Topic-Guided Variational Autoencoders for Text Generation

Wenlin Wang, Zhe Gan, Hongteng Xu, Ruiyi Zhang, Guoyin Wang, Dinghan Shen, Changyou Chen, Lawrence Carin

Duke University, Microsoft Dynamics 365 AI Research, Infinia ML, Inc, University at Buffalo

November 7, 2019

Overview

Framework for Variational Inference

- MLE: $\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}}\ p(x;\theta)$
  - Intractable when latent variables z with prior $p(z;\theta)$ are introduced
  - Marginal likelihood: $p(x;\theta) = \int p(x,z;\theta)dz = \int p(x|z;\theta)p(z;\theta)dz$
  - Posterior: $p(z|x;\theta) = \frac{p(x|z;\theta)p(z;\theta)}{p(x;\theta)}$ also intractable
- Assumptions
  - $p(z;\theta) \sim \mathcal{N}(0, I)$
  - $p(x|z;\theta) \sim \mathcal{N}(f(z), cI)$
- This implies posterior $p(z|x;\theta)$ also follow a Gaussian distribution
  - The mean and variance depends on the function f which can be complex

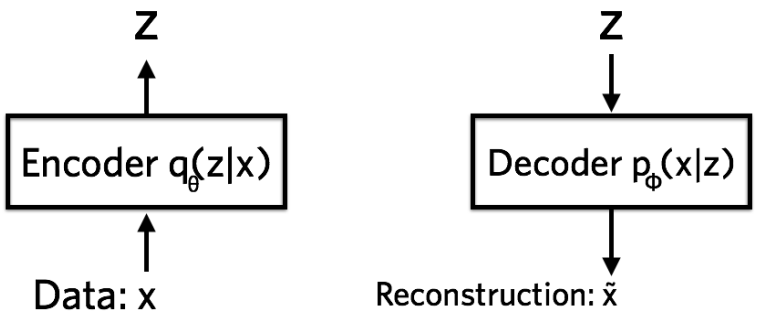
VI Formulation

- Variational Inference (VI) is a technique to approximate complex distributions
- Approximate $p(z|x;\theta)$ by $q(z|x;\phi) \sim \mathcal{N}(g(x), h(x))$
- Find best approximation by minimizing the Kullback-Leibler divergence between the $q(z|x;\phi)$ and the target $p(z|x;\theta)$

$$
\begin{aligned}
\phi^* &= \operatorname*{argmin}_{\phi} \mathrm{KL}(q(z|x;\phi)||p(z|x;\theta)) \\
&= \operatorname*{argmin}_{\phi} \mathbb{E}_{q(z|x;\phi)} \log q(z|x;\phi) - \mathbb{E}_{q(z|x;\phi)} \log \frac{p(x|z;\theta)p(z;\theta)}{p(x;\theta)} \\
&= \operatorname*{argmin}_{\phi} \mathbb{E}_q \log q_\phi(z|x) - \mathbb{E}_q p_\theta(x|z) - \mathbb{E}_q p_\theta(z) + \mathbb{E}_q p_\theta(x) \\
&= \operatorname*{argmax}_{\phi} \mathbb{E}_q p_\theta(x|z) - \mathrm{KL}(q_\phi(z|x)||p_\theta(z)) \\
&= \operatorname*{argmax}_{\phi} \mathbb{E}_q(-\frac{\|x - f(z)\|^2}{2c}) - \mathrm{KL}(q_\phi(z|x)||p_\theta(z))
\end{aligned}
$$

Variational Autoencoder (VAE)

- VAE is a autoencoder that employs re-parameterization of variational lower bound and optimized using standard gradient methods



- Define $p_\theta(x|z)$ to be the decoder network parameterized by θ, and $q_\phi(z|x)$ be the encoder network parameterized by ϕ
- VAE encode an input as a distribution over the latent space instead of a single point
- Regularity of the latent space allows the generative process to be possible through two main properties
  - Continuity
  - Completeness
- Want to regularise both the covariance matrix and the mean of the distributions returned by the encoder (Standard Gaussian)

Evidence Lower Bound (ELBO)

