Inference and Introspection in Deep Generative Models of Sparse Data

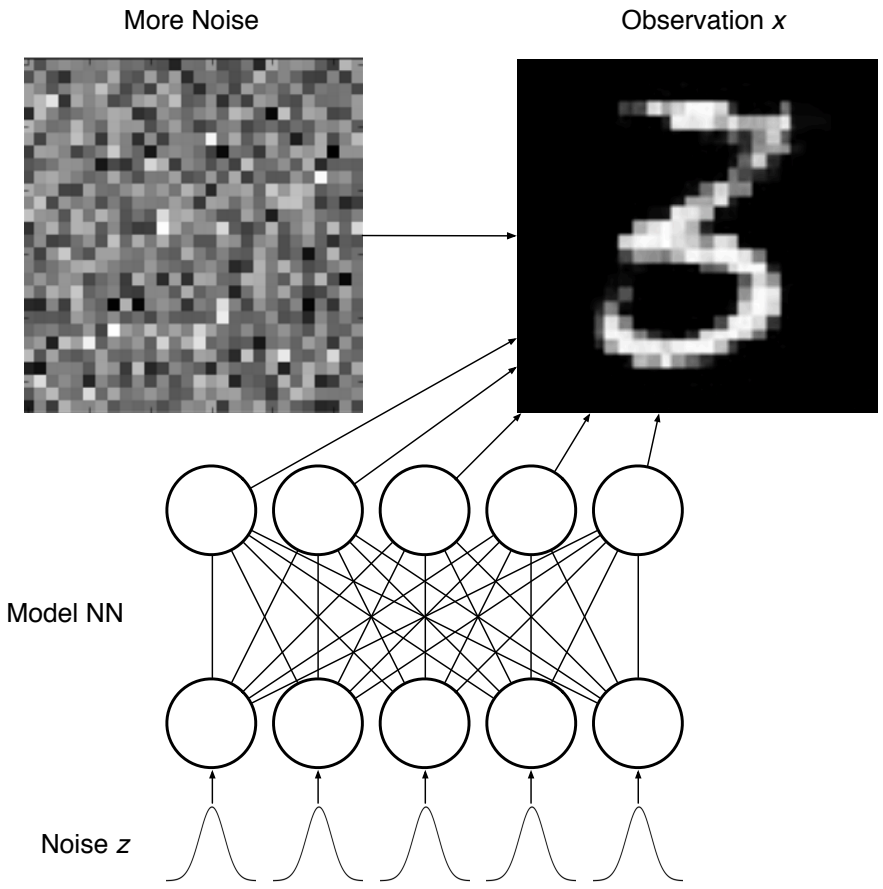
Rahul Krishnan
NYU

Matthew D. Hoffman
Adobe Research

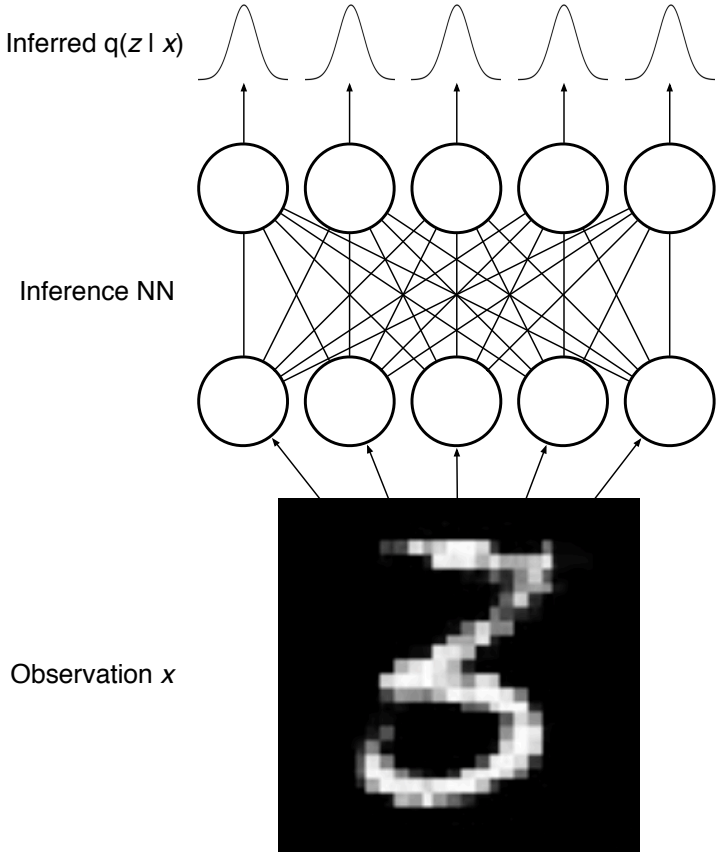
(Presenting)

Background: Variational Autoencoders (VAEs)

Deep Latent Gaussian Model



Recognition Network



Sparse Data and Deep Generative Models

VAEs/DLGMs are almost exclusively applied to images, not text or other high-dimensional, sparse data (some exceptions: Miao et al., 2016; Bowman et al., 2016).

