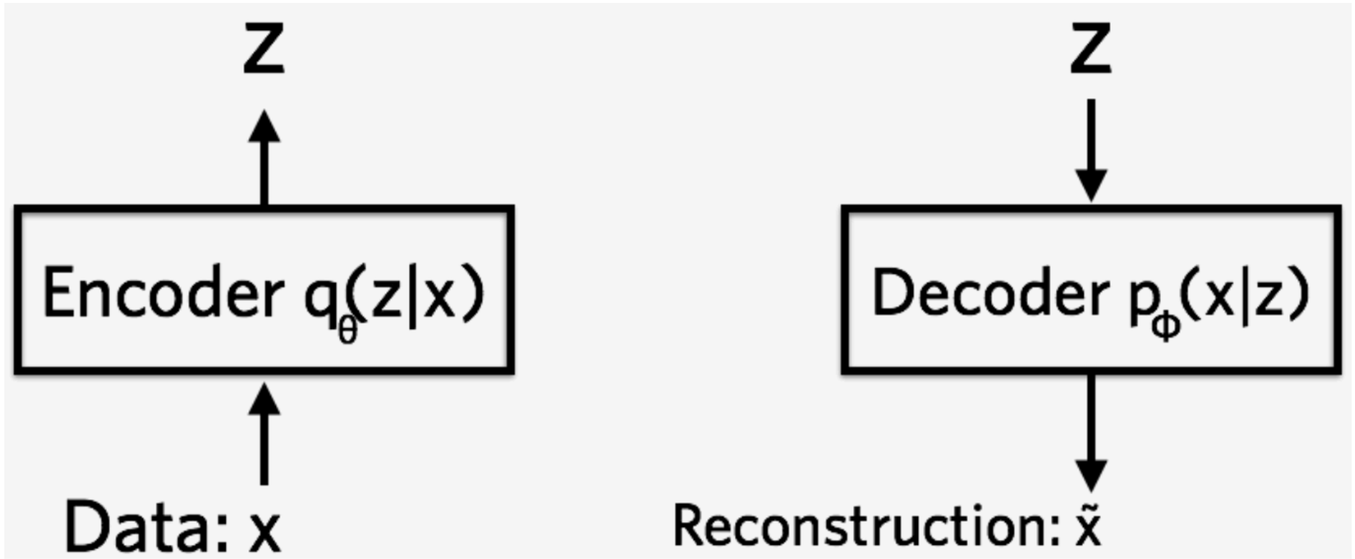
Variational Auto-encoding

NLP: Fall 2023

Anoop Sarkar

Encoder-Decoder neural nets

- An encoder q takes an input x and encodes it into a hidden representation z using some parameters θ. Encoder: $q_\theta(z \mid x)$
- An decoder p takes a hidden layer z and decodes it into an output $\tilde{x}$ using some parameters ϕ. Decoder: $\tilde{x} \sim p_\phi(x \mid z)$
- The output $\tilde{x}$ should be similar to but not necessarily identical to the true x



Autoencoder loss

- How much information is lost by going from x to z and then back to $\tilde{x}$?
- We measure the information loss by representing using z using reconstruction log-likelihood
- $\log p_\phi(x \mid z)$ measured in nats (bits are base 2, nats are base e)
- The loss function for an *variational* autoencoder is the negative log likelihood with a regularizer
- For single data point x_i we compute the above loss l_i.
- Total loss for the dataset: $\sum_i l_i$

Variational autoencoder loss

- Loss function l_i for datapoint x_i is
- $l_i(\theta, \phi) = -E_{z \sim q_\theta(z|x_i)}[\log p_\phi(x_i \mid z)] + KL(q_\theta(z \mid x_i) \| p(z))$
- First term is the expected negative log-likelihood of the data point x_i
- We want to place the most probability mass on the true output x_i
- Second term is the regularizer: the Kullback-Leibler divergence between the encoder distribution $q_\theta(z \mid x)$ and $p(z)$
- $p(z)$ is used to reward "good" values of the hidden representation that are efficient, can be sampled from easily and do not memorize the dataset.