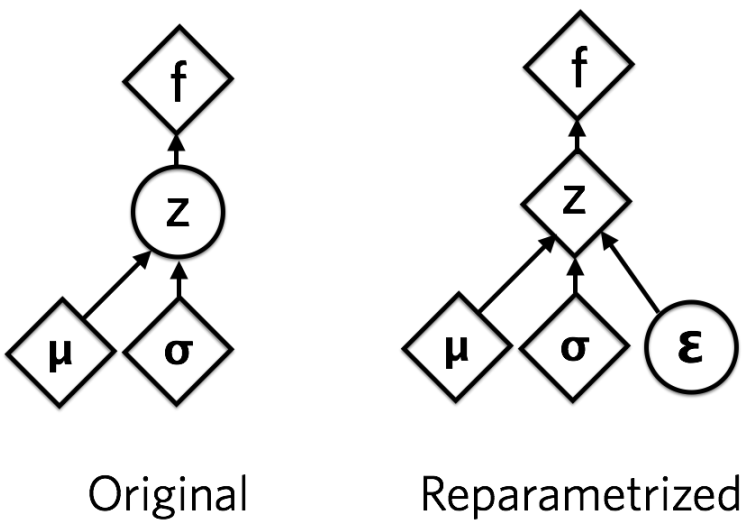
- Integrating over evidence $p(x) = \int p(x|z)p(z)dz$ requires exponential time to compute
- Since from Jensen's inequality we know that when f is concave, $f(E[X]) \geq E[f(X)]$, then we have

$$
\begin{aligned}
\log p(x) &= \log \int_z p(x, z) \\
&= \log \int_z p(x, z) \frac{q_\phi(z|x)}{q_\phi(z|x)} \\
&= \log(\mathbb{E}_{q_\phi} \left[\frac{p(x, z)}{q_\phi(z|x)} \right]) \\
&\geq \mathbb{E}_{q_\phi} \left[\log(\mathrm{p(x,\ z)}) \right] - \mathbb{E}_{q_\phi} \left[\log(q_\phi(z|x)) \right]
\end{aligned}
$$

- Let $ELBO = \mathbb{E}_{q_\phi} \left[\log(q_\phi(z|x)) \right] - \mathbb{E}_{q_\phi} \left[\log(\mathrm{p(x,\ z)}) \right]$, then maximizing $ELBO$ is equivalent to minimizing $\mathrm{KL}(q(z|x; \phi)||p(z|x; \theta))$

Re-parametrization Trick

- Implementing VAE requires taking the derivatives with respect to the parameters of a stochastic variable z drawn from distribution $q_\phi(z|x)$
- Want the samples to deterministically depend on the parameters of the distribution in order to perform back-propagation
- For the normal distribution $q_\phi(z|x)$, we can sample z such that
  - $z = \mu + \sigma \odot \epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$



- Allows backpropagation with respect to θ through the objective (the ELBO) which is a function of samples of the latent variables z

Overview

Motivation

- VAE is of particular interest when one desires not only text generation, but also the capacity to infer meaningful latent codes from text
- Semantically meaningful latent codes can provide high-level guidance while generating sentences
  - E.g. the vocabulary could potentially be narrowed down if the input latent code corresponds to a certain topic

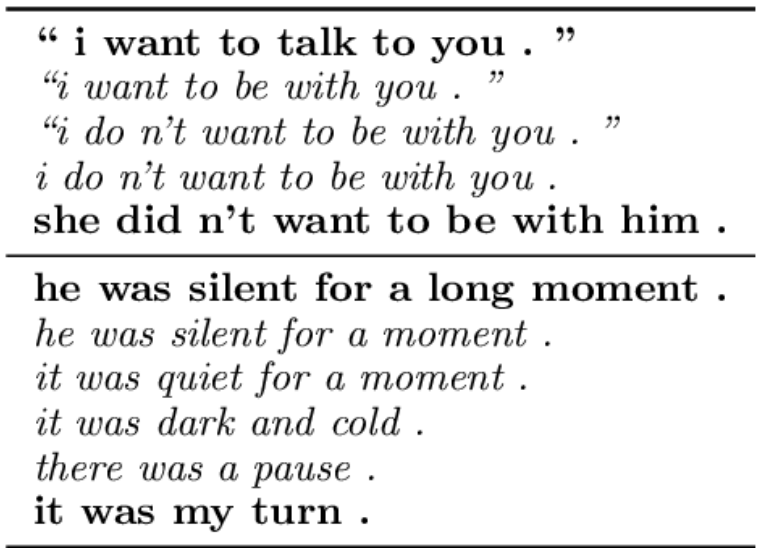
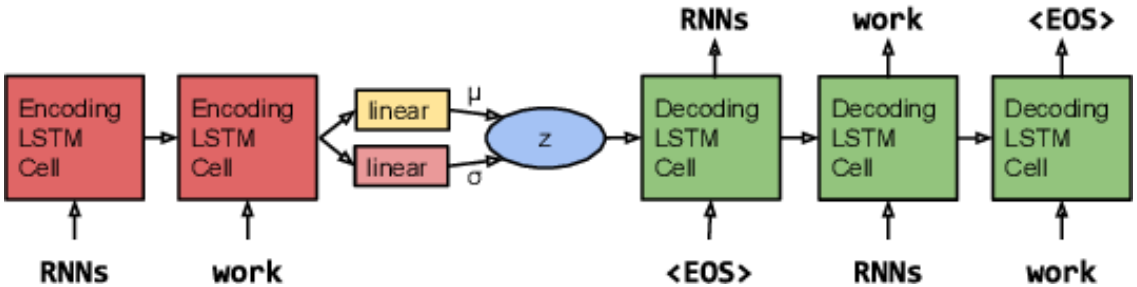


