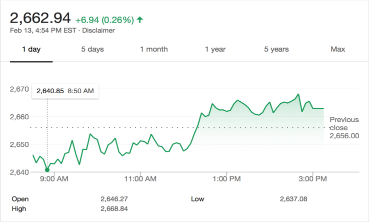
Deep Sequence Models

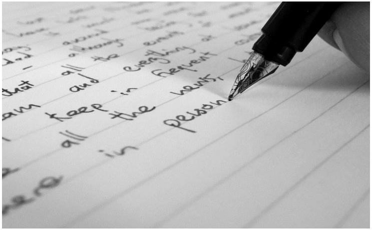
Context Representation, Regularization, and Application to Language

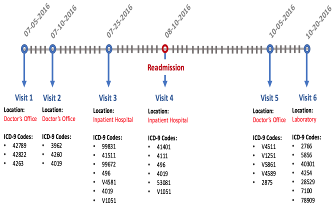
Adji Bousso Dieng

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

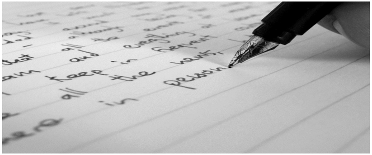
All Data Are Born Sequential

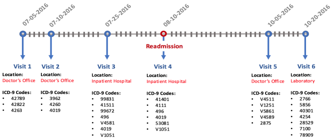






"Time underlies many interesting human behaviors."– Elman, 1990.





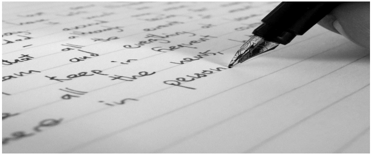
Why model these data?

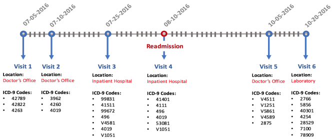
$\rightarrow$ to help in decision making

$\rightarrow$ to generate more of it

$\rightarrow$ to predict and forecast

$\rightarrow$ ... for science





How do we model these data?

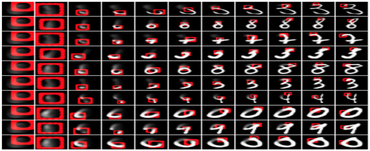
$\rightarrow$ need to capture all the dependencies

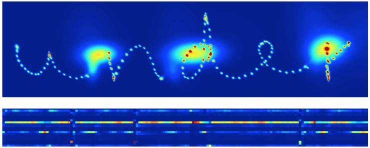
$\rightarrow$ need to account for dimensionality

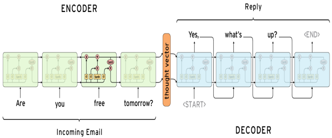
$\rightarrow$ need to account for seasonality

... It's complicated.

Recurrent Neural Networks: Successes





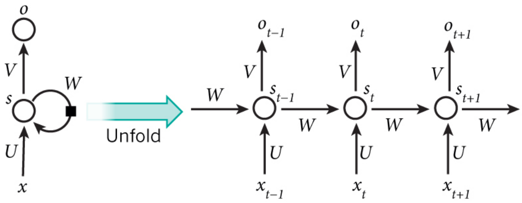


→ Image generation (Gregor+, 2015)

→ Text generation (Graves, 2013)

→ Machine translation (Sutskever+, 2014)

Recurrent Neural Networks: Challenges



$$s_t = f_W(x_t, s_{t-1})$$
$$s_t = g(s_0, x_t, x_{t-1}, ..., x_0) \text{ and } g = f(f(f(...)))$$
$$o_t = \text{softmax}(V s_t)$$

→ Vanishing and exploding gradients.

→ V can be very high-dimensional

→ Hidden state has limited capacity.

→ The RNN is trying to do too many things at once.

Context Representation

What Is Context?

The U.S. presidential race is not only drawing attention and controversy in the United States – it is being closely watched across the globe. But what does the rest of the world think about a campaign that has already thrown up one surprise after another? CNN asked 10 journalists for their take on the race so far, and what their country might be hoping for in America's next –

→ local context:

few words preceding the word to predict

order matters.

defines syntax

What Is Context?