Reparametrize z

- We want to use gradient descent to learn $q_\theta(z \mid x)$
- Need to take derivative of $p(z)$ wrt θ
- We reparametrize z
- $z = \mu + \sigma \circ \epsilon$ where $\epsilon \sim$ Normal$(0,1)$ and $\circ$ is element wise multiplication
- Now we can take derivatives of $p(z)$ wrt μ and σ
- Output of $q_\theta(z \mid x)$ is a vector of μ's and σ's

Variational autoencoder loss

- The regularizer term keeps the representation of z sufficiently diverse
- Without the regularizer, given large enough z the encoder-decoder would simply memorize the entire dataset
- Two different x_i and x_j that are actually very close to each other would end up learning very different z_i and z_j which defeats the purpose of modeling similarity between inputs.
- The regularizer would make sure z_i and z_j cannot get too far from each other unless x_i is very different from x_j
- The variational autoencoder (vae) is trained using gradient descent

Variational autoencoder loss

- Unfortunately, gradient descent requires computing distribution $q_\theta(z \mid x)$
- This is exponential because it is over all configurations of latent variable z
- Variational inference approximates this using a distribution $q_\lambda(z \mid x)$
- λ is the variational parameter which indexes a family of distributions
- If q is a normal distribution then λ_{x_i} would be the mean μ and variance σ^2 for each data point x_i
- $\lambda_{x_i} = (\mu_{x_i}, \sigma^2_{x_i})$

Tractable variational inference

- We want to measure how well does the variational distribution $q_\lambda(z \mid x)$ approximate the true distribution $q(z \mid x)$
- We use the KL divergence again: KL$(q_\lambda(z \mid x) \| q(z \mid x))$
- The optimal approximate distribution involves finding the optimal variational parameters λ
- $q_\lambda^*(z \mid x) = \arg\min_\lambda KL(q_\lambda(z \mid x) \| q(z \mid x))$
- Unfortunately, this is still intractable

Tractable variational inference

- Define ELBO(λ) the Evidence Lower BOund of λ
- ELBO(λ) = $E_{z\sim q_\lambda}[\log p(x \mid z)] - E_{z\sim q_\lambda}[\log q_\lambda(z \mid x)]$
- Minimizing KL($q_\lambda \| q$) wrt λ is equivalent to maximizing ELBO(λ)
- For each data point x_i
- ELBO$_i(\lambda) = E_{z\sim q_\lambda(z|x_i)}[\log p_\phi(x_i \mid z)] - \mathrm{KL}(q_\lambda(z \mid x_i) \| p(z)]$
- Maximizing ELBO$_i(\lambda)$ is equivalent to minimizing $l_i(\theta, \phi) = -E_{z\sim q_\theta(z|x_i)}[\log p_\phi(x_i \mid z)] + KL(q_\theta(z \mid x_i) \| p(z))$

Applications: Image generation

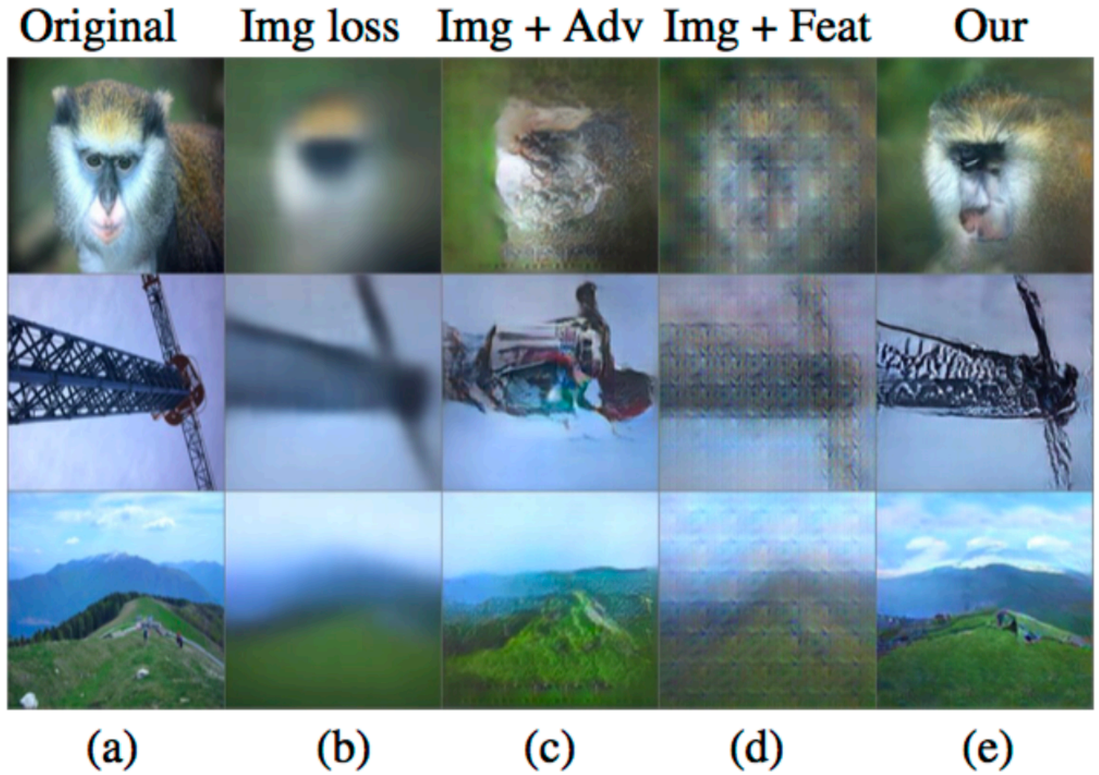
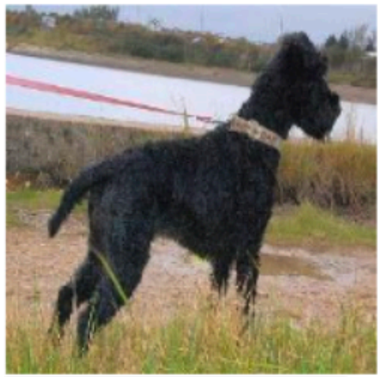


