Boosting Summarization with Normalizing Flows and Aggressive Training

Yu Yang & Xiaotong Shen

School of Statistics
University of Minnesota

Why Variational Models?

Issues with most Transformer-based models

- exposure bias
- lack of model variability
- insufficient capturing of semantic information

Variational Models

Introduce uncertainty by learning a probability distribution over the latent variables.

- smoother output spaces, reducing the exposure bias.
- diverse summaries
- better semantic capturing

Challenges

1. Simple Gaussian inflexible to capture the latent intricacies
2. Posterior collapse

Proposal

Normalizing Flows + Training Techniques

Normalizing Flows (NF)

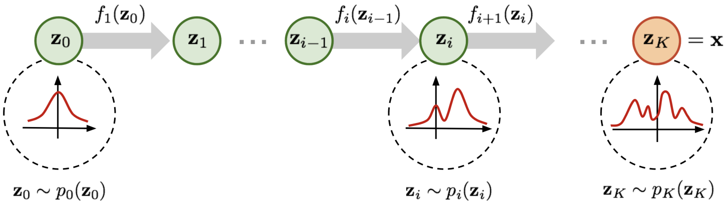


Figure 1: Illustration of a Normalizing Flow Model[1]

$$z_i = f_i\left(z_{i-1}\right) \Leftrightarrow z_{i-1} = f_i^{-1}\left(z_i\right)$$

$$z_K = f_K \circ f_{K-1} \circ \cdots \circ f_1\left(z_0\right)$$

$$\log p_K\left(z_K\right) = \log p_0\left(z_0\right) - \sum_{i=1}^{K} \log \left|\det \frac{df_i}{dz_{i-1}}\right|$$

[1]Image source: https://lilianweng.github.io/posts/2018-10-13-flow-models/.

Variational Encoder Decoder (VED)

Given x, we assume there exists a latent variable $z \sim p(z|x)$ and that $y \sim p(y|x,z)$.

$$p(\mathrm{y} \mid \mathrm{x}) = \int p(\mathrm{z} \mid \mathrm{x}) p(\mathrm{y} \mid \mathrm{x}, \mathrm{z}) dz$$

- a variational posterior $q_\psi(z|x,y) \to p(z|x,y)$
- a model of $p_\theta(y|x,z)$

Model: FlowSUM

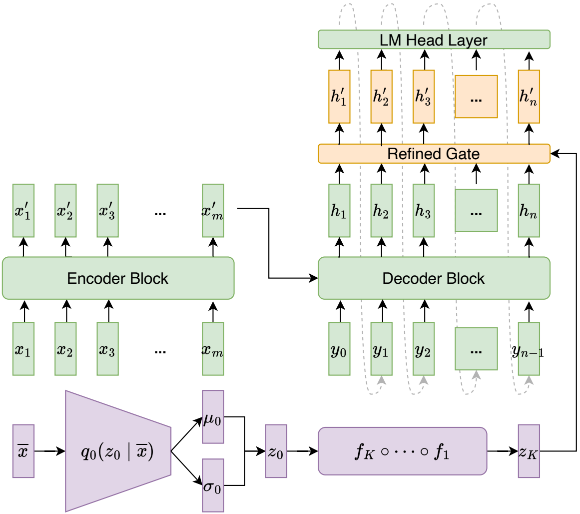
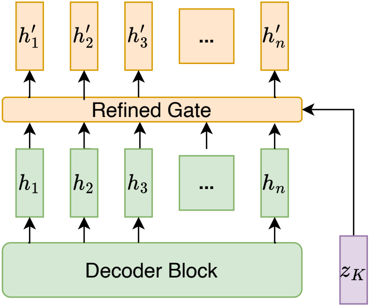
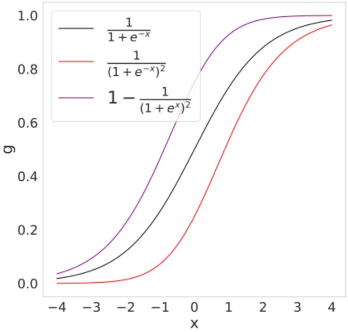