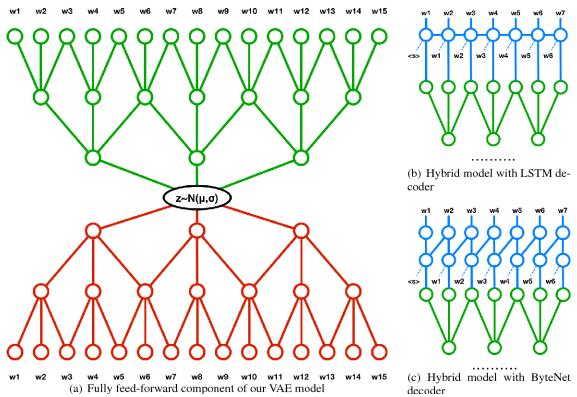
Figure: Paths between pairs of random points in VAE latent space

VAE for text-generation

- (Bowman et al. 2015) proposed using LSTM-RNN to form a sequence autoencoder with the Gaussian prior acting as a regularizer on the hidden code.



- (Semeniuta, Severyn, and Barth 2017) proposed a convolutional architecture similar to the models in the vision domain



- (Kikuchi et al. 2016) attempted to control output length for abstractive summarization

Challenges

- **Over-regularization**
  - Difficult for the simple Gaussian prior to indicate the semantic structure among sentences or text
  - Reduce the generative power of the decoder
- **Posterior collapse**
  - A strong decoder network $p(x|z)$ ignore the information from the latent code and merely depends on previous generated tokens for prediction
  - Signal from input x to posterior parameters is either too weak or too noisy

Overview

Introduction

- (Wang et al. 2019) proposed a topic-guided variational autoencoder (TGVAE) to generate semantically-meaningful sentences from different topics
- Attepmts to address the previous two challenges by:
  - Permitting text generation with designated topic by specifies a Gaussian mixture model (GMM) as the prior of the latent code
  - Householder flow is introduced to transform the mixture distribution into a flexible approximate posterior
- Comprised of two modules, a neural topic model (NTM) and a neural sequence model (NSM)

Neural Topic Model (NTM)

- Use a bag-of-words representation of the document, such that $d \in \mathbb{Z}_+^D$
- Let $\theta \in \mathcal{N}(0, I)$ and $t = g(\theta)$, where $g(\theta) = \mathrm{softmax}(\hat{W}\theta + \hat{b})$ is the function that maps sample θ to topic embedding t
- Let a_n be the topic assignment for word w_n, and $\beta_{a_n} \in \mathbb{R}^D$ represents the distribution over words for topic a_n. Then the marginal likelihood for document d, $p_\theta(d|\beta)$ is

$$
\begin{aligned}
p_\theta(d|\beta) &= \int_t p(t) \prod_n \sum_{a_n} p(w_n|\beta_{a_n}) p(a_n|t) dt \\
&= \int_t p(t) \prod_n p(w_n|\beta, t) dt \\
&= \int_t p(t) p(d|\beta, t) dt = \int_\theta p(\theta) p(d|\beta, \theta) d\theta
\end{aligned}
$$

- Note that $a_n \sim \mathrm{discrete}(t)$ so $p(w_n|\beta, t) = \sum_{a_n} p(w_n|\beta_{a_n}) p(a_n|t)$

Neural Sequence Model (NSM)

- Assumes each z is sampled from a topic dependent GMM using word distribution β and topic usage t from NTM, such that