Why? Sparse "texty" data is everywhere! (Documents, social networks, ratings/views/listens, medical diagnoses, etc., etc.)

Our Hypothesis: Local Optima

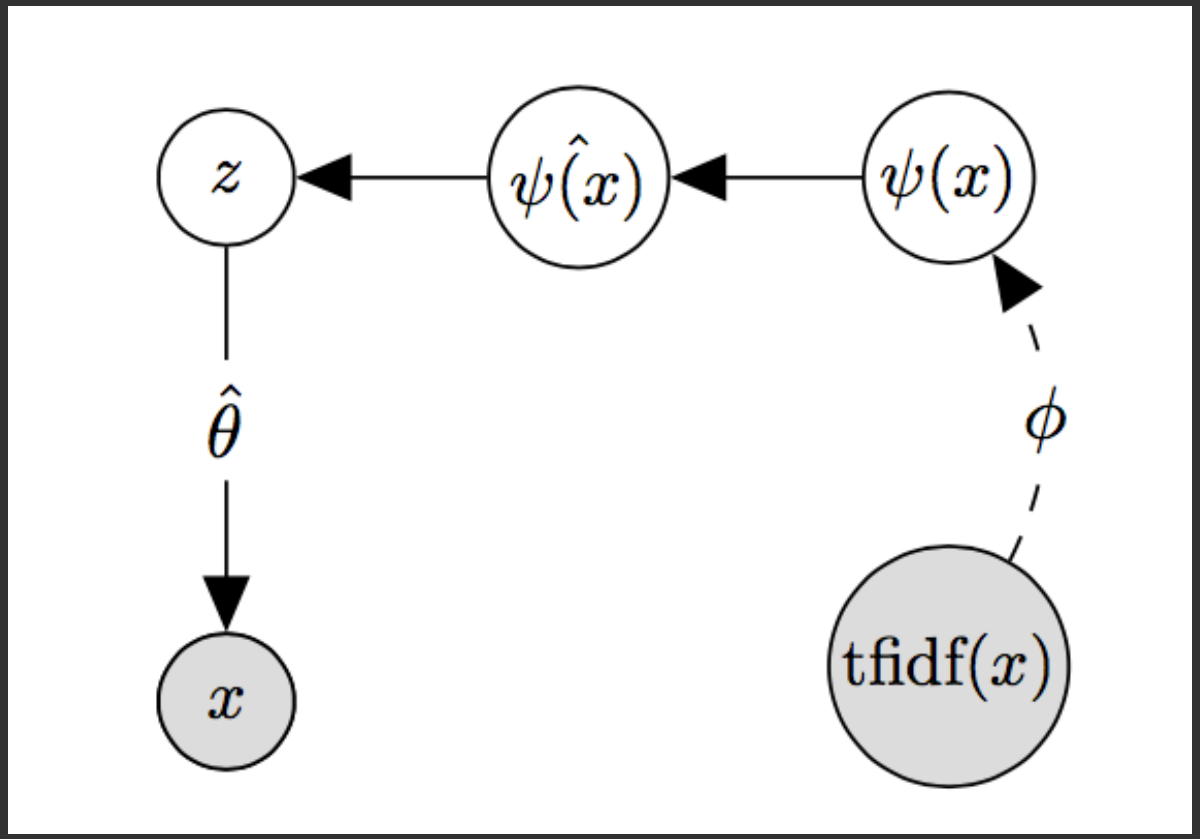
Many observed words are rare.

The inference network needs to see each rare word a few times to learn to interpret that word.

But if the inference network's inferences are bad, the generative network's gradient signal will be bad.

Our Approach

Use the inference network as an *initializer* and optimize from there. (Cf. Hjelm et al. 2016)



Experimental Results: Held-Out Perplexity

Deeper is better for large datasets.

TF-IDF often helps, never hurts.

Secondary optimization *at training time* always helps deeper models.

Secondary optimization *at test time* is very important.

Come to the poster for details.