Figure 2: FlowSUM Model Architecture

Refined Gate Mechanism





Mitigate the saturation problem and allows for better gradient flow (Gu et al., 2020).

$$\begin{cases} z'_K = W^z z_K \\ f_j = \delta\left(W^f\,[h_j; z'_K]\right) \\ h'_j = (1-f_j)\cdot h_j + f_j \cdot z'_K \end{cases} \Rightarrow \begin{cases} z'_K = W^z z_K \\ f_j = \delta\left(W^f\,[h_j; z'_K]\right) \\ r_j = \delta\left(W^r\,[h_j; z'_K]\right) \\ g_j = f_j + f_j(1-f_j)(2r_j - 1) \\ h'_j = (1-g_j)\cdot h_j + g_j \cdot z'_K \end{cases}$$

Objective Function

$$\begin{aligned}\mathrm{ELBO}_{\text{NF-VED}} =& \mathbb{E}_{q_0(z_0)}\Big[\log p\left(y \mid x, z_K\right)+\log p\left(z_K \mid x\right)\Big] \\ & -\mathbb{E}_{q_0(z_0)}\left[\log q_0\left(z_0\right)-\sum_{k=1}^{K} \log \left|\det J_{f_k}\left(z_{k-1}\right)\right|\right],\end{aligned} \tag{1}$$

$$\begin{aligned}\mathcal{L} =& -\sum_{j=1}^{n} \log p\left(y_j \mid \{x_i\}_{i=1}^{m}, z_K, y_{<j}\right)-\log p\left(z_K \mid x\right) \\ & +\log q_0\left(z_0\right)-\sum_{k=1}^{K} \log \left|\det J_{f_k}\left(z_{k-1}\right)\right|\end{aligned} \tag{2}$$

where q_0 is the probability density function for z_0.

Mitigating Posterior Collapse

Algorithm 1 Controlled Alternate Aggressive Training (CAAT)

Input: number of aggressive training steps n_{agg}; maximum number of training steps n_{max}; number of alternating steps n_{alt}.

$\boldsymbol{\theta}, \boldsymbol{\psi} \leftarrow$ Initialize encoder-decoder parameters and variational parameters respectively
for $i = 1, 2, \cdots, n_{agg}$ **do**
    $\mathbf{X} \leftarrow$ Random data minibatch
    if $i \mod n_{alt} = 0$ **then**
        Compute $\boldsymbol{g}_{\boldsymbol{\theta},\boldsymbol{\psi}} \leftarrow \nabla_{\boldsymbol{\psi},\boldsymbol{\theta}} \mathcal{L}(\mathbf{X}; \boldsymbol{\theta}, \boldsymbol{\psi})$
        Update $\boldsymbol{\theta}, \boldsymbol{\psi}$ using gradients $\boldsymbol{g}_{\boldsymbol{\theta},\boldsymbol{\psi}}$
    else
        Compute $\boldsymbol{g}_{\psi} \leftarrow \nabla_{\boldsymbol{\psi}} \mathcal{L}(\mathbf{X}; \boldsymbol{\theta}, \boldsymbol{\psi})$
        Update $\boldsymbol{\psi}$ using graidents $\boldsymbol{g}_{\boldsymbol{\psi}}$
for $i = n_{agg}, n_{agg} + 1, \cdots, n_{max}$ **do**
    $\mathbf{X} \leftarrow$ Random data minibatch
    Compute $\boldsymbol{g}_{\boldsymbol{\theta},\boldsymbol{\psi}} \leftarrow \nabla_{\boldsymbol{\psi},\boldsymbol{\theta}} \mathcal{L}(\mathbf{X}; \boldsymbol{\theta}, \boldsymbol{\psi})$
    Update $\boldsymbol{\theta}, \boldsymbol{\psi}$ using gradients $\boldsymbol{g}_{\boldsymbol{\theta},\boldsymbol{\psi}}$
    if early stopping criterion is met **then**
        break

Datasets

- **CNN/Daily Mail**
  - 312,085 online news articles paired with multi-sentence summaries.
- **Multi-News**
  - 56k pairs of multi-document news articles and multi-sentence summaries.
- **arXiv, PubMed**
  - two scientific paper document datasets from arXiv.org (113k) and PubMed (215k).
- **XSum**
  - 227k BBC articles, each summarized in a single sentence.
- **SAMSum**
  - 16k conversations annotated with summaries by linguists.