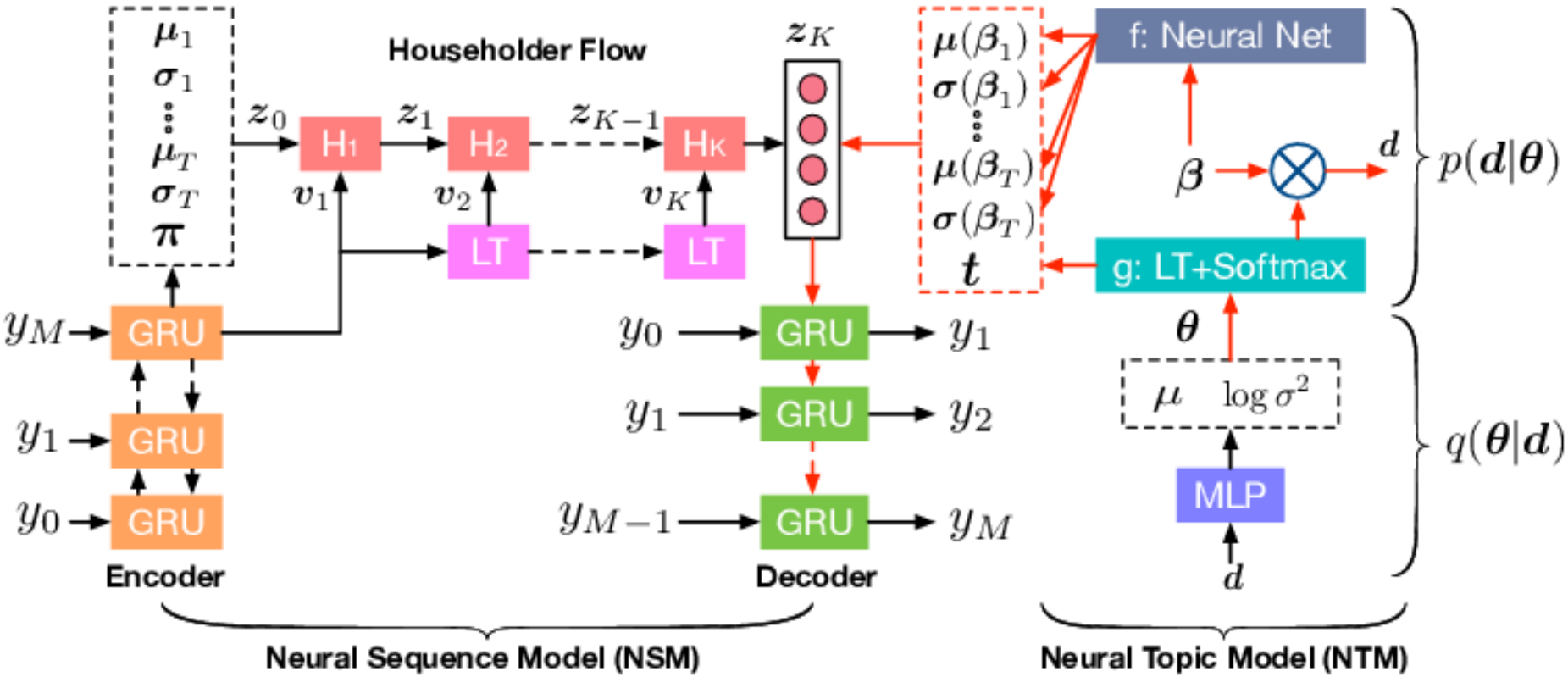
$$p(z|\beta, t) = \sum_{i=1}^{T} t_i \, \mathcal{N}(\mu(\beta_i), \sigma^2(\beta_i))$$
$$\mu(\beta_i) = f_\mu(\beta_i)$$
$$\sigma^2(\beta_i) = \mathrm{diag}(\exp(f_\sigma(\beta_i)))$$

- Both f_μ and f_σ are MLPs parameterized by W_μ and W_σ, respectively
- **Decoder**: The likelihood of a word sequence $p(\hat{x}|z)$ is

$$p(\hat{x}|z) = p(\hat{x}_1|z) \prod_{m=2} p(\hat{x}_m|h_m)$$

  - Where $h_m = \mathrm{GRU}(h_{m-1}, \hat{x}_{m-1}, z)$

Architecture



Inference

- The joint marginal likelihood can be written as

$$p(\hat{x}, d|\beta) = \int_{\theta} \int_{z} p(\theta) p(d|\beta, \theta)\ p(z|\beta, \theta) p(\hat{x}|z) d\theta dz$$

- Since direct optimization is intractable, use variation distribution of θ and z given by $q(\theta|d)$ and $q(z|x)$
- Variational objective ELBO given as

$$\begin{aligned}\mathcal{L} = &\mathbb{E}_{q(\theta|d)} \left[\log\, p(d|\theta\beta)\right] - \mathrm{KL}(q(\theta|d)||p(\theta)) \\ &+ \mathbb{E}_{q(z|x)} \left[\log\, p(x|z)\right] - \mathbb{E}_{q(\theta|d)} \left[\mathrm{KL}(q(z|x)||p(z|\beta, \theta))\right]\end{aligned}$$

