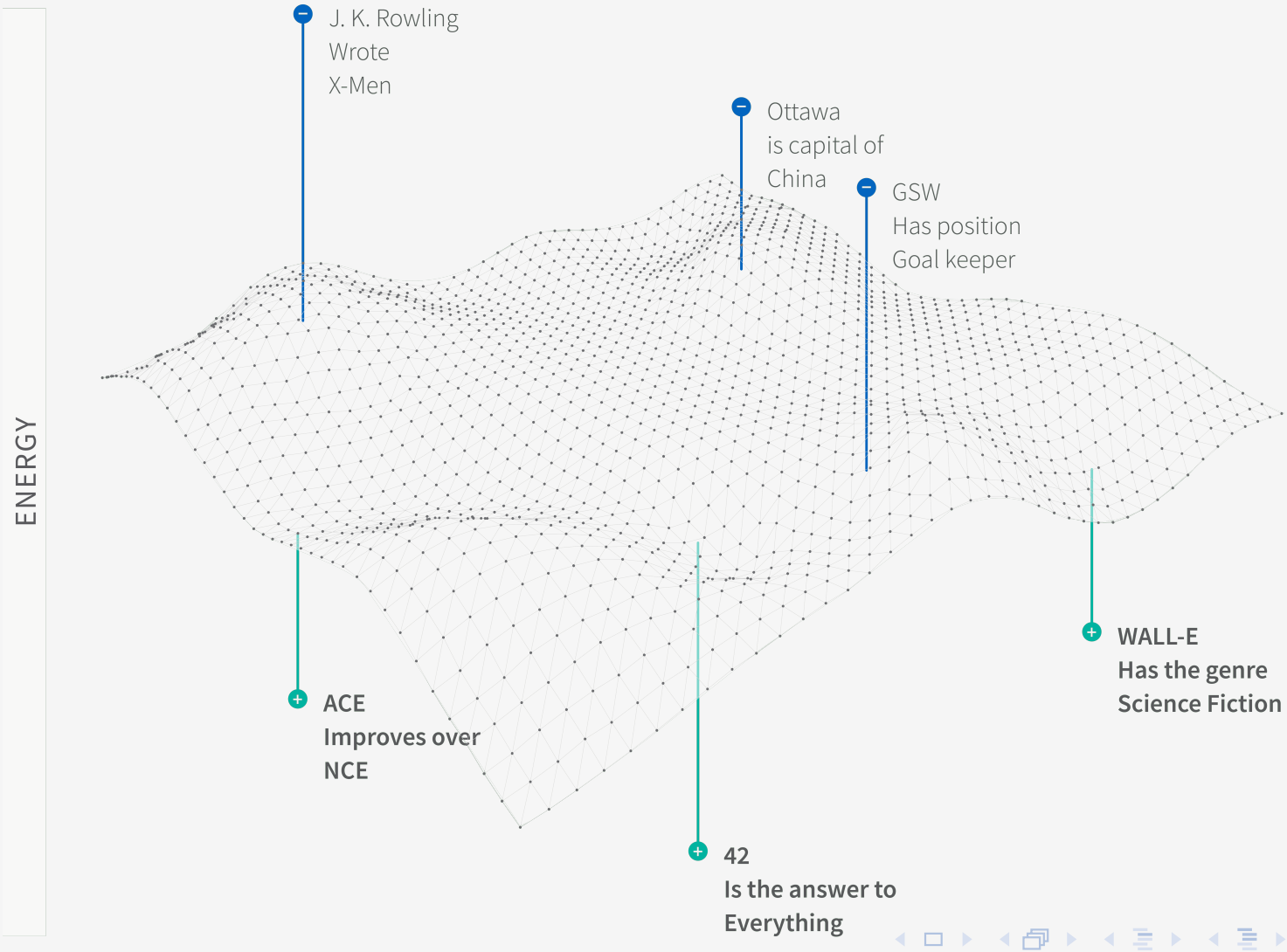
Negative Triplet: Either negative head or tail is sampled i.e. $\xi^- = (h^-, r^+, t^+)$ or $\xi^- = (h^+, r^+, t^-)$

Loss Function:

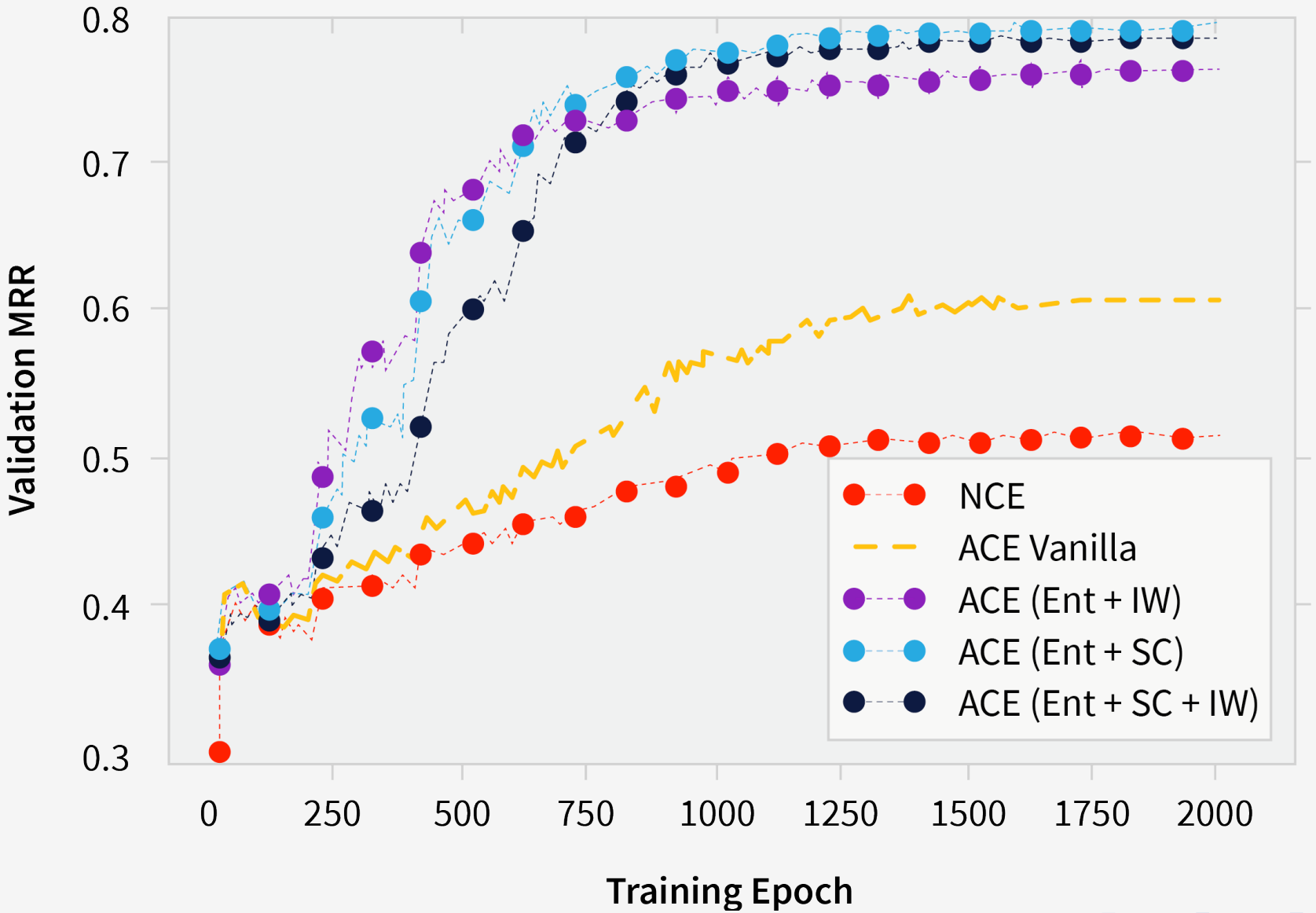
$$L = \max(0, \eta + s_\omega(\xi^+) - s_\omega(\xi^-)) \tag{3}$$

ACE Generator: $g_\theta(t^-|r^+, h^+)$ or $g_\theta(h^-|r^+, t^+)$ parametrized by a feed forward neural net.

