Диффузионные модели для текстовых данных

ВШЭ ФКН, Методы предобучения без учителя

Шабалин Александр

Multinomial Diffusion

Категориальное распределение

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{C}(\boldsymbol{x}_t|(1-\beta_t)\boldsymbol{x}_{t-1} + \beta_t/K)$$

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{C}(\boldsymbol{x}_t|\bar{\alpha}_t\boldsymbol{x}_0 + (1-\bar{\alpha}_t)/K)$$

https://arxiv.org/pdf/2102.05379.pdf

Multinomial Diffusion

Категориальное распределение

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{C}(\boldsymbol{x}_t|(1-\beta_t)\boldsymbol{x}_{t-1} + \beta_t/K)$$

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{C}(\boldsymbol{x}_t|\bar{\alpha}_t\boldsymbol{x}_0 + (1-\bar{\alpha}_t)/K)$$

$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{C}(\boldsymbol{x}_{t-1}|\boldsymbol{\theta}_{\text{post}}(\boldsymbol{x}_t, \boldsymbol{x}_0)), \quad \text{where} \quad \boldsymbol{\theta}_{\text{post}}(\boldsymbol{x}_t, \boldsymbol{x}_0) = \tilde{\boldsymbol{\theta}} / \sum_{k=1}^{K} \tilde{\theta}_k$$

$$\text{and} \quad \tilde{\boldsymbol{\theta}} = [\alpha_t\boldsymbol{x}_t + (1-\alpha_t)/K] \odot [\bar{\alpha}_{t-1}\boldsymbol{x}_0 + (1-\bar{\alpha}_{t-1})/K].$$

https://arxiv.org/pdf/2102.05379.pdf

Multinomial Diffusion

$$p(\boldsymbol{x}_0|\boldsymbol{x}_1) = \mathcal{C}(\boldsymbol{x}_0|\hat{\boldsymbol{x}}_0)$$

$$p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{C}(\boldsymbol{x}_{t-1}|\boldsymbol{\theta}_{\text{post}}(\boldsymbol{x}_t, \hat{\boldsymbol{x}}_0))$$

$$\hat{\boldsymbol{x}}_0 = \mu(\boldsymbol{x}_t, t)$$

https://arxiv.org/pdf/2102.05379.pdf

Multinomial Diffusion

$$p(\boldsymbol{x}_0|\boldsymbol{x}_1) = \mathcal{C}(\boldsymbol{x}_0|\hat{\boldsymbol{x}}_0)$$

$$p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{C}(\boldsymbol{x}_{t-1}|\boldsymbol{\theta}_{\text{post}}(\boldsymbol{x}_t, \hat{\boldsymbol{x}}_0))$$

$$\hat{\boldsymbol{x}}_0 = \mu(\boldsymbol{x}_t, t)$$

$$\mathrm{KL}\big(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)|p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)\big) = \mathrm{KL}\big(\mathcal{C}(\boldsymbol{\theta}_{\text{post}}(\boldsymbol{x}_t, \boldsymbol{x}_0))|\mathcal{C}(\boldsymbol{\theta}_{\text{post}}(\boldsymbol{x}_t, \hat{\boldsymbol{x}}_0))\big) =$$
$$= \sum_k \boldsymbol{\theta}_{\text{post}}(\boldsymbol{x}_t, \boldsymbol{x}_0))_k \cdot \log \frac{\boldsymbol{\theta}_{\text{post}}(\boldsymbol{x}_t, \boldsymbol{x}_0))_k}{\boldsymbol{\theta}_{\text{post}}(\boldsymbol{x}_t, \hat{\boldsymbol{x}}_0))_k}$$

https://arxiv.org/pdf/2102.05379.pdf

Multinomial Diffusion

$$p(\boldsymbol{x}_0|\boldsymbol{x}_1) = \mathcal{C}(\boldsymbol{x}_0|\hat{\boldsymbol{x}}_0)$$

$$p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{C}(\boldsymbol{x}_{t-1}|\boldsymbol{\theta}_{\text{post}}(\boldsymbol{x}_t, \hat{\boldsymbol{x}}_0))$$