The U.S. presidential race is not only drawing attention and controversy in the United States – it is being closely watched across the globe. But what does the rest of the world think about a campaign that has already thrown up one surprise after another? CNN asked 10 journalists for their take on the race so far, and what their country might be hoping for in America's next –

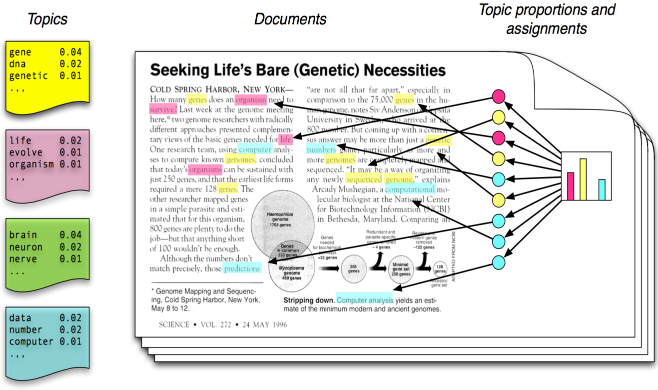
$\rightarrow$ global context:

words in the same document as the word to predict

order does not matter.

defines semantic

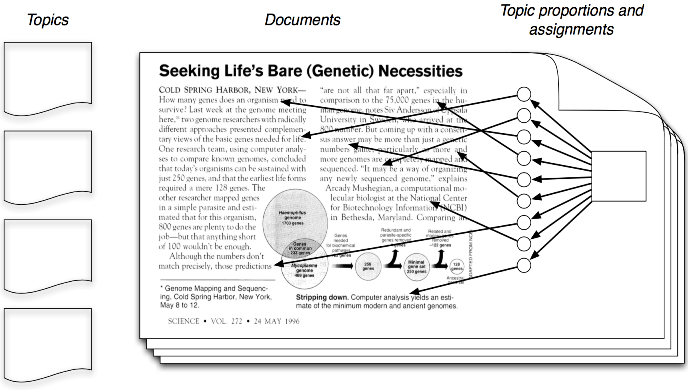
Topics As Context (1/3)



Generative process

source: David Blei

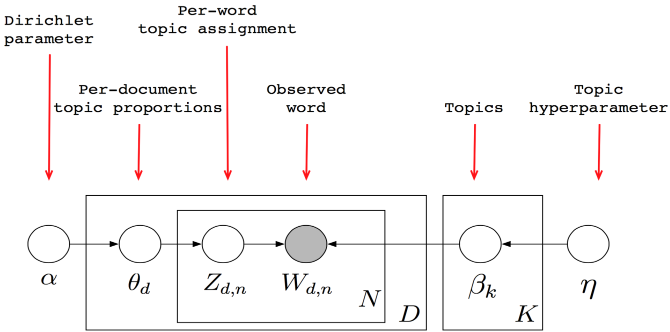
Topics As Context (2/3)



Posterior inference

source: David Blei

Topics As Context (3/3)

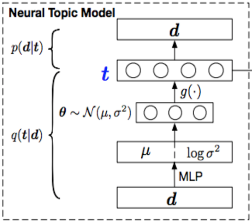


source: David Blei

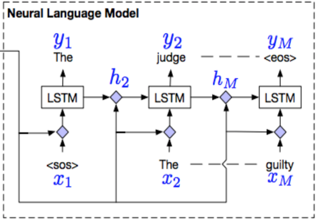
$$\theta_d \sim \text{Dir}(\alpha) \ ; \ \beta_k \sim \text{Dir}(\eta) \ ; \ z_{dn} \sim \text{Multinomial}(\theta_d)$$

$$w_{dn} \sim \text{Multinomial}(\beta_{z_{dn}})$$

Composing Topics And RNNs (1/3)



Topic	Proportion
Law	0.36
Art	0.03
Market	0.10
Travel	0.07
Company	0.09
Politics	0.15
Sport	0.01
Education	0.11
Medical	0.02
Army	0.06



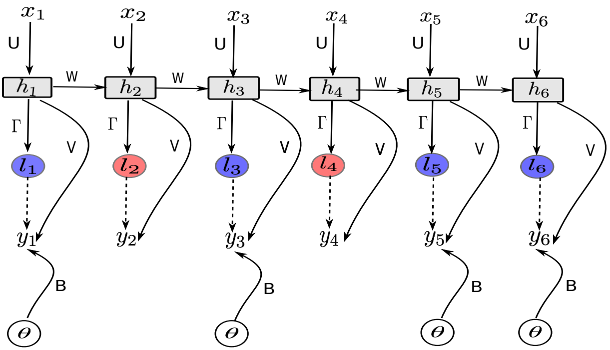
source: Wang+, 2017

→ RNN focuses on capturing local correlations (syntax model)