  - Under the assumption that $q(\theta|d) \sim \mathcal{N}(\theta|g_{\mu}(d), \mathrm{diag}(\exp(g_{\sigma}(d))))$

Householder Transformation

- We know that any hyper plane can be defined by a orthogonal vector v such that $\|v\|_2 = 1$ (orthonormal)
- Let z_k be the reflection of vector z_{k-1} about the hyperplane defined by v_k, then we have $z_k = z_{k-1} - 2v_k^T z_{k-1} v_k$
- To capture the transformation as a matrix, we have

$$z_k = (I - 2\frac{v_k^T v_k}{\|v_k\|^2})z_{k-1} = H_k z_{k-1}$$

  - Where $H_k = (I - 2\frac{v_k^T v_k}{\|v_k\|^2})$ is called the Householder matrix
- For $k = 1, \ldots, K$, vector v_k is produced by a linear layer with input v_{k-1}, and $v_0 = h$ is the last hidden vector of the encoder RNN sentence y

Householder Flow for Approximate Posterior

- (Tomczak and Welling 2016) showed that Householder flow can transform a simple posterior to an arbitrarily complex distribution
- To construct posterior $p_K(z_K|x)$ from z_0, we can use the property of Jacobians of invertible functions and apply chain rule

$$\log q_K(z_K|x) = \log q_0(z_0|x) - \sum_{k=1}^{K} \log|\det \frac{\partial f_k}{\partial z_{k-1}}|$$

Here we have $q_O(z_0|x)$ is specified as a GMM, where

$$q_O(z_0|x) = \sum_{i=1}^{T} \pi_i(x)\, \mathcal{N}(\mu_i(x), \sigma^2(x))$$

  - Where mixture components π, μ, and σ are produced by the encoder network with respect to h

Extension to Text Summarization

- For text summarization, interested in modelling $p(y, d|x)$ where (x, y) denotes the document-summary pair, where

$$p(y, d|x) = \int_{\theta} \int_{z} p(\theta)p(d|\beta, \theta)p(z|\beta, \theta)p(y|x, z)d\theta dz$$