Models in Comparison

Deterministic Models

- PG+Cov (See et al., 2017)
- BERT2BERT (Rothe et al., 2020)
- BERTSUM (Liu and Lapata, 2019)
- BART (Lewis et al., 2020)
- PEGASUS (J. Zhang et al., 2020)

Variational Models

- VHTM (X. Fu et al., 2020)
- TAS (Zheng et al., 2020)
- PEGASUS+Flow-NTM (Nguyen et al., 2021)

Proposed Models

- VEDSUM
- FlowSUM

Normalizing Flows Types

- Planar flow (Rezende and Mohamed, 2015)
- Radial flow (Rezende and Mohamed, 2015)
- Sylvester flow (van den Berg et al., 2018)
- RealNVP (Dinh et al., 2017)
- Inverse Autoregressive flow (IAF) (Kingma et al., 2016)
- Rational-Linear Neural Splines flows (RLNSF) (Dolatabadi et al., 2020)
- Rational-Quadratic Neural Spline Flows (RQNSF) (Durkan et al., 2019)

Evaluation Metrics

- ROUGE scores (Lin, 2004)
  - ROUGE-1: overlap of unigrams
  - ROUGE-2: overlap of bigrams
  - ROUGE-L: overlap of the longest common sequence
- BERTScore (T. Zhang et al., 2020)
- Repetition measure rep-w (Z. Fu et al., 2021)
  - the proportions of words that occur in the previous w words

Quantitative Analysis I

Model	ROUGE ↑		
	1	2	L
PG+Cov	39.53	17.28	36.38
BERT2BERT	41.28	18.69	38.09
BERTSUM	42.13	19.60	39.18
BART	44.16	21.28	40.90
PEGASUS	44.17	21.47	41.11
VHTM	40.57	18.05	37.18
TAS	44.38	21.19	41.33
PEGASUS+NTM	44.52	**21.95**	41.39
VEDSUM (BERT2BERT)	40.89	18.28	37.95
FlowSUM (BERT2BERT)	41.51	18.81	38.56
VEDSUM (BART)	44.36	21.09	41.37
FlowSUM (BART)	**44.64**	21.36	**41.65**

Table 1: Comparison with baselines on CNN/DM.

Quantitative Analysis II

Model	ROUGE 1/2/L ↑	BERTScore ↑	rep-w ↓	Length
CNN/DM				
BART	44.16/21.28/40.90	89.40	**8.31**	84.11
VEDSUM	44.34/21.09/41.37	89.20	8.43	88.63
FlowSUM	**44.64/21.36/41.65**	**89.46**	8.43	92.24
Multi-News				
BART	42.56/15.34/36.67	86.69	**9.76**	133.42
VEDSUM	43.91/16.68/38.10	87.04	9.95	128.79
FlowSUM	**44.42/17.01/38.36**	**87.09**	9.91	128.87
arXiv				
BART	42.55/15.92/37.89	85.35	17.23	130.68
VEDSUM	43.05/**16.34**/38.26	85.44	16.63	130.92
FlowSUM	**43.11**/16.26/**38.31**	**85.45**	**16.55**	132.88
PubMed				
BART	41.57/16.72/36.94	84.65	13.26	136.10
VEDSUM	44.21/19.20/39.32	85.07	12.76	138.70
FlowSUM	**44.55/19.50/39.59**	**85.16**	**12.59**	138.09

Table 2: Comparison of BART, VEDSUM, and FlowSUM on long-summary benchmarks.