Figure 1: Reconstructions from AlexNet FC6 with different components of the loss.

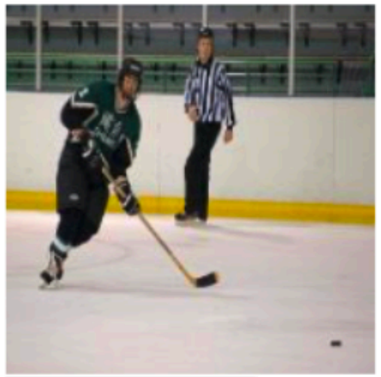
A. Dosovitskiy and T. Brox. Generating images with perceptual similarity metrics based on deep networks. *arXiv preprint arXiv*:1602.02644, 2016.

Applications: caption generation

a man with a snowboard next to a man with glasses



a big black dog standing on the grass



a player is holding a hockey stick

a desk with a keyboard

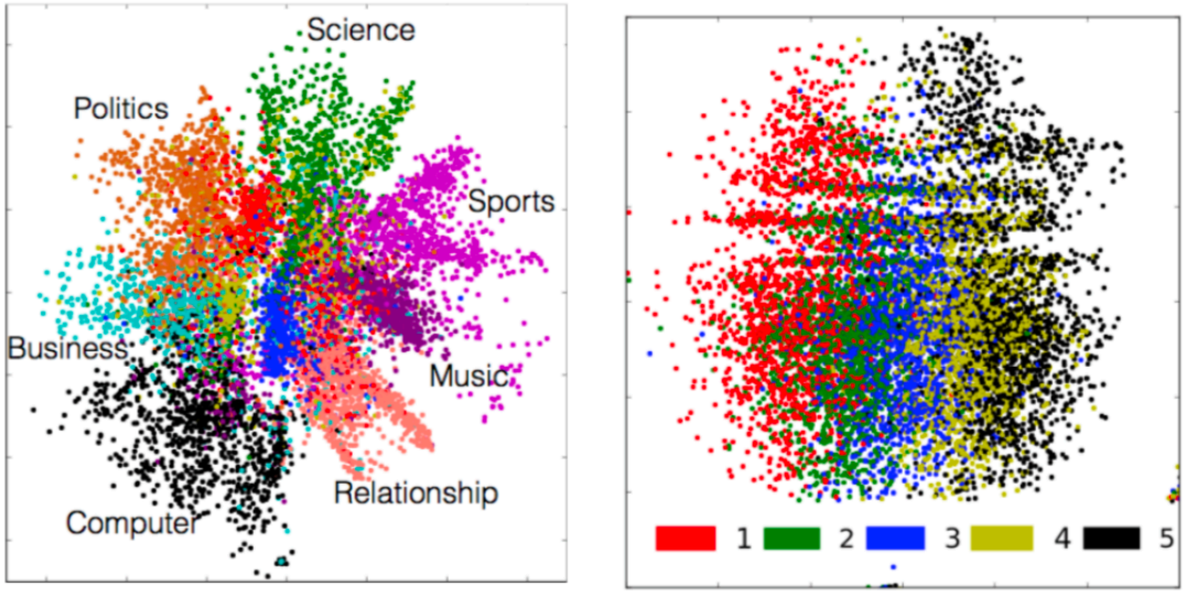
a man is standing next to a brown horse

a box full of apples and oranges

Figure 2: Examples of generated caption from unseen images on the validation dataset of ImageNet.