ACE for Knowledge Graph Embeddings



Experimental Result: Ablation Study

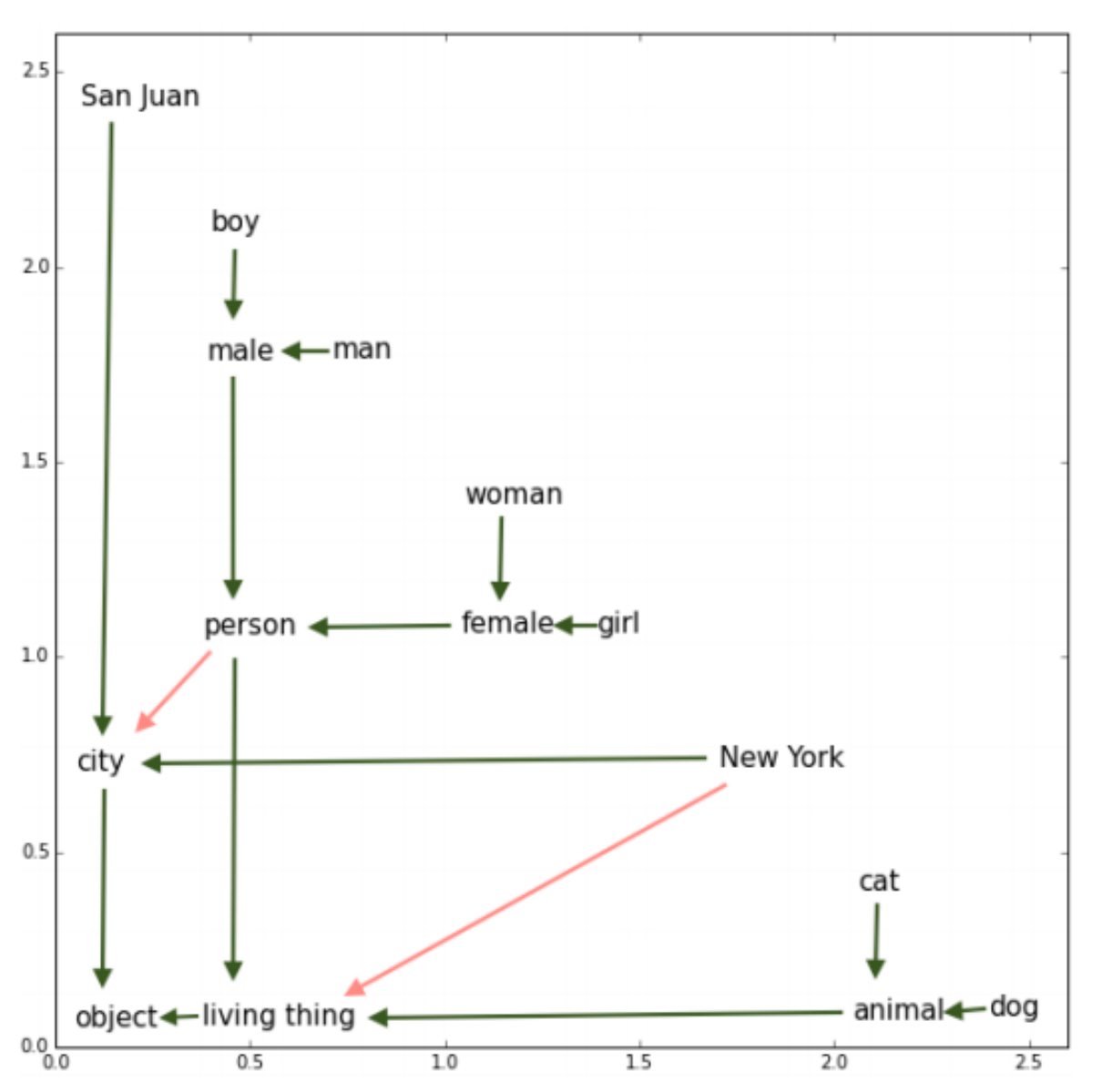


ACE for Order Embeddings

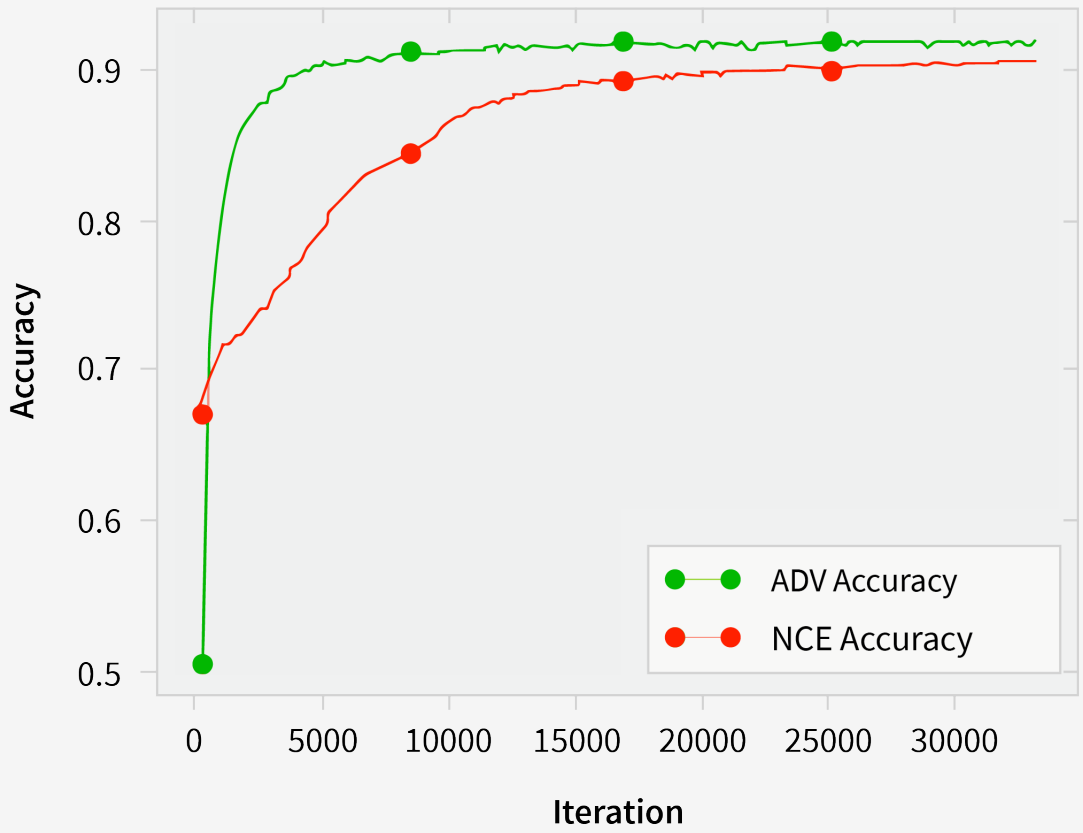
Hypernym Prediction: A hypernym pair is a pair of concepts where the first concept is a specialization or an instance of the second.

- Learning embeddings that are hierarchy preserving. The Root Node is at the origin and all other embeddings lie on the positive semi-space
- Constraint enforces the magnitude of the parent's embedding to be smaller than child's in every dimension
- Sibling nodes are not subjected to this constraint.