$$\hat{\boldsymbol{x}}_0 = \mu(\boldsymbol{x}_t, t)$$

$$\mathrm{KL}\big(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)|p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)\big) = \mathrm{KL}\big(\mathcal{C}(\boldsymbol{\theta}_{\text{post}}(\boldsymbol{x}_t, \boldsymbol{x}_0))|\mathcal{C}(\boldsymbol{\theta}_{\text{post}}(\boldsymbol{x}_t, \hat{\boldsymbol{x}}_0))\big) =$$
$$= \sum_k \boldsymbol{\theta}_{\text{post}}(\boldsymbol{x}_t, \boldsymbol{x}_0))_k \cdot \log \frac{\boldsymbol{\theta}_{\text{post}}(\boldsymbol{x}_t, \boldsymbol{x}_0))_k}{\boldsymbol{\theta}_{\text{post}}(\boldsymbol{x}_t, \hat{\boldsymbol{x}}_0))_k}$$

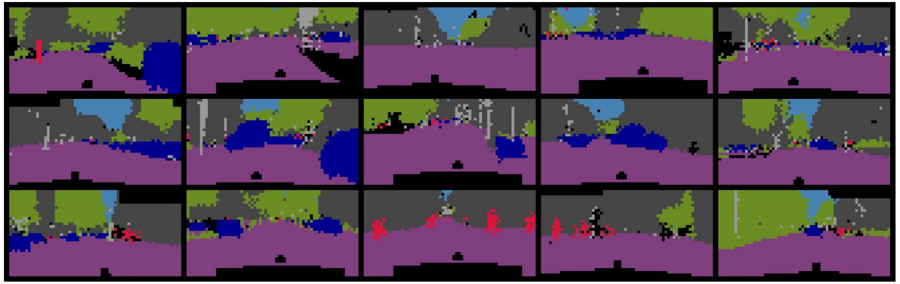
$$\log p(\boldsymbol{x}_0|\boldsymbol{x}_1) = \sum_k \boldsymbol{x}_{0,k} \log \hat{\boldsymbol{x}}_{0,k}$$

https://arxiv.org/pdf/2102.05379.pdf

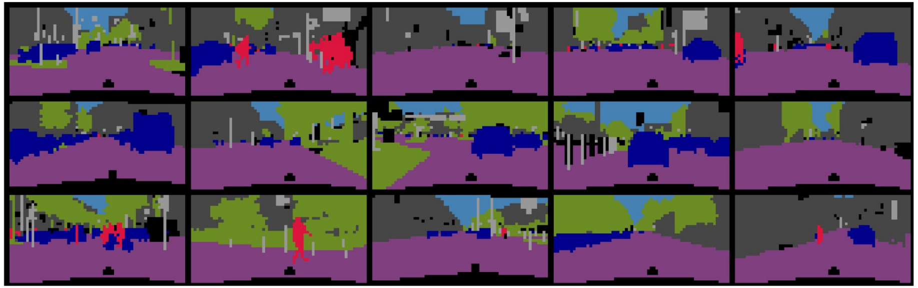
Multinomial Diffusion

heartedness frege thematically infered by the famous existence of a fu
nction f from the laplace definition we can analyze a definition of bin
ary operations with additional size so their functionality cannot be re
viewed here there is no change because its

otal cost of learning objects from language to platonic linguistics exa
mines why animate to indicate wild amphibious substances animal and mar
ine life constituents of animals and bird sciences medieval biology bio
logy and central medicine full discovery re



(b) Samples from the Multinomial Diffusion model.