Y. Pu, Z. Gan, R. Henao, X. Yuan, C. Li, A. Stevens, and L. Carin. Variational autoencoder for deep learning of images, labels and captions. In NIPS, 2016.

Applications: document clustering



(a) Yahoo (b) Yelp

Figure 3: Visualizations of learned latent representations.

Z. Yang, Z. Hu, R. Salakhutdinov, and T. Berg-Kirkpatrick. Improved variational autoencoders for text modeling using dilated convolutions. In *Proceedings of The 34rd International Conference on Machine Learning*, 2017.

Applications: sign clustering

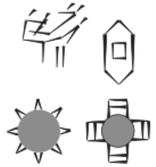
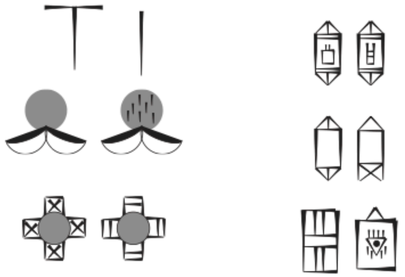
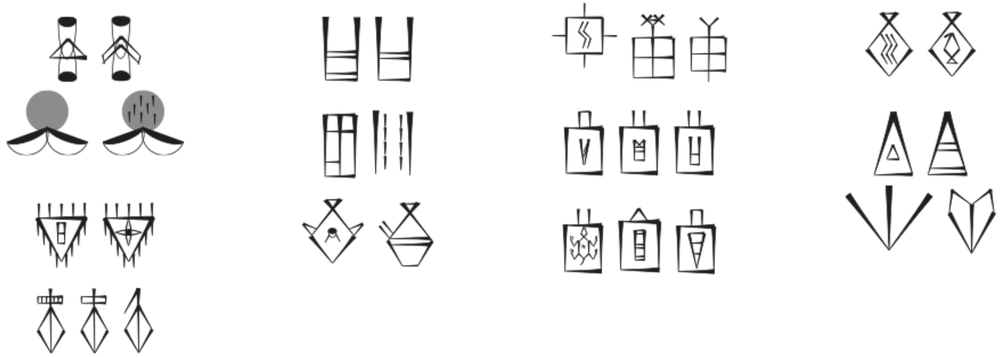
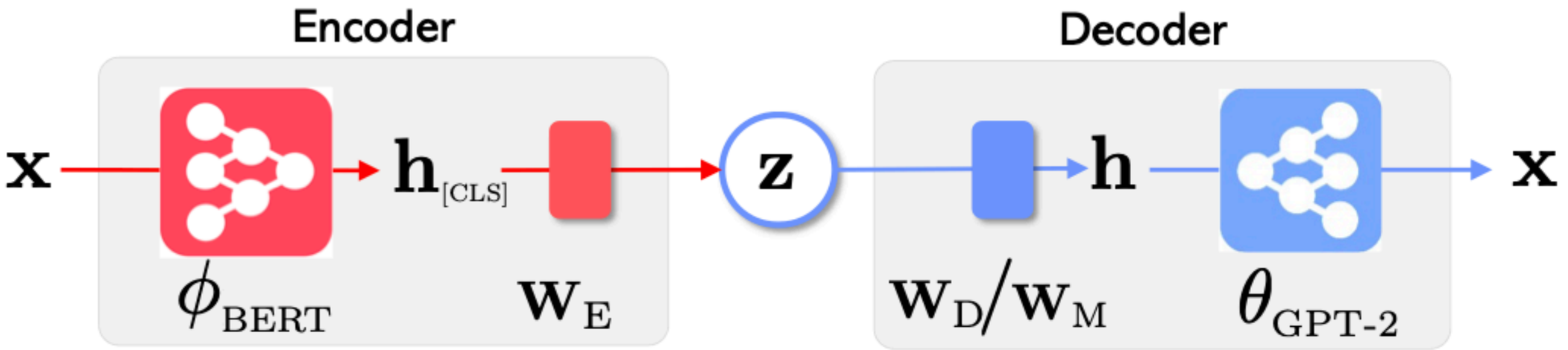
VAE+Neighbor	VAE+LSTM	VAE+Transformer
		

Table 3: Pairs/triplets of character images which have distinct labels in the working signlist, but which our models merge into single clusters.