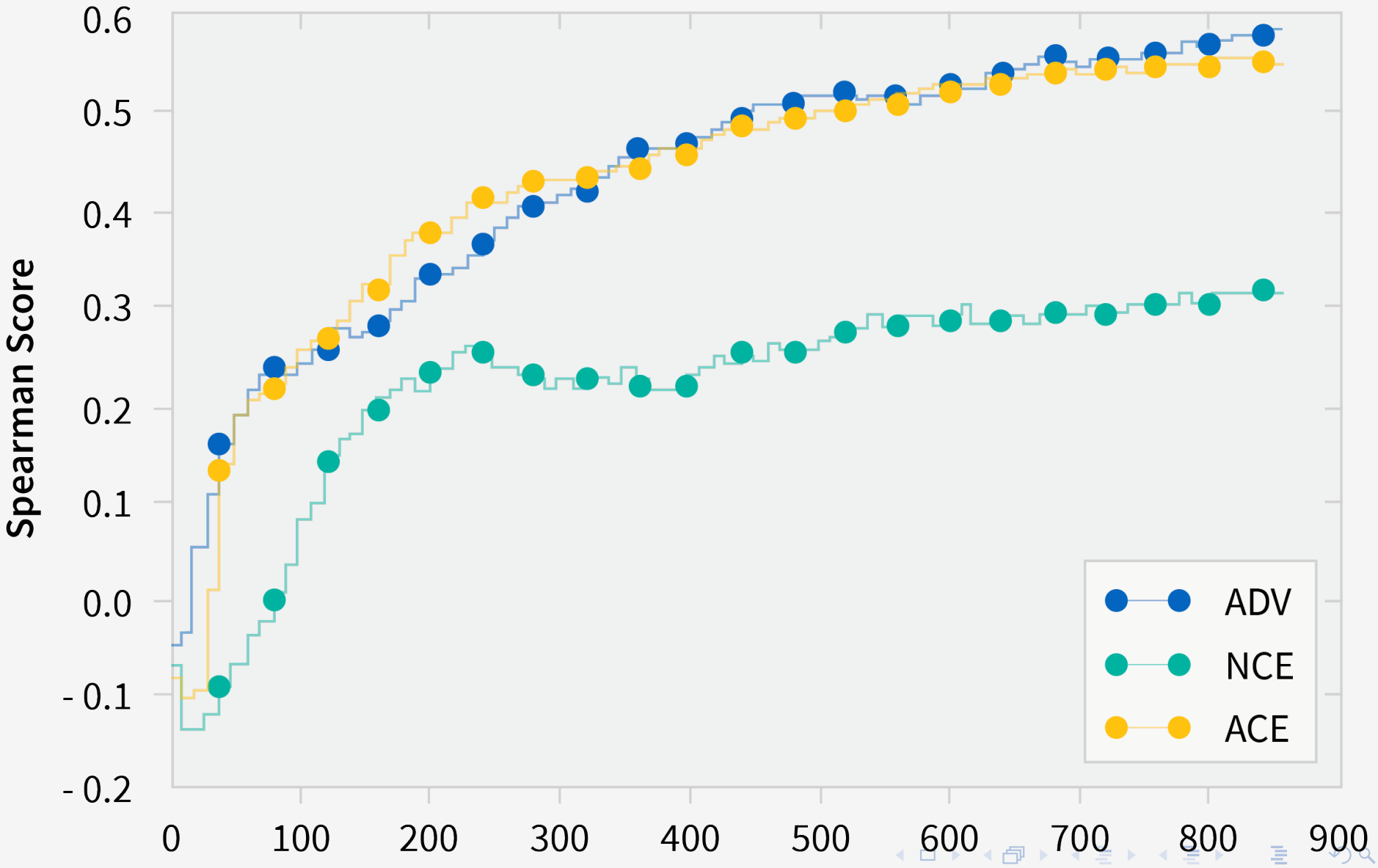
Order Embeddings (Vendrov et. al 2016)