Quantitative Analysis III

Model	ROUGE 1/2/L ↑	BERTScore ↑	rep-w ↓	Length
		XSum		
BART	45.14/**22.27**/**37.25**	**92.16**	**4.63**	25.54
VEDSUM	43.62/20.27/35.06	91.75	5.96	31.22
FlowSUM	**45.26**/22.12/37.00	92.13	4.95	28.71
		SAMSum		
BART	**53.16**/28.19/**49.03**	**92.68**	6.71	30.00
VEDSUM	51.91/26.74/47.41	92.40	7.53	30.92
FlowSUM	53.13/**28.49**/49.00	92.67	**6.59**	29.77

Table 3: Comparison of BART, VEDSUM, and FlowSUM on short-summary benchmarks.

Effect of NF Types

Model	ROUGE 1/2/L ↑	BERTScore ↑	rep-w ↓
BART	42.56/15.35/36.67	86.69	**9.76**
VEDSUM	43.91/16.68/38.10	87.04	9.95
FlowSUM (Planar)	43.85/16.61/37.97	87.03	10.04
FlowSUM (Radial)	43.84/16.68/37.98	87.04	9.92
FlowSUM (Sylvester)	44.18/16.71/38.15	87.08	9.80
FlowSUM (RealNVP)	44.19/16.64/38.15	87.05	9.81
FlowSUM (IAF)	**44.42/17.01/38.36**	**87.09**	9.91
FlowSUM (RLNSF)	44.25/16.86/38.14	87.06	9.80
FlowSUM (RQNSF)	44.31/16.98/38.27	87.07	9.91

Table 4: Effect of NF Types on Multi-News.

Effect of NF Depth

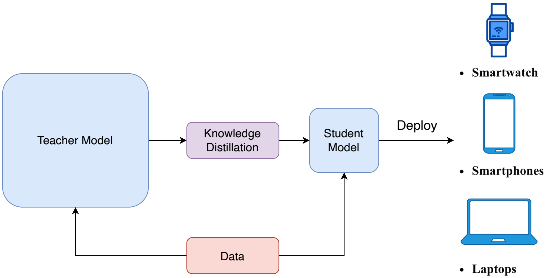
Model	ROUGE 1/2/L ↑	BERTScore ↑	rep-w ↓
FlowSUM (IAF-4)	44.30/**17.03**/38.22	87.05	**9.82**
FlowSUM (IAF-6)	**44.42**/17.01/**38.36**	**87.09**	9.91
FlowSUM (IAF-8)	44.18/16.90/38.16	87.04	9.88
FlowSUM (RQNSF-2)	44.15/16.88/38.20	87.04	9.94
FlowSUM (RQNSF-4)	**44.31**/**16.98**/**38.27**	**87.07**	9.91
FlowSUM (RQNSF-6)	44.15/16.88/38.18	87.06	**9.87**

Table 5: Effect of NF Depth on Multi-News.

Effect of Training Strategies

Model	Training	ROUGE ↑			KL
		1	2	L	Divergence
VEDSUM	standard	43.91	16.68	38.10	0.0117
VEDSUM	β_C-VAE	43.78	16.54	37.96	0.0082
FlowSUM (Planar)	standard	43.85	16.61	37.97	0.2719
FlowSUM (Planar)	β_C-VAE	43.68	16.47	37.85	0.1815
FlowSUM (Radial)	standard	43.63	16.37	37.82	0.0121
FlowSUM (Radial)	β_C-VAE	43.84	16.68	37.98	0.0096
FlowSUM (Sylvester)	standard	43.68	16.51	37.87	0.0841
FlowSUM (Sylvester)	β_C-VAE	44.18	16.71	38.15	0.0348
FlowSUM (RealNVP)	standard	44.19	16.64	38.15	4.7986
FlowSUM (RealNVP)	β_C-VAE	43.71	16.54	37.85	7.8938
FlowSUM (RealNVP)	CAAT	44.12	16.82	38.11	5.2107
FlowSUM (IAF)	standard	43.87	16.62	37.97	3.9146
FlowSUM (IAF)	β_C-VAE	43.81	16.58	37.91	3.9128
FlowSUM (IAF)	CAAT	44.30	17.03	38.22	2.1108
FlowSUM (RLNSF)	standard	44.25	16.86	38.14	104.9667
FlowSUM (RLNSF)	β_C-VAE	44.25	16.86	38.14	104.9667
FlowSUM (RLNSF)	CAAT	44.14	16.82	38.05	95.3774
FlowSUM (RQNSF)	standard	44.18	16.76	38.18	127.8106
FlowSUM (RQNSF)	β_C-VAE	44.18	16.76	38.18	127.8106
FlowSUM (RQNSF)	CAAT	44.31	16.98	38.27	107.0794