OPTIMUS: Organizing Sentences via Pre-trained Modeling of a Latent Space

Chunyuan Li, Xiang Gao, Yuan Li, Baolin Peng, Xiujun Li, Yizhe Zhang, Jianfeng Gao
Microsoft Research, Redmond
{chunyl, xiag, v-liyua, bapeng, xiul, yizzhang, jfgao}@microsoft.com

Optimus



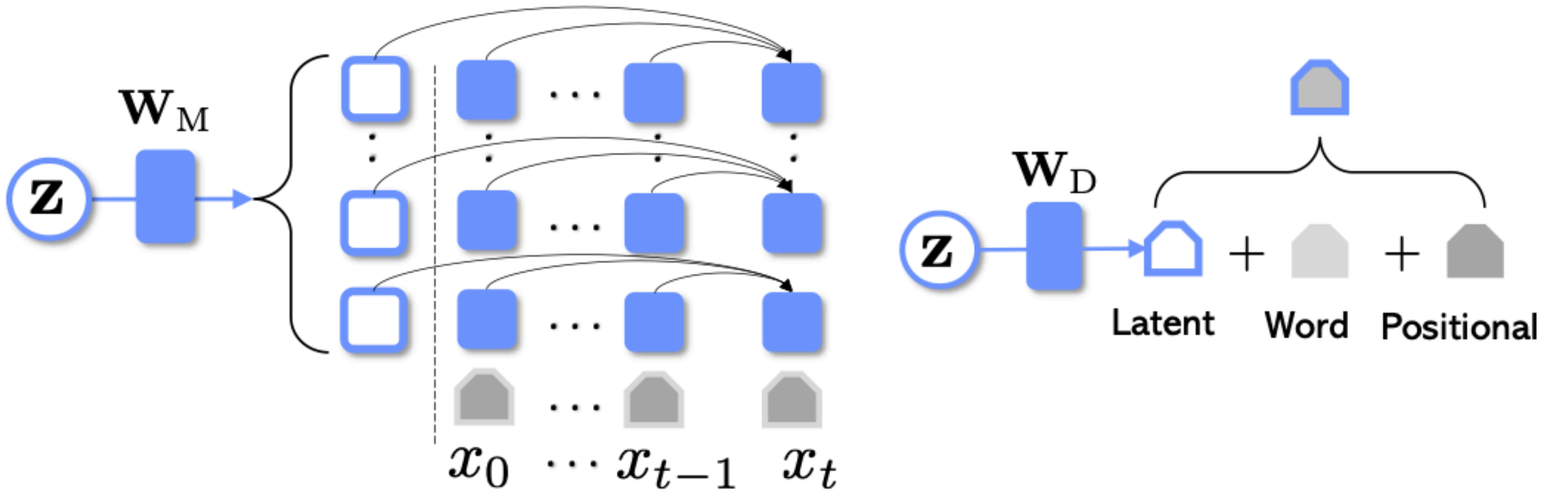
Latent variable

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) = \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(x_t|x_{<t}, \boldsymbol{z}).$$

all tokens before t

Latent vector injection

- The encoder is a BERT model with sentence embedding [CLS]: $h_{\text{[CLS]}} \in \mathbb{R}^H$
- Latent representation $z \in \mathbb{R}^P$ and $W_E \in \mathbb{R}^{P \times H}$: $z = W_E h_{\text{[CLS]}}$
- Two ways to use z in GPT2 decoding (where GPT2 has L layers):
  - **Memory**:
    - $h_{\text{Mem}} = W_M z$ where $W_M \in \mathbb{R}^{LH \times P}$ is a weight matrix
    - $h_{\text{Mem}} \in \mathbb{R}^{LH}$ is divided into L vectors of size H
  - **Embedding**:
    - New embedding representation $h'_{\text{Emb}} = h_{\text{Emb}} + W_D z$ where $W_D \in \mathbb{R}^{H \times P}$



(a) Memory (b) Embedding

Figure 2: Illustration of two schemes to inject latent vector. (a) Memory: x_t attends both $x_{<t}$ and $\boldsymbol{h}_{\texttt{Mem}}$; (b) Embedding: latent embedding is added into old embeddings to construct new token embedding $\boldsymbol{h}'_{\texttt{Emb}}$.