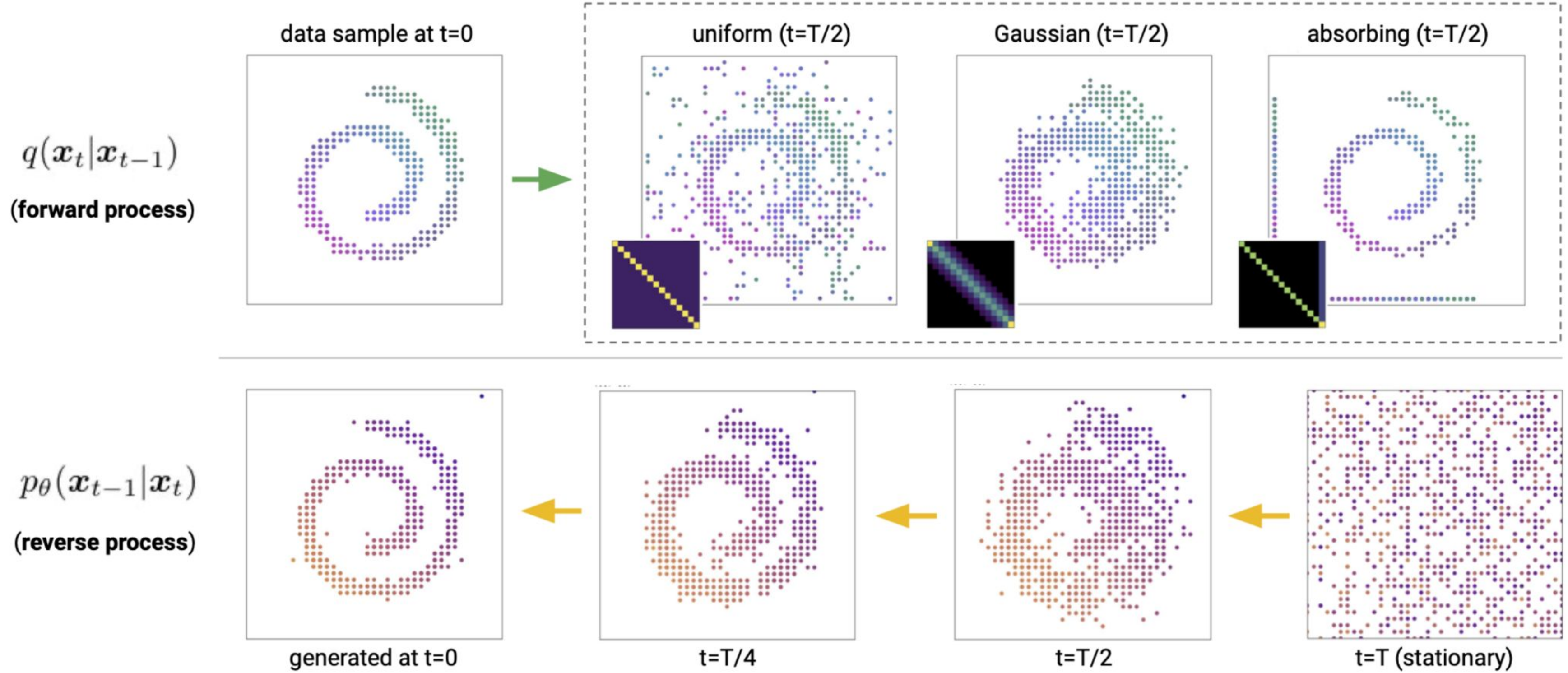


(c) Cityscapes data.

https://arxiv.org/pdf/2102.05379.pdf

D3PM

Discrete Denoising Diffusion Probabilistic Model



https://arxiv.org/pdf/2107.03006.pdf

D3PM

Discrete Denoising Diffusion Probabilistic Model

$$x_t, x_{t-1} \in 1, ..., K$$

$$[\boldsymbol{Q}_t]_{ij} = q(x_t = j | x_{t-1} = i)$$

https://arxiv.org/pdf/2107.03006.pdf

D3PM
Discrete Denoising Diffusion Probabilistic Model

$$x_t, x_{t-1} \in 1, ..., K$$

$$[\boldsymbol{Q}_t]_{ij} = q(x_t = j | x_{t-1} = i)$$

$\boldsymbol{x}$ – one-hot row vector

$$q(\boldsymbol{x}_t | \boldsymbol{x}_{t-1}) = \mathrm{Cat}(\boldsymbol{x}_t; \boldsymbol{p} = \boldsymbol{x}_{t-1} \boldsymbol{Q}_t)$$

https://arxiv.org/pdf/2107.03006.pdf

D3PM

Discrete Denoising Diffusion Probabilistic Model

$$x_t, x_{t-1} \in 1, ..., K$$

$$[\boldsymbol{Q}_t]_{ij} = q(x_t = j | x_{t-1} = i)$$

$\boldsymbol{x}$ – one-hot row vector

$$q(\boldsymbol{x}_t | \boldsymbol{x}_{t-1}) = \mathrm{Cat}(\boldsymbol{x}_t; \boldsymbol{p} = \boldsymbol{x}_{t-1} \boldsymbol{Q}_t)$$