Introspection and Interpretability

What makes "shallow" models (e.g., LDA) interpretable?

A hypothesis: It's because the parameters encode a *linear* relationship between latent vectors z and observations x:

$$\mathbb{E}[x|z] = \theta z; \quad \frac{\partial \mathbb{E}[x|z]}{\partial z} = \theta.$$

Introspection and Interpretability

What makes "shallow" models (e.g., LDA) interpretable?

A hypothesis: It's because the parameters encode a *linear* relationship between latent vectors z and observations x:

$$\mathbb{E}[x|z] = \theta z; \quad \frac{\partial \mathbb{E}[x|z]}{\partial z} = \theta.$$

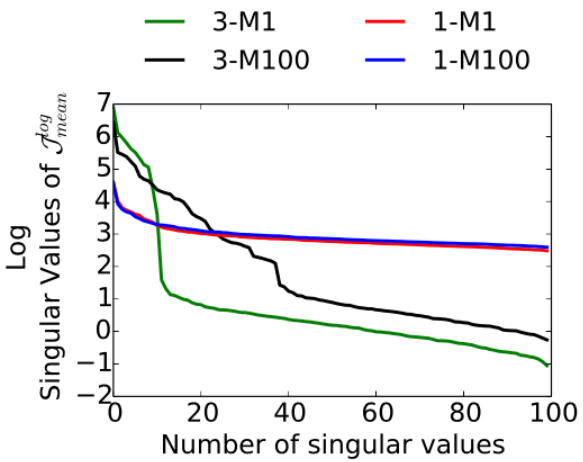
For nonlinear models this Jacobian depends on z, so we average over samples of z:

$$\mathcal{J}^{\text{prob}}_{\text{mean}} \triangleq \int_z p(z) \frac{\partial \mathbb{E}[x|z]}{\partial z} dz; \quad \mathcal{J}^{\text{log}}_{\text{mean}} \triangleq \int_z p(z) \frac{\partial \log \mathbb{E}[x|z]}{\partial z} dz.$$

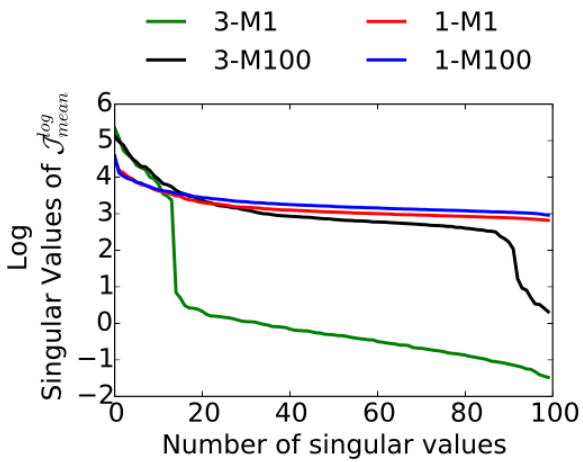
Diagnosing Pruning

The singular value spectrum of J tells us how many latent dimensions are being used.

We find that secondary optimization dramatically reduces overpruning in deep models.



(a) RCV2



(b) Wikipedia

Word Embeddings

We can use the rows of J as word embeddings, with sensible results:

Query	**Neighborhood**
intelligence	espionage, secrecy, interrogation
zen	dharma, buddhism, buddhas, meditation
artificial	artificially, molecules, synthetic, soluble
military	civilian, armys, commanders, infantry

Query	**Neighborhood**
Superman II	Superman: The Movie, Superman III, Superman IV: The Quest for Peace
Casablanca	Citizen Kane, The Treasure of the Sierra Madre, Working with Orson Welles, The Millionairess
The Princess Bride	The Breakfast Club, Sixteen Candles, Groundhog Day, Beetlejuice
12 Angry Men	To Kill a Mockingbird, Rear Window, Mr. Smith Goes to Washington, Inherit the Wind

Contextual Word Embeddings

Using the Jacobian matrix for a particular z gives context-dependent embeddings:

Word	**Context**	**Neighboring Words**
crane	construction	lifting, usaaf, spanned, crushed, lift
	bird	erected, parkland, locally, farmland
bank	river	watershed, footpath, confluence, drains
	money	banking, government, bankers, comptroller
fires	burn	ignition, combustion, engines, fuel, engine
	layoff	thunderstorm, grassy, surrounded, walkway

Summary

We addressed some issues with fitting VAEs to sparse, high-dimensional "texty" data.

We proposed a simple way to examine what deep models of text are learning.

Come to the poster and let's talk about details, extensions, and applications!