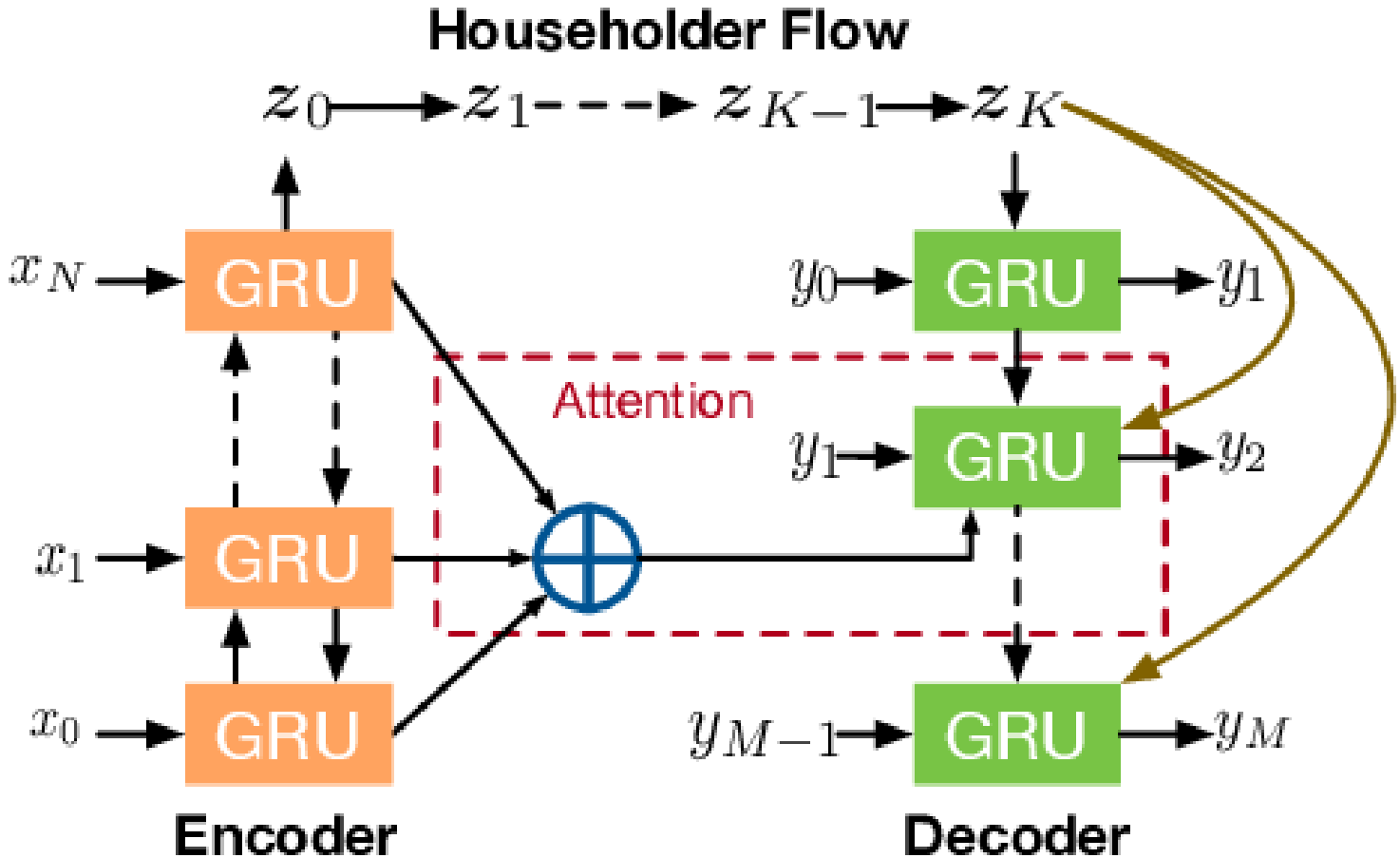
- The ELBO loss is reconstructed as

$$\mathcal{L} = \mathcal{L}_t + \mathbb{E}_{q(z|x)} \left[\log p(y|x, z)\right] - \mathbb{E}_{q(\theta|d)} \left[\mathbf{KL}(q(z|x)||p(z|\beta, \theta))\right]$$

  - Where $\mathcal{L}_t$ is the NTM loss as earlier

- The main difference compared to the unconditioned text generation being the usage of z_K to generate y at every time-step to provide topic guidance

Text Summarization Architecture



Experiments: Text Generation

Metric	Methods	APNEWS				IMDB				BNC			
		B-2	B-3	B-4	B-5	B-2	B-3	B-4	B-5	B-2	B-3	B-4	B-5
test-BLEU	VAE	0.564	0.278	0.192	0.122	0.597	0.315	0.219	0.147	0.479	0.266	0.169	0.117
	VAE+HF (K=1)	0.566	0.280	0.193	0.124	0.593	0.317	0.218	0.148	0.475	0.268	0.165	0.112
	VAE+HF (K=10)	0.570	0.279	0.195	0.123	0.610	0.322	0.221	0.147	0.483	0.270	0.169	0.110
	TGVAE (K=0, T=10)	0.582	0.320	0.203	0.125	0.627	0.362	0.223	0.159	0.517	0.282	0.181	0.115
	TGVAE (K=1, T=10)	0.581	0.326	0.202	0.124	0.623	0.358	0.224	0.160	0.519	0.282	0.182	0.118
	TGVAE (K=10, T=10)	0.584	0.327	0.202	0.126	0.621	0.357	0.223	0.159	0.518	0.283	0.173	0.119
	TGVAE (K=10, T=30)	0.627	0.335	0.207	0.131	**0.655**	0.369	**0.243**	**0.165**	0.528	**0.291**	0.182	0.119
	TGVAE (K=10, T=50)	**0.629**	**0.340**	**0.210**	**0.132**	0.652	**0.372**	0.239	0.160	**0.535**	0.290	**0.188**	**0.120**
self-BLEU	VAE	0.866	0.531	0.233	-	0.891	0.632	0.275	-	0.851	0.51	0.163	-
	VAE+HF (K=1)	0.865	0.533	0.241	-	0.899	0.641	0.278	-	0.854	0.515	0.163	-
	VAE+HF (K=10)	0.873	0.552	0.219	-	0.902	0.648	0.262	-	0.854	0.520	0.168	-
	TGVAE (K=0, T=10)	0.847	0.499	0.161	-	0.878	0.572	0.234	-	0.832	0.488	0.160	-
	TGVAE (K=1, T=10)	0.847	0.495	0.160	-	0.871	0.571	0.233	-	0.828	0.483	0.150	-
	TGVAE (K=10, T=10)	0.839	0.512	0.172	-	0.889	0.577	0.242	-	0.829	0.488	0.151	-
	TGVAE (K=10, T=30)	0.811	0.478	0.157	-	0.850	0.560	0.231	-	0.806	0.473	0.150	-
	TGVAE (K=10, T=50)	**0.808**	**0.476**	**0.150**	-	**0.842**	**0.559**	**0.227**	-	**0.793**	**0.469**	**0.150**	-