$$q(\boldsymbol{x}_t | \boldsymbol{x}_0) = \mathrm{Cat}\left(\boldsymbol{x}_t; \boldsymbol{p} = \boldsymbol{x}_0 \overline{\boldsymbol{Q}}_t\right), \quad \text{with} \quad \overline{\boldsymbol{Q}}_t = \boldsymbol{Q}_1 \boldsymbol{Q}_2 \dots \boldsymbol{Q}_t$$

$$q(\boldsymbol{x}_{t-1} | \boldsymbol{x}_t, \boldsymbol{x}_0) = \frac{q(\boldsymbol{x}_t | \boldsymbol{x}_{t-1}, \boldsymbol{x}_0) q(\boldsymbol{x}_{t-1} | \boldsymbol{x}_0)}{q(\boldsymbol{x}_t | \boldsymbol{x}_0)} = \mathrm{Cat}\left(\boldsymbol{x}_{t-1}; \boldsymbol{p} = \frac{\boldsymbol{x}_t \boldsymbol{Q}_t^\top \odot \boldsymbol{x}_0 \overline{\boldsymbol{Q}}_{t-1}}{\boldsymbol{x}_0 \overline{\boldsymbol{Q}}_t \boldsymbol{x}_t^\top}\right)$$

https://arxiv.org/pdf/2107.03006.pdf

D3PM

Discrete Denoising Diffusion Probabilistic Model

Uniform diffusion

$$[\boldsymbol{Q}_t]_{ij} = \begin{cases} 1 - \frac{K-1}{K}\beta_t & \text{if} \quad i = j \\ \frac{1}{K}\beta_t & \text{if} \quad i \neq j \end{cases}$$

https://arxiv.org/pdf/2107.03006.pdf

D3PM
Discrete Denoising Diffusion Probabilistic Model

Uniform diffusion

$$[\boldsymbol{Q}_t]_{ij} = \begin{cases} 1 - \frac{K-1}{K}\beta_t & \text{if} \quad i = j \\ \frac{1}{K}\beta_t & \text{if} \quad i \neq j \end{cases}$$

Diffusion with an absorbing state

$$[\boldsymbol{Q}_t]_{ij} = \begin{cases} 1 & \text{if} \quad i = j = m \\ 1 - \beta_t & \text{if} \quad i = j \neq m \\ \beta_t & \text{if} \quad j = m, i \neq m \end{cases}$$

https://arxiv.org/pdf/2107.03006.pdf

D3PM

Discrete Denoising Diffusion Probabilistic Model

Uniform diffusion

$$[\boldsymbol{Q}_t]_{ij} = \begin{cases} 1 - \frac{K-1}{K}\beta_t & \text{if} \quad i = j \\ \frac{1}{K}\beta_t & \text{if} \quad i \neq j \end{cases}$$

Discretized Gaussian transition matrices

$$[\boldsymbol{Q}_t]_{ij} = \begin{cases} \dfrac{\exp\left(-\frac{4|i-j|^2}{(K-1)^2\beta_t}\right)}{\sum_{n=-(K-1)}^{K-1} \exp\left(-\frac{4n^2}{(K-1)^2\beta_t}\right)} & \text{if} \quad i \neq j \\ 1 - \sum_{l=0, l\neq i}^{K-1} [\boldsymbol{Q}_t]_{il} & \text{if} \quad i = j \end{cases}$$

Diffusion with an absorbing state

$$[\boldsymbol{Q}_t]_{ij} = \begin{cases} 1 & \text{if} \quad i = j = m \\ 1 - \beta_t & \text{if} \quad i = j \neq m \\ \beta_t & \text{if} \quad j = m, i \neq m \end{cases}$$

https://arxiv.org/pdf/2107.03006.pdf

D3PM

Discrete Denoising Diffusion Probabilistic Model

Uniform diffusion

$$[\boldsymbol{Q}_t]_{ij} = \begin{cases} 1 - \frac{K-1}{K}\beta_t & \text{if} \quad i = j \\ \frac{1}{K}\beta_t & \text{if} \quad i \neq j \end{cases}$$

Diffusion with an absorbing state

$$[\boldsymbol{Q}_t]_{ij} = \begin{cases} 1 & \text{if} \quad i = j = m \\ 1 - \beta_t & \text{if} \quad i = j \neq m \\ \beta_t & \text{if} \quad j = m, i \neq m \end{cases}$$