→ Topic model captures global dependencies (semantic model)

→ Combine both to make predictions

Composing Topics And RNNs (2/3)

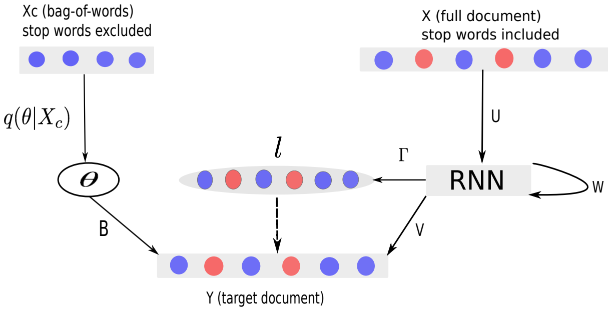


source: Dieng+, 2017

$$h_t = f_W(x_t, h_{t-1}) \ ; \ l_t \sim \text{Bernoulli}(\sigma(\Gamma^\top h_t))$$

$$y_t \sim \text{softmax}(V^\top h_t + (1 - l_t)B^\top \theta)$$

Composing Topics And RNNs (3/3)



source: Dieng+, 2017

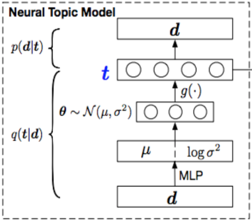
$\rightarrow$ Choose $q(\theta \mid X_c)$ to be an MLP

$\rightarrow$ Choose $p(\theta)$ to be standard Gaussian: $\theta = g(\mathcal{N}(0, I_K))$

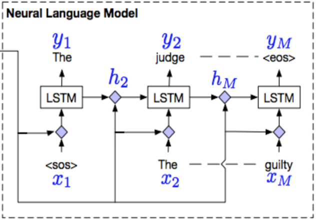
$\rightarrow$ Maximize the ELBO:

$$\text{ELBO} = E_{q(\theta \mid X_c)} \left[\sum_{t=1}^{T} \log p(y_t, l_t | \theta; h_t) \right] - KL\left(q(\theta \mid X_c) \,\|\, p(\theta)\right)$$

Composing Topics And RNNs (3/3)



Topic	Proportion
Law	0.36
Art	0.03
Market	0.10
Travel	0.07
Company	0.09
Politics	0.15
Sport	0.01
Education	0.11
Medical	0.02
Army	0.06



source: Wang+, 2017

→ has been extended to mixture of experts (Wang+, 2017)

→ has been applied to conversation modeling (Wen+, 2017)

Some Results On Language Modeling (1)

10 Neurons	Valid	Test
RNN (no features)	239.2	225.0
RNN (LDA features)	197.3	187.4
TopicRNN	184.5	172.2
TopicLSTM	188.0	175.0
TopicGRU	178.3	**166.7**

100 Neurons	Valid	Test
RNN (no features)	150.1	142.1
RNN (LDA features)	132.3	126.4
TopicRNN	128.5	122.3
TopicLSTM	126.0	118.1
TopicGRU	118.3	**112.4**

300 Neurons	Valid	Test
RNN (no features)	–	124.7
RNN (LDA features)	–	113.7
TopicRNN	118.3	112.2
TopicLSTM	104.1	99.5
TopicGRU	99.6	**97.3**

source: Dieng+, 2017

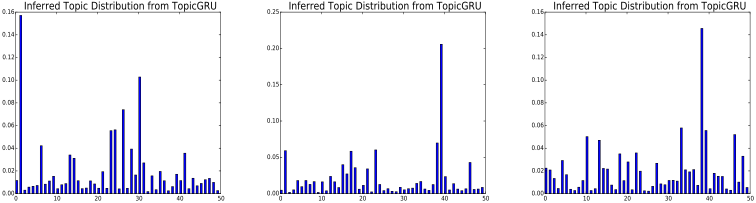
$\rightarrow$ Perplexity on Penn Treebank dataset (the lower the better)

$\rightarrow$ Three different network capacity

$\rightarrow$ Adding topic features is always better

$\rightarrow$ Doing so jointly is even better

Some Results On Language Modeling (2)



source: Dieng+, 2017

$\rightarrow$ Document distribution for 3 different documents with TopicGRU

$\rightarrow$ Different topics get picked up for different documents

Some Results On Language Modeling (3)