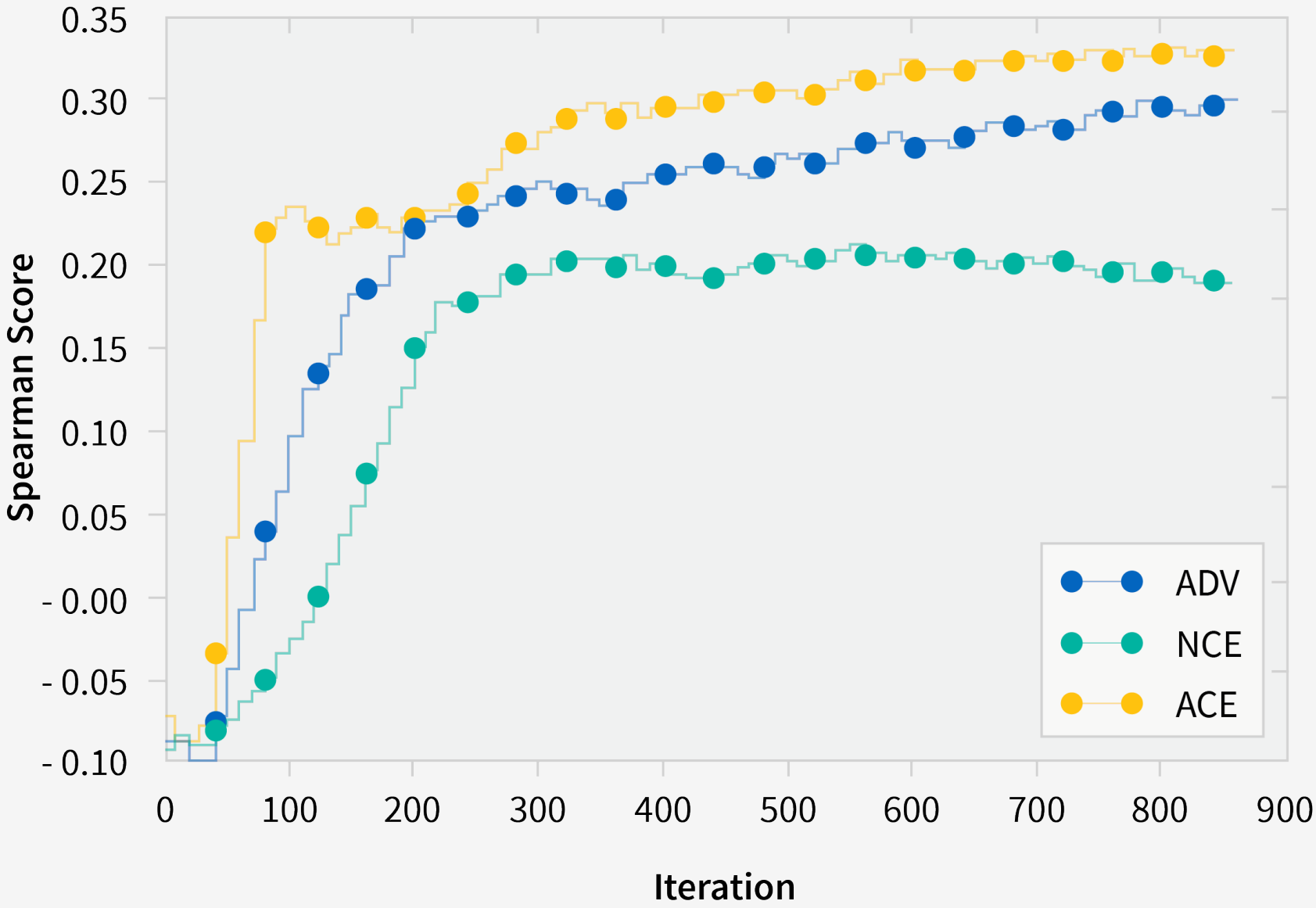
ACE for Order Embeddings



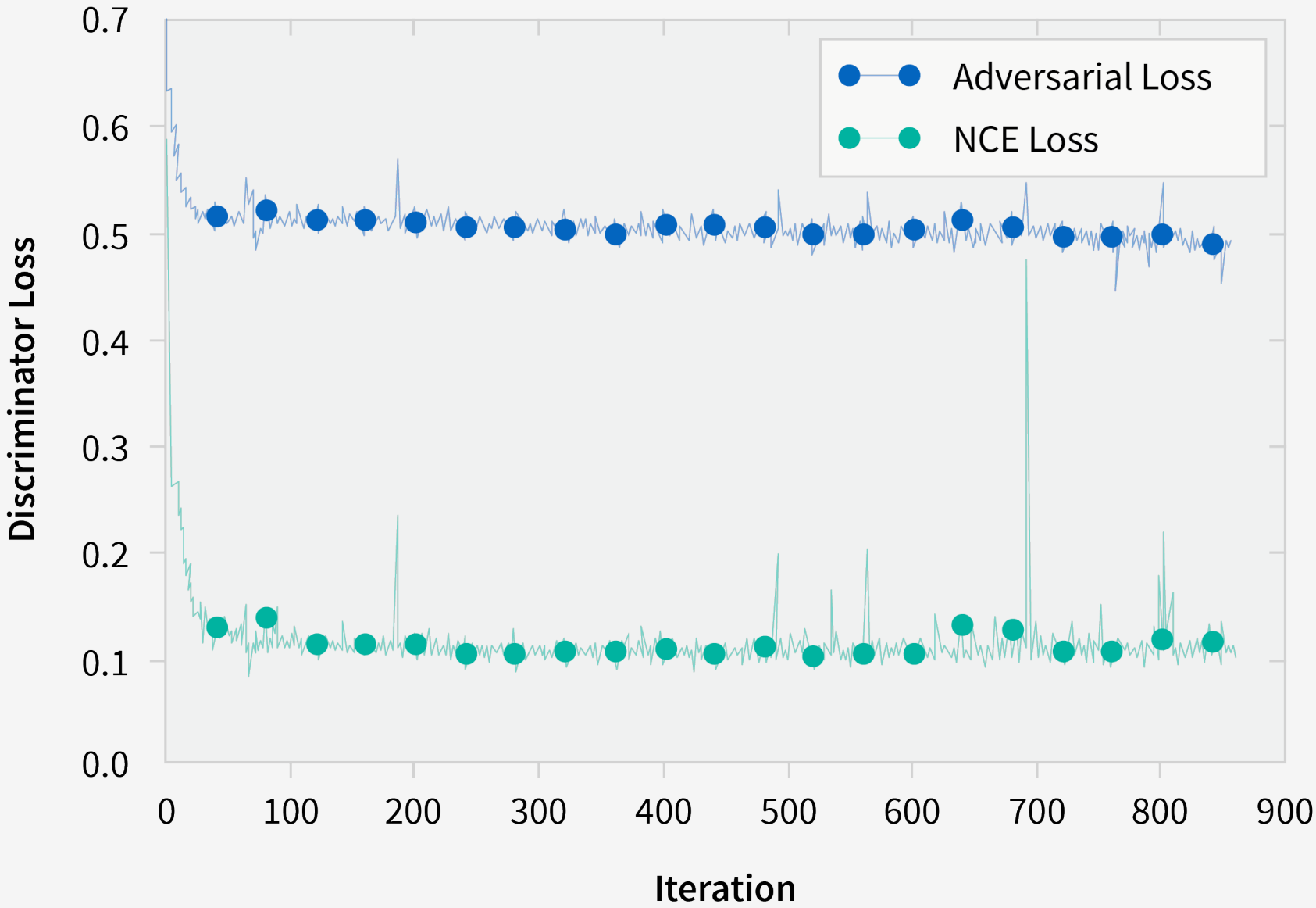
ACE for Word Embeddings: WordSim353



ACE for Word Embeddings: Stanford Rare Word



Discriminator Loss on NCE vs. Adversarial Examples



Nearest Neighbors for NCE vs. ACE

	Queen	King	Computer	Man	Woman
Skip-Gram NCE Top 5	princess	prince	computers	woman	girl
	king	queen	computing	boy	man
	empress	kings	software	girl	prostitute
	pxqueen	emperor	microcomputer	stranger	person
	monarch	monarch	mainframe	person	divorcee
Skip-Gram NCE Top 45-50	sambiria	eraric	hypercard	angiomata	suitor
	phongsri	mumbere	neurotechnology	someone	nymphomaniac
	safrit	empress	lgp	bespectacled	barmaid
	mcelvoy	saxonvm	pcs	hero	redheaded
	tsarina	pretender	keystroke	clown	jew
Skip-Gram ACE Top 5	princess	prince	software	woman	girl
	prince	vi	computers	girl	herself
	elizabeth	kings	applications	tells	man
	duke	duke	computing	dead	lover
	consort	iii	hardware	boy	tells
Skip-Gram ACE Top 45-50	baron	earl	files	kid	aunt
	abbey	holy	information	told	maid
	throne	cardinal	device	revenge	wife
	marie	aragon	design	magic	lady
	victoria	princes	compatible	angry	bride

Table 1: Top 5 Nearest Neighbors of Words followed by Neighbors 45-50 for different Models.

Questions?

BlogPost:

http://borealisai.com/2018/07/13/adversarial-contrastive-estimation-harder-better-faster-stronger/