Language modeling results vs. GPT2

	Dataset	PTB			YELP			YAHOO			SNLI		
		LM	Repr.		LM	Repr.		LM	Repr.		LM	Repr.	
	Method	PPL ↓	MI ↑	AU ↑	PPL ↓	MI ↑	AU ↑	PPL ↓	MI ↑	AU ↑	PPL ↓	MI ↑	AU ↑
OPTIMUS	$\lambda=0.05$	**23.58**	3.78	32	**21.99**	2.54	32	22.34	5.34	32	**13.47**	3.49	32
	$\lambda=0.10$	23.66	4.29	32	21.99	2.87	32	22.56	5.80	32	13.48	4.65	32
	$\lambda=0.25$	24.34	5.98	32	22.20	5.31	32	22.63	7.42	32	14.08	7.22	32
	$\lambda=0.50$	26.69	7.64	32	22.79	7.67	32	23.11	8.85	32	16.67	8.89	32
	$\lambda=1.00$	35.53	**8.18**	32	24.59	**9.13**	32	24.92	**9.18**	32	29.63	**9.20**	32
	GPT-2	24.23	-	-	23.40	-	-	22.00	-	-	19.68	-	-
	LSTM-LM	100.47	-	-	42.60	-	-	60.75	-	-	21.44	-	-
	LSTM-AE	-	8.22	32	-	9.24	32	-	9.26	32	-	9.18	32

Optimus

Source $\boldsymbol{x}_A$	**Target** $\boldsymbol{x}_B$
a girl makes a silly face	two soccer players are playing soccer
Input $\boldsymbol{x}_C$	**Output** $\boldsymbol{x}_D$
• a girl poses for a picture	• two soccer players are at a soccer game.
• a girl in a blue shirt is taking pictures of a microscope	• two football players in blue uniforms are at a field hockey game
• a woman with a red scarf looks at the stars	• two men in white uniforms are field hockey players
• a boy is taking a bath	• two baseball players are at the baseball diamond
• a little boy is eating a bowl of soup	• two men are in baseball practice

Table 2: Sentence transfer via arithmetic $\boldsymbol{z}_D = \boldsymbol{z}_B - \boldsymbol{z}_A + \boldsymbol{z}_C$. The output sentences are in blue.

Finetuning Pretrained Transformers into Variational Autoencoders

Seongmin Park **Jihwa Lee**
ActionPower
Seoul, Republic of Korea
{seongmin.park, jihwa.lee}@actionpower.kr

Posterior Collapse

- Text variational autoencoders suffer from *posterior collapse*
- The model learns to ignore the latent variable from the encoder especially when the decoder is a large language model and very expressive
- Optimus mitigates posterior collapse using massive pre-training
- This paper presents a simple two-phrase training scheme to convert an encoder-decoder Transformer into a VAE by using only *fine-tuning*
- Many proposed techniques to avoid posterior collapse do not work on large language models

