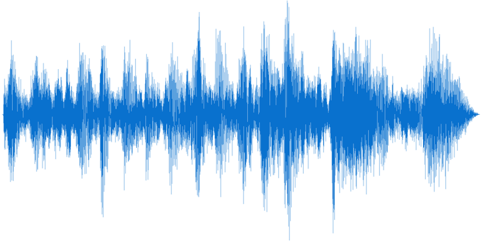
Latent Normalizing Flows for Discrete Sequences

Zachary M. Ziegler, Alexander M. Rush

School of Engineering and Applied Sciences, Harvard University

Poster #3 @ Pacific Ballroom

Motivation: Normalizing flows



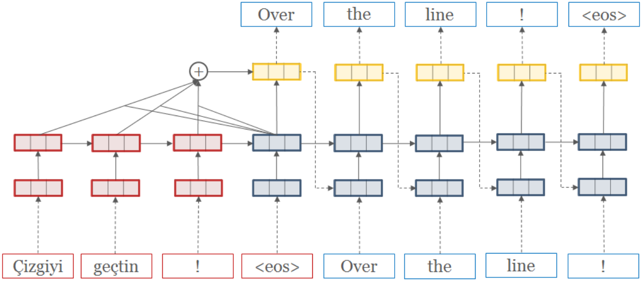
For invertible $f_\theta : \epsilon \to \mathcal{Z}$ and base density $p_\epsilon(\boldsymbol{\epsilon})$,

$$p_Z(\boldsymbol{z}) = p_\epsilon(f_\theta^{-1}(\boldsymbol{z})) \left| \det \frac{\partial f_\theta^{-1}(\boldsymbol{z})}{\partial \boldsymbol{z}} \right|$$

- Flows generalize autoregressive models for continuous data, allowing **increased model flexibility** and **non-autoregressive generation**.

Kingma and Dhariwal 2018, van den Oord et al. 2017, Rezende and Mohamed 2015

Goal: Flows for discrete data

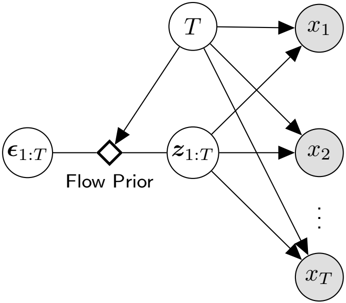


- For discrete sequences MLE autoregressive models are ubiquitous. Can flows go beyond AR models for discrete sequences?

Figure: OpenNMT

Challenges and approach

1. Discrete change of variables poses theoretical and practical challenges compared to continuous change of variables.



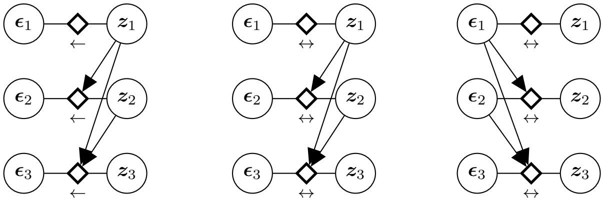
$$\boldsymbol{x} \in \mathcal{V}^T \quad \boldsymbol{z} \in \mathcal{R}^{T \times H}$$

- Latent variable model where prior $p(\boldsymbol{z}_{1:T})$ captures dynamics of discrete data over time.
- Key: weak conditionally independent emission model.
- VAE for inference, optimize ELBO.

Challenges and approach

1. Discrete change of variables poses theoretical and practical challenges compared to continuous.
2. Discrete data is inherently highly multimodal.

- Specialized flows for multimodal sequences:
  - Model dependencies across dimension and across time.

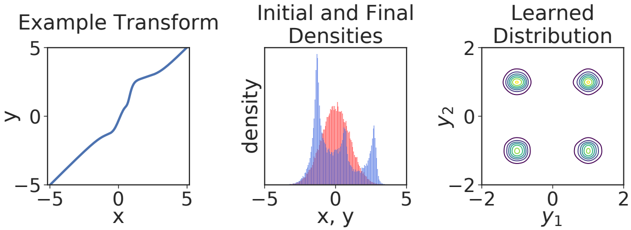


Autoregressive ($\leftarrow$) | Autoregressive in time ($\leftarrow$) | Non-autoregressive ($\rightarrow$)

Challenges and approach

1. Discrete change of variables poses theoretical and practical challenges compared to continuous.
2. Discrete data is inherently highly multimodal.

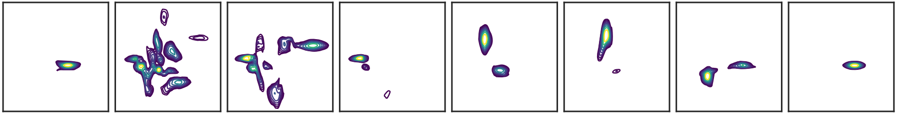
- Specialized flows for multimodal sequences:
  - Model dependencies across dimension and across time.
  - Replace underlying affine transformation with non-linear transformation.



Experiments: Character-level LM, PTB

Model	Test NLL	Reconst.	KL
LSTM	**1.41**	-	-
Independent-across-time flow	2.90	0.15	2.77
Autoregressive ($\leftarrow$)	**1.42**	0.10	**1.37**
Autoregressive in time ($\leftarrow$)	1.46	0.10	1.43
Non-autoregressive ($\rightarrow$)	1.63	0.21	1.55

- KL always makes up $> 90\%$ of loss, indicating continuous flow models vast majority of uncertainty.
- Additional experiments on polyphonic music generation.



_ g r o u p s _

Conclusions

- Latent variable model for discrete sequences modeling discrete dynamics in continuous latent space with continuous flows.
- See poster for details of approach, more experimental results, and generation speed comparison.

Poster #3 @ Pacific Ballroom, for details and more experiments