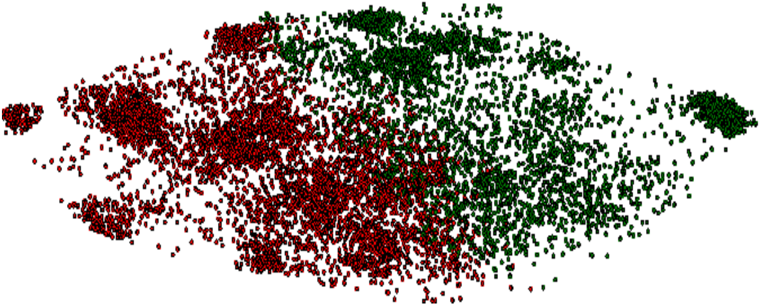
Dataset	army	animal	medical	market	lottery	terrorism	law	art	transportation	education
APNEWS	afghanistan	animals	patients	zacks	casino	syria	lawsuit	album	airlines	students
	veterans	dogs	drug	cents	mega	iran	damages	music	fraud	math
	soldiers	zoo	fda	earnings	lottery	militants	plaintiffs	film	scheme	schools
	brigade	bear	disease	keywords	gambling	al-qaida	filed	songs	conspiracy	education
	infantry	wildlife	virus	share	jackpot	korea	suit	comedy	flights	teachers
	horror	**action**	**family**	**children**	**war**	**detective**	**sci-fi**	**negative**	**ethic**	**epsiode**
IMDB	zombie	martial	rampling	kids	war	eyre	alien	awful	gay	season
	slasher	kung	relationship	snoopy	che	rochester	godzilla	unfunny	school	episodes
	massacre	li	binoche	santa	documentary	book	tarzan	sex	girls	series
	chainsaw	chan	marie	cartoon	muslims	austen	planet	poor	women	columbo
	gore	fu	mother	parents	jews	holmes	aliens	worst	sex	batman
	environment	**education**	**politics**	**business**	**facilities**	**sports**	**art**	**award**	**expression**	**crime**
BNC	pollution	courses	elections	corp	bedrooms	goal	album	john	eye	police
	emissions	training	economic	turnover	hotel	score	band	award	looked	murder
	nuclear	students	minister	unix	garden	cup	guitar	research	hair	killed
	waste	medau	political	net	situated	ball	music	darlington	lips	jury
	environmental	education	democratic	profits	rooms	season	film	speaker	stared	trail

source: Wang+, 2017

→ Topics for three different datasets

→ Shows top five words of ten random topics

Some Results On Document Classification



source: Dieng+, 2017

→ Sentiment classification on IMDB

→ Feature extraction: concatenate RNN feature and Topic feature

→ PCA + K-Means

Some Results On Document Classification

Model	Reported Error rate
BoW (bnc) (Maas et al., 2011)	12.20%
BoW ($b\Delta$ tć) (Maas et al., 2011)	11.77%
LDA (Maas et al., 2011)	32.58%
Full + BoW (Maas et al., 2011)	11.67%
Full + Unlabelled + BoW (Maas et al., 2011)	11.11%
WRRBM (Dahl et al., 2012)	12.58%
WRRBM + BoW (bnc) (Dahl et al., 2012)	10.77%
MNB-uni (Wang & Manning, 2012)	16.45%
MNB-bi (Wang & Manning, 2012)	13.41%
SVM-uni (Wang & Manning, 2012)	13.05%
SVM-bi (Wang & Manning, 2012)	10.84%
NBSVM-uni (Wang & Manning, 2012)	11.71%
seq2-bown-CNN (Johnson & Zhang, 2014)	14.70%
NBSVM-bi (Wang & Manning, 2012)	8.78%
Paragraph Vector (Le & Mikolov, 2014)	7.42%
SA-LSTM with joint training (Dai & Le, 2015)	14.70%
LSTM with tuning and dropout (Dai & Le, 2015)	13.50%
LSTM initialized with word2vec embeddings (Dai & Le, 2015)	10.00%
SA-LSTM with linear gain (Dai & Le, 2015)	9.17%
LM-TM (Dai & Le, 2015)	7.64%
SA-LSTM (Dai & Le, 2015)	7.24%
Virtual Adversarial (Miyato et al. 2016)	**5.91%**
TopicRNN	**6.28%**

source: Dieng+, 2017

Regularization

Co-adaptation

"When a neural network overfits badly during training, its hidden states depend very heavily on each other."