Discretized Gaussian transition matrices

$$[\boldsymbol{Q}_t]_{ij} = \begin{cases} \dfrac{\exp\left(-\frac{4|i-j|^2}{(K-1)^2\beta_t}\right)}{\sum_{n=-(K-1)}^{K-1}\exp\left(-\frac{4n^2}{(K-1)^2\beta_t}\right)} & \text{if} \quad i \neq j \\ 1 - \sum_{l=0, l\neq i}^{K-1}[\boldsymbol{Q}_t]_{il} & \text{if} \quad i = j \end{cases}$$

Structured diffusion in text

$[\mathbf{G}]_{ij} = 1$ if w_i is a k-nearest neighbor of w_j else 0

$$\mathbf{A} = (\mathbf{G} + \mathbf{G}^T)/(2k)$$

$$[\boldsymbol{R}]_{ij} = \begin{cases} -\sum_{l\neq i} A_{il} & \text{if} \quad i = j \\ A_{ij} & \text{otherwise} \end{cases}$$

$$\mathbf{Q}_t = \exp(\alpha_t \mathbf{R}) = \sum_{n=0}^{\infty} \frac{\alpha_t^n}{n!}\boldsymbol{R}^n$$

https://arxiv.org/pdf/2107.03006.pdf

Diffusion-LM

$$\textsc{Emb}(\mathbf{w}) = [\textsc{Emb}(w_1), \ldots, \textsc{Emb}(w_n)] \in \mathbb{R}^{nd}$$

$$q_\phi(\mathbf{x}_0|\mathbf{w}) = \mathcal{N}(\textsc{Emb}(\mathbf{w}), \sigma_0 I) \qquad p_\theta(\mathbf{w} \mid \mathbf{x}_0) = \prod_{i=1}^{n} p_\theta(w_i \mid x_i)$$

https://arxiv.org/pdf/2205.14217.pdf

Diffusion-LM

$$\textsc{Emb}(\mathbf{w}) = [\textsc{Emb}(w_1), \ldots, \textsc{Emb}(w_n)] \in \mathbb{R}^{nd}$$

$$q_\phi(\mathbf{x}_0|\mathbf{w}) = \mathcal{N}(\textsc{Emb}(\mathbf{w}), \sigma_0 I) \qquad p_\theta(\mathbf{w} \mid \mathbf{x}_0) = \prod_{i=1}^{n} p_\theta(w_i \mid x_i)$$

$$\mathcal{L}_{\text{simple}}(\mathbf{x}_0) = \sum_{t=1}^{T} \mathop{\mathbb{E}}_{q(\mathbf{x}_t|\mathbf{x}_0)} ||\mu_\theta(\mathbf{x}_t, t) - \hat{\mu}(\mathbf{x}_t, \mathbf{x}_0)||^2$$

$$\mathcal{L}^{\text{e2e}}_{\text{simple}}(\mathbf{w}) = \mathop{\mathbb{E}}_{q_\phi(\mathbf{x}_{0:T}|\mathbf{w})} \left[\mathcal{L}_{\text{simple}}(\mathbf{x}_0) + ||\textsc{Emb}(\mathbf{w}) - \mu_\theta(\mathbf{x}_1, 1)||^2 - \log p_\theta(\mathbf{w}|\mathbf{x}_0)\right]$$

https://arxiv.org/pdf/2205.14217.pdf

DiffuSeq

$$\textsc{Emb}(\mathbf{w}^{x \oplus y}) \ = \ [\textsc{Emb}(w_1^x), ..., \textsc{Emb}(w_m^x), \textsc{Emb}(w_1^y), ..., \textsc{Emb}(w_n^y)] \in \mathbb{R}^{(m+n) \times d}$$

https://arxiv.org/pdf/2210.08933.pdf

DiffuSeq

$$\text{EMB}(\mathbf{w}^{x \oplus y}) \ = \ [\text{EMB}(w_1^x), ..., \text{EMB}(w_m^x), \text{EMB}(w_1^y), ..., \text{EMB}(w_n^y)] \in \mathbb{R}^{(m+n) \times d}$$

$$q_\phi(\mathbf{z}_0 | \mathbf{w}^{x \oplus y}) = \mathcal{N}(\text{EMB}(\mathbf{w}^{x \oplus y}), \beta_0 \mathbf{I})$$