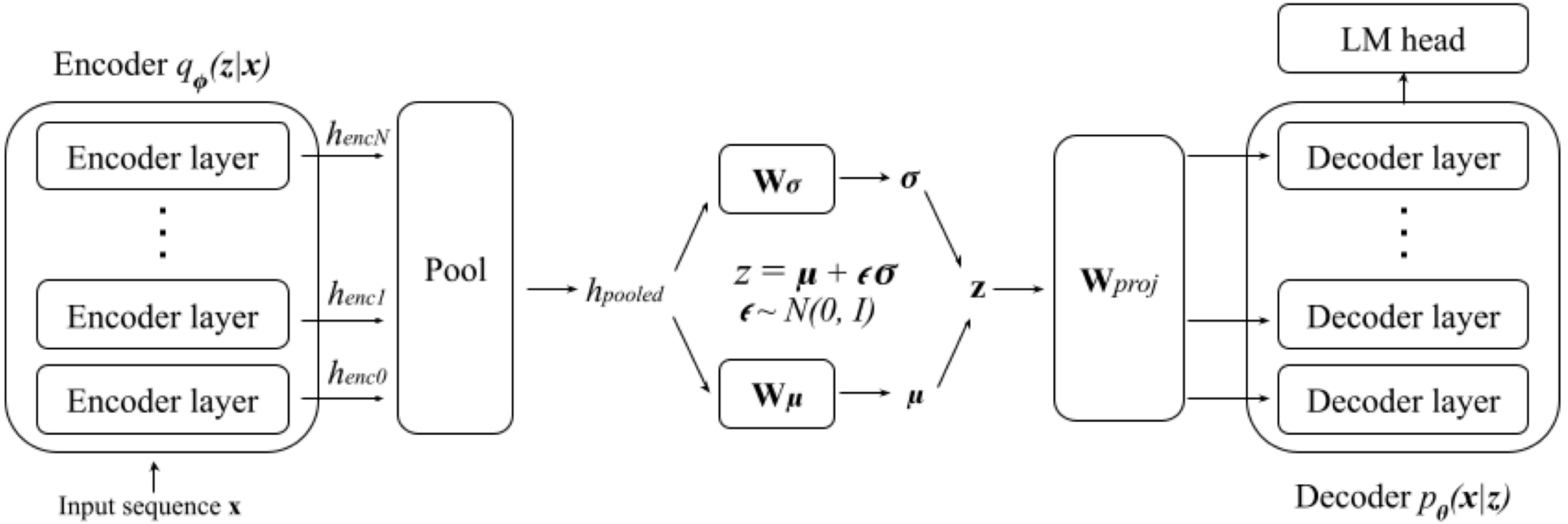


Figure 1: Transformer VAE architecture. A "bottleneck" step (W_σ and W_μ) is placed between the encoder and the decoder of T5. Latent information from pooled encoder hidden states is captured in the bottleneck layer before being passed to the decoder. The network is optimized against regularization loss in the bottleneck and reconstruction loss at the decoder.

Model	PPL↓	KL	-ELBO↓	MI↑	AU↑
Optimus ($\lambda = 0.5$) (Li et al., 2020)	23.11	17.45	301.21	8.85	32
GPT-2 (Radford et al., 2019)	22.00	-	-	-	-
Encoder pretraining ($\lambda = 3$) (Li et al., 2019)	59.24	7.44	328.73	6.41	32
Ours (Max pool)	20.90	0.21	343.02	0.04	0
Ours (Max pool + Denoise)	30.13	41.49	301.86	1.32	24
Ours (Max pool + Denoise + KLT)	60.44	119.89	223.69	4.73	29
Ours (Max pool + Denoise + KLT + Deep)	54.40	155.50	140.57	5.43	28

Table 1: Phase 2 results on Yahoo. Due to space constraints, we report experimental results on other datasets in the appendix. Results on baselines are quoted from Li et al. (2020) and (Li et al., 2019). KLT denotes KL thresholding with $\lambda = 3$. Our models are finetuned from a pretrained 6-layer T5, except the deep variant with 12 layers.

CTRL: A Conditional Transformer Language Model for Controllable Generation

Nitish Shirish Keskar[*], **Bryan McCann**[*], **Lav R. Varshney**, **Caiming Xiong**, **Richard Socher**
Salesforce Research[†]

$$p(x|c) = \prod_{i=1}^{n} p(x_i|x_{<i}, c) \qquad \mathcal{L}(D) = -\sum_{k=1}^{|D|} \log p_\theta(x_i^k|x_{<i}^k, c^k)$$

Control Code	Description
Wikipedia	English Wikipedia
Books	Books from Project Gutenberg
Reviews	Amazon Reviews data (McAuley et al., 2015)
Links	OpenWebText (See Sec. 3.2)
Translation	WMT translation date (Barrault et al., 2019)
News	News articles from CNN/DailyMail Nallapati et al. (2016), New York Times and Newsroom (Grusky et al., 2018)
multilingual	Wikipedias in German, Spanish and French
Questions	(Questions and answers only) MRQA shared task (See Section 3.1)
Explain	(Only main post) (Fan et al., 2019)