– Hinton, 2012

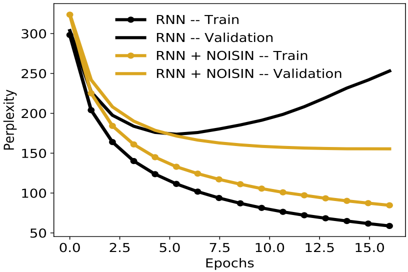
Noise As Regularizer

$\rightarrow$ Define a noise-injected RNN as:

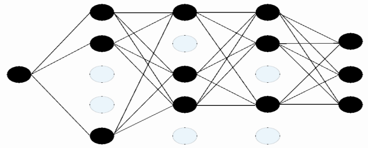
$\epsilon_{1:T} \sim \varphi(\cdot; \mu, \gamma)$; $z_t = g_W(x_t, z_{t-1}, \epsilon_t)$ and $p(y_t \,|\, y_{1:t-1}) = p(y_t \,|\, z_t)$

$\rightarrow$ The likelihood $p(y_t \,|\, z_t)$ is in the exponential family

$\rightarrow$ Different noise ϵ at each layer



Dropout



$\rightarrow$ For the LSTM this is:

$$
\begin{aligned}
f_t &= \sigma(W_{x1}^\top x_{t-1} \odot \epsilon_t^{xf} + W_{h1}^\top h_{t-1} \odot \epsilon_t^{hf}) \\
i_t &= \sigma(W_{x2}^\top x_{t-1} \odot \epsilon_t^{xi} + W_{h2}^\top h_{t-1} \odot \epsilon_t^{hi}) \\
o_t &= \sigma(W_{x4}^\top x_{t-1} \odot \epsilon_t^{xo} + W_{h4}^\top h_{t-1} \odot \epsilon_t^{ho}) \\
c_t &= f_t \odot c_{t-1} + i_t \odot \tanh(W_{x3}^\top x_{t-1} \odot \epsilon_t^{xc} + W_{h3}^\top h_{t-1} \odot \epsilon_t^{hc}) \\
z_t^{dropout} &= o_t \odot \tanh(c_t).
\end{aligned}
$$

NOISIN: Unbiased Noise Injection

→ *Strong unbiasedness* condition

$$\mathbb{E}_{p(z_t(\epsilon_{1:t}) \mid z_{t-1})}\left[z_t(\epsilon_{1:t})\right] = s_t$$

→ *Weak unbiasedness* condition

$$\mathbb{E}_{p(z_t(\epsilon_{1:t}) \mid z_{t-1})}\left[z_t(\epsilon_{1:t})\right] = f_W(x_{t-1}, z_{t-1})$$

→ Under unbiasedness the underlying RNN is preserved

→ Examples: additive and multiplicative noise

$$g_W(x_{t-1}, z_{t-1}, \epsilon_t) = f_W(x_{t-1}, z_{t-1}) + \epsilon_t$$
$$g_W(x_{t-1}, z_{t-1}, \epsilon_t) = f_W(x_{t-1}, z_{t-1}) \odot (1 + \epsilon_t)$$

→ Dropout does not meet this requirement; it is *biased*

NOISIN: The Objective

→ NOISIN maximizes the following objective

$$\mathcal{L} = E_{p(\epsilon_{1:T})}\left[\log p(x_{1:T} | z_{1:T}(\epsilon_{1:T}))\right]$$

→ In more detail this is

$$\mathcal{L} = \sum_{t=1}^{T} E_{p(\epsilon_{1:t})}\Big[\log p(x_t | z_t(\epsilon_{1:t}))\Big]$$

→ Notice this objective is a Jensen bound on the marginal log-likelihood of the data,

$$\mathcal{L} \leq \log E_{p(\epsilon_{1:T})}\left[p(x_{1:T} | z_{1:T}(\epsilon_{1:T}))\right] = \log p(x_{1:T})$$

NOISIN: Connections

$$\mathcal{L} = \sum_{t=1}^{T} E_{p(\epsilon_{1:t})}\Big[\log p(x_t | z_t(\epsilon_{1:t}))\Big]$$

$\rightarrow$ Ensemble method

average the predictions of infinitely many RNNs

at each time step

$\rightarrow$ Empirical Bayes

estimate the parameters of the prior on the hidden states

Some Results On Language Modeling (1/2)