NF-enhanced Knowledge Distillation



Knowledge Distillation (Shleifer and Rush, 2020)

- Shrink and Fine-Tune (SFT): shrinks the teacher model and re-finetunes the shrunk model
- Pseudo-labels (PL): initializes the student model with the compressed version produced by SFT and then fine-tunes on the pseudo-labeled data generated by the teacher model

NF-enhanced Knowledge Distillation

- FlowSUM $>$ dBART $\Rightarrow$ NF helps with SFT
- FlowSUM-PLKD $>$ FlowSUM $\Rightarrow$ NF boosts further with PL.

Model	ROUGE ↑ 1/2/L	BERT-Score ↑	Length	# Params (MM)	Inference Time (MS) ↓
		dBART-6-6			
dBART-6-6	42.78/20.24/39.72	88.98	67.42	230	170.5
FlowSUM	43.41/20.33/40.41	89.18	91.25	238	234.9
FlowSUM-PLKD	**43.70/20.71/40.73**	**89.24**	91.10	238	239.7
		dBART-12-3			
dBART-12-3	43.39/20.57/40.44	89.20	85.48	255	199.6
FlowSUM	43.53/20.61/40.59	89.28	83.74	263	190.7
FlowSUM-PLKD	**44.05/21.06/41.07**	**89.37**	84.48	263	200.4

Table 6: Knowledge Distillation on DistilBART on CNN/DM.

NF-enhanced Knowledge Distillation

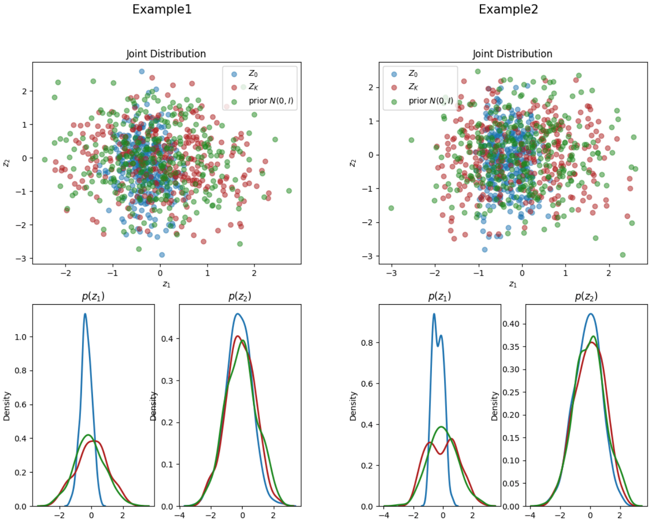


Figure 4: Visualization of the first two dimensions of z_0, z_K, and $N(0, I)$ from FlowSUM-PLKD on CNN/DM.

Conclusion

1. Propose FlowSUM, a normalizing flows-based variational encoder-decoder framework for Transformer-based models.
2. Improve the training efficacy with a training strategy and a refined gate mechanism.
3. Investigate the operating characteristics of normalizing flows in summarization.
4. Identify normalizing flows' benefits for knowledge distillation.

Thank You!