Generation on Given Topic

Data	Topic	Sentences
APNEWS	education	• the commission has approved a bill that would make state funding available for the city 's new school .
	animal	•the feline did n't survive fence hangars at the lake .
	crime	• the jury found the defense was not a <unk> , <unk> 's ruling and that the state 's highest court has been convicted of first-degree murder .
	weather	• forecasters say they 're still trying to see the national weather service watch for the latest forecast for friday evening .
	lottory	• she hopes the jackpot now exceeds $ 9 million .
	education+law	• an alabama law professor thomas said monday that the state's open court claims it takes an emotional matter about issuing child molesters based on religion.
	animal+medicine	• the study says the animal welfare department and others are not sure to make similar cases to the virus in the zoo.
IMDB	war	• after watching the movie , there is a great documentary about the war in the years of the israeli war .
	children	• the entire animation was great at times as to the readings of disney favorites .
	epsiode	• the show would have warranted for 25 episodes and it does help immediately .
	name	• she steals the other part where norma 's <unk> husband (crawford) (as at his part , sh*t for the road) .
	detective	• holmes shouted just to be as much as the movie 's last scene where there were <unk> pills to nab the <unk> .
	horror + negative	• the movie about a zombie is the worst movie i have ever seen.
	detective + children	• my favorite childhood is that rochester takes the character in jane's way, playing the one with hamlet.
BNC	medical	• here mistaking ' causes ' drugs as the problem although both economically ill patients arising from a local job will be in traumatic dangers .
	education	• he says the sale is given to five students ' award off : out at a laboratory after the three watts of the hours travelling in and chairman store the bank of the <unk> sutcliffe .
	religion	• schoolchildren will either go or back to church in his place every year in the savoy .
	entertainment	• 100 company and special lace with <unk> garland for tea our garden was filmed after a ceremony
	IT	• ibm also has shut all the big macs in the 60mhz ncube , represent on the acquisition and mips unix .
	environment + crime	• the earth's environmental protection agency said that the government was still being shut down by the police.
	education+entertainment	• the school is 55 and hosts one of a musician's theme charities festival.