	Medium			Large		
Method	γ	Dev	Test	γ	Dev	Test
None	$--$	115	109	$--$	123	123
Gaussian	1.10	76.2	71.8	1.37	73.2	69.1
Logistic	1.06	76.4	72.3	1.39	73.6	69.3
Laplace	1.06	76.6	72.4	1.39	73.7	69.4
Gamma	1.06	78.2	74.5	1.39	73.6	69.5
Bernoulli	0.41	**75.7**	**71.4**	0.33	**72.8**	**68.3**
Gumbel	1.06	76.2	72.7	1.39	73.5	69.5
Beta	1.07	76.0	71.4	1.50	74.4	70.2
Chi	1.50	84.5	80.7	1.20	79.2	75.5

	Medium			Large		
Method	γ	Dev	Test	γ	Dev	Test
Dropout (D)	$--$	80.2	77.0	$--$	78.6	75.3
D + Gaussian	0.53	73.4	70.4	0.92	**70.0**	**66.1**
D + Logistic	0.53	73.0	69.9	0.84	69.8	66.4
D + Laplace	0.53	73.1	70.0	0.92	69.9	66.6
D + Gamma	0.38	73.5	70.3	0.92	71.1	68.2
D + Bernoulli	0.80	73.3	70.1	0.50	**70.0**	**66.1**
D + Gumbel	0.46	74.5	71.2	0.92	70.2	67.1
D + Beta	0.20	**73.0**	**69.2**	0.70	70.0	66.2
D + Chi	0.29	76.1	72.8	0.82	73.0	70.0

$\rightarrow$ Perplexity on the Penn Treebank (lower the better)

$\rightarrow$ D + Distribution is Dropout-LSTM with NOISIN

$\rightarrow$ Studied many noise distributions: only variance matters

$\rightarrow$ Noise is scaled to enjoy unbounded variance

Some Results On Language Modeling (2/2)

	Medium			Large		
Method	γ	Dev	Test	γ	Dev	Test
None	−−	141	136	−−	176	140
Gaussian	1.00	92.7	87.8	1.37	87.7	83.4
Logistic	1.00	93.2	88.4	1.28	88.1	83.5
Laplace	1.00	95.3	89.8	1.28	88.0	83.4
Gamma	0.72	97.6	92.9	1.39	89.2	84.5
Bernoulli	0.54	91.2	86.6	0.41	86.9	83.0
Gumbel	1.00	95.4	90.9	1.28	88.7	84.0
Beta	0.80	**91.1**	**87.2**	1.50	**86.9**	**82.9**
Chi	0.20	111	105	1.50	99.0	92.9

	Medium			Large		
Method	γ	Dev	Test	γ	Dev	Test
Dropout (D)	−−	88.7	84.8	−−	95.0	91.0
D + Gaussian	0.50	86.3	82.3	0.69	81.4	77.7
D + Logistic	0.40	86.4	82.5	0.77	81.6	78.1
D + Laplace	0.40	**85.6**	**82.1**	0.61	83.2	79.1
D + Gamma	0.30	86.5	82.4	0.61	85.5	81.3
D + Bernoulli	0.50	100.6	94.4	0.64	**80.8**	**76.8**
D + Gumbel	0.30	86.4	82.4	0.53	83.7	80.1
D + Beta	0.10	86.2	82.3	0.60	81.5	77.9
D + Chi	0.20	92.0	87.4	0.29	87.1	82.8

$\rightarrow$ Perplexity on the Wikitext-2 (lower the better)

$\rightarrow$ D + Distribution is Dropout-LSTM with NOISIN

$\rightarrow$ Studied many noise distributions: only variance matters

$\rightarrow$ Noise is scaled to enjoy unbounded variance

Lessons Learned So Far

Context representation

$\rightarrow$ Need to rethink long-term dependencies (for language)

$\rightarrow$ Combine a syntax model and a semantic model

$\rightarrow$ Topic models are good semantic models

$\rightarrow$ TopicRNN is a deep generative model that uses topics as context for RNNs

Regularization

$\rightarrow$ Noise can be used to avoid co-adaptation

$\rightarrow$ It should be injected *unbiasedly* into the hidden units of the RNN

$\rightarrow$ This is some form of model averaging and is like empirical Bayes

$\rightarrow$ NOISIN is simple yet significantly improves RNN-based models

More Challenges to Tackle

- $\rightarrow$ Scalability
- $\rightarrow$ Incorporating prior knowledge
- $\rightarrow$ Improving generation