Hallucination vs Truthfulness

- Temperature controlled stochastic sampling is used to generate text from a trained LM
- $T = 1$ is greedy while $T \to \infty$ makes the distribution more uniform
- Clip to top-k or use nucleus sampling to clip to all tokens above some probability
- Penalized sampling
- Near greedy sampling by penalizing repetitions
- Given a list of generated tokens g we compute p_i

$$p_i = \frac{\exp(x_i/T)}{\sum_j \exp(x_j/T)}.$$

$$p_i = \frac{\exp(x_i/(T \cdot I(i \in g))}{\sum_j \exp(x_j/(T \cdot I(j \in g))}$$

$$I(c) = \theta \text{ if c is True else } 1$$

Wikipedia *Anarchism is* a political philosophy that advocates the abolition of all forms of hierarchy and domination, including capitalism, patriarchy, racism, sexism, heterosexism and other oppressive social structures.\n\nThe term ”anarchism” was coined by Pierre-Joseph Proudhon in his book ”The Poverty of Philosophy” (1844). It has been used to describe various movements within anarchism since then. In its modern sense, it refers to an ideology or movement advocating for social, political, economic and/or cultural change through direct action against existing institutions and practices.

Books *Anarchism is* the \n only true and practical form of Socialism. It has been said that Socialism \n will never be established until it can be proved to work in practice. The \n same may be said, with equal truth, about Anarchism. But if we are not able \n to prove its practicability at present, then let us wait for a better time \n when our present conditions shall have become so much more favorable than \n they now are, that we shall no longer need any such proof. Let us, then, \n patiently await this time, which I believe will come sooner or later. In \n the meantime, as long as there remains one man who

Horror *A knife* handle pulled through the open hole in the front. I jumped when the knife hit.\n\nEyes widened in horror. Her scream was the only sound I heard besides her sobs.\n\nThe spider touched her feet as it started to dig into the top of her arch. The creature's tears began to flow.\n\nThe spider looked up to her and looked back at me with eyes filled with burning tears. My heart started to race. . .

Reviews *A knife* is a tool and this one does the job well.\n\nRating: 4.0\n\nI bought these for my husband who has been using them to cut up his own meat since he got them. He says they are very sharp so be careful when you use them, but that doesn't seem like much of an issue because he's used it on everything from chicken breasts to beef tenderloin. . .

`Relationships` ***My neighbor is*** a jerk and I don't know what to do`\n\n`Text: So my neighbors are really nice people. They have been for years. We live in an apartment complex so we get along great.`\n\n`But recently they started acting like jerks. . .

`Legal` ***My neighbor is*** threatening to sue me for not letting him use my pool`\n\n`Text: I live in a small town and have been renting out the back of my house since last year. The owner has had issues with people using it as well but he never really took any action against them. Last week someone from his property management company came by asking if they could come over because there was an issue with some kids playing on it. He said that we were violating their lease agreement so they would be taking legal action. . .

Links https://www.cnn.com/2014/09/20/us-president-meets-british-pm JUST WATCHED\n\nObama meets with British PM\n\nMUST WATCH\n\nStory highlights\n\nPresident Barack Obama met with Britain's Prime Minister David Cameron

Links https://www.cnn.com/2018/09/20/us-president-meets-british-pm JUST WATCHED\n\nTrump and May meet for first time\n\nMUST WATCH\n\nWashington (CNN) President Donald Trump, who has been criticized by some in the UK over his decision to leave the European Union, met with British Prime Minister Theresa May, a White House official said on Thursday.

Links https://www.cnn.com/09/20/2018/star-spotted JUST WATCHED\n\nStar spotted: The world's most elusive star\n\nMUST WATCH\n\nStory highlights\n\nA new study suggests that the Milky Way is home to a supermassive black hole\n\nIt may be one of our closest galactic neighbors\n\n(CNN) Astronomers have discovered what they believe could be an extremely rare astronomical event: An object so bright it can only be seen with powerful telescopes.\n\nThis object, dubbed SN2014J, was first detected by NASA's Spitzer Space Telescope in 2014. . . .

Questions *Q: What is the capital of India?* A: New Delhi\n\nQ: Which city in India has a large population of people from Gujarat?\n\nA: Ahmedabad\n\nQ: How many languages are spoken by Indians living abroad?\n\nA: more than 100. . .

Translation English : *We release a new model for coherent language generation* ; French : Nous publions un nouveau modéle de génération cohérente du langage\n\nEnglish : The European Parliament has just adopted the report on the proposal for a regulation of the European Parliament and of the Council establishing an action programme to promote multilingualism in Europe. ; French : Le Parlement européen vient d'adopter le rapport sur la proposition de règlement du Parlement et du Conseil établissant un programme-cadre pour promouvoir le pluriling uisme dans l'Union européenne. . .

Table 4: More complex templatized control codes are used for task-specific generation.