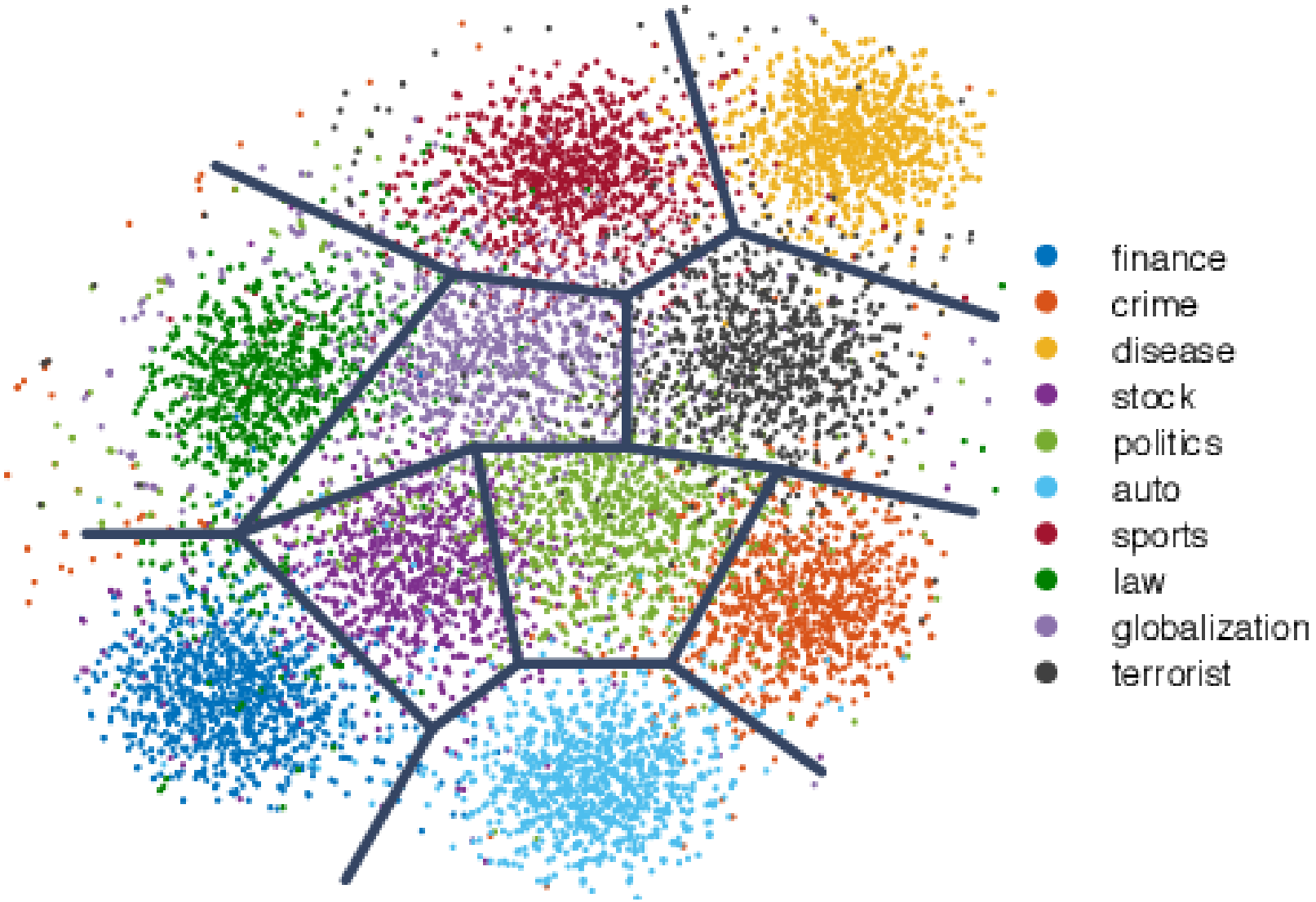
Summarization Results

Methods	GIGAWORDS			DUC-2004		
	RF-1	**RF-2**	**RF-L**	**RR-1**	**RR-2**	**RR-L**
ABS	29.55	11.32	26.42	26.55	7.06	22.05
ABS+	29.78	11.89	26.97	28.18	8.49	23.81
RAS-LSTM	32.55	14.70	30.03	28.97	8.26	24.06
RAS-Elman	33.78	15.97	31.15	27.41	7.69	23.06
lvt2k-lsent	32.67	15.59	30.64	28.35	9.46	24.59
lvt5k-lsent	35.30	16.64	32.62	28.61	9.42	25.24
ASC+FSC	34.17	15.94	31.92	26.73	8.39	23.88
Seq2Seq	34.03	15.93	31.68	28.39	9.26	24.83
Var-Seq2Seq	34.00	15.97	31.85	28.11	9.24	24.86
Var-Seq2Seq-HF (K=1)	34.04	15.98	31.84	28.18	9.27	24.84
Var-Seq2Seq-HF (K=10)	34.22	16.10	32.13	28.78	9.11	24.96
TGVAE (K=0, T=10)	35.34	16.68	32.69	28.99	9.21	24.89
TGVAE (K=1, T=10)	35.35	16.70	32.64	29.02	9.24	24.93
TGVAE (K=10, T=10)	35.40	16.77	32.71	29.07	9.32	25.17
TGVAE (K=10, T=30)	35.59	17.18	32.89	29.38	**9.60**	25.22
TGVAE (K=10, T=50)	**35.63**	**17.27**	**33.02**	**29.65**	9.55	**25.38**

Summarization Examples

Sample of Summaries
D: a court here thursday sentenced a ##-year-old man to ## years in jail after he admitted pummelling his baby son to death to silence him while watching television . **R**: man who killed baby to hear television better gets ## years. **Seq2Seq**: man sentenced to ## years after the son 's death **Ours**: a court sentenced a man ## years in jail
D: european stock markets advanced strongly thursday on some bargain-hunting and gains by wall street and japanese shares ahead of an expected hike in us interest rates , dealers said **R**: european stocks bounce back UNK UNK with closing levels **Seq2Seq**: european stocks advance ahead of us interest rate hike **Ours**: european stocks rise on bargain-hunting, dealer said friday
D: the democratic people 's republic of korea whitewashed south korea in the women 's team semi-finals at the world table tennis championships here on sunday **R**: dpr korea sails into women 's team final **Seq2Seq**: dpr korea whitewash south korea in women 's team final **Ours**: dpr korea beat south korea in table tennis worlds

Latent Space Visualization



Reference

Samuel R. Bowman et al. "Generating Sentences from a Continuous Space". In: *CoNLL*. 2015.

Yuta Kikuchi et al. "Controlling Output Length in Neural Encoder-Decoders". In: *EMNLP*. 2016.

Stanislau Semeniuta, Aliaksei Severyn, and Erhardt Barth. "A Hybrid Convolutional Variational Autoencoder for Text Generation". In: *EMNLP*. 2017.

Jakub M. Tomczak and Max Welling. "Improving Variational Auto-Encoders using Householder Flow". In: *ArXiv* abs/1611.09630 (2016).

Wenlin Wang et al. "Topic-Guided Variational Autoencoders for Text Generation". In: *NAACL-HLT*. 2019.

Overview

The End