$$p_\theta(\mathbf{z}_{t-1} | \mathbf{z}_t) \ = \ \mathcal{N}(\mathbf{z}_{t-1}; \mu_\theta(\mathbf{z}_t, t), \sigma_\theta(\mathbf{z}_t, t))$$

https://arxiv.org/pdf/2210.08933.pdf

DiffuSeq

$$\textsc{Emb}(\mathbf{w}^{x\oplus y}) \ = \ [\textsc{Emb}(w_1^x), ..., \textsc{Emb}(w_m^x), \textsc{Emb}(w_1^y), ..., \textsc{Emb}(w_n^y)] \in \mathbb{R}^{(m+n)\times d}$$

$$q_\phi(\mathbf{z}_0|\mathbf{w}^{x\oplus y}) = \mathcal{N}(\textsc{Emb}(\mathbf{w}^{x\oplus y}), \beta_0\mathbf{I})$$

$$p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t) \ = \ \mathcal{N}(\mathbf{z}_{t-1}; \mu_\theta(\mathbf{z}_t, t), \sigma_\theta(\mathbf{z}_t, t))$$

$$\mathcal{L}_{\text{VLB}} = \mathbb{E}_{q(\mathbf{z}_{1:T}|\mathbf{z}_0)}\Bigg[\underbrace{\log\frac{q(\mathbf{z}_T|\mathbf{z}_0)}{p_\theta(\mathbf{z}_T)}}_{\mathcal{L}_T} + \sum_{t=2}^{T}\underbrace{\log\frac{q(\mathbf{z}_{t-1}|\mathbf{z}_0,\mathbf{z}_t)}{p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t)}}_{\mathcal{L}_{t-1}}$$
$$+\underbrace{\log\frac{q_\phi(\mathbf{z}_0|\mathbf{w}^{x\oplus y})}{p_\theta(\mathbf{z}_0|\mathbf{z}_1)}}_{\mathcal{L}_0} \underbrace{- \log p_\theta(\mathbf{w}^{x\oplus y}|\mathbf{z}_0)}_{\mathcal{L}_{\text{round}}}\Bigg].$$

https://arxiv.org/pdf/2210.08933.pdf

DiffuSeq

$$\textsc{Emb}(\mathbf{w}^{x\oplus y}) \;=\; [\textsc{Emb}(w_1^x), ..., \textsc{Emb}(w_m^x), \textsc{Emb}(w_1^y), ..., \textsc{Emb}(w_n^y)] \in \mathbb{R}^{(m+n)\times d}$$

$$q_\phi(\mathbf{z}_0|\mathbf{w}^{x\oplus y}) = \mathcal{N}(\textsc{Emb}(\mathbf{w}^{x\oplus y}), \beta_0\mathbf{I})$$

$$p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t) \;=\; \mathcal{N}(\mathbf{z}_{t-1}; \mu_\theta(\mathbf{z}_t, t), \sigma_\theta(\mathbf{z}_t, t))$$

$$\mathcal{L}_{\text{VLB}} = \mathbb{E}_{q(\mathbf{z}_{1:T}|\mathbf{z}_0)}\Bigg[\underbrace{\log\frac{q(\mathbf{z}_T|\mathbf{z}_0)}{p_\theta(\mathbf{z}_T)}}_{\mathcal{L}_T} + \sum_{t=2}^{T}\underbrace{\log\frac{q(\mathbf{z}_{t-1}|\mathbf{z}_0,\mathbf{z}_t)}{p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t)}}_{\mathcal{L}_{t-1}}$$

$$+\underbrace{\log\frac{q_\phi(\mathbf{z}_0|\mathbf{w}^{x\oplus y})}{p_\theta(\mathbf{z}_0|\mathbf{z}_1)}}_{\mathcal{L}_0} - \underbrace{\log p_\theta(\mathbf{w}^{x\oplus y}|\mathbf{z}_0)}_{\mathcal{L}_{\text{round}}}\Bigg].$$

$$\min_\theta\ \mathcal{L}_{\text{VLB}} = \min_\theta\left[\sum_{t=2}^{T}||\mathbf{z}_0 - f_\theta(\mathbf{z}_t, t)||^2 + ||\textsc{Emb}(\mathbf{w}^{x\oplus y}) - f_\theta(\mathbf{z}_1, 1)||^2 - \log p_\theta(\mathbf{w}^{x\oplus y}|\mathbf{z}_0)\right]$$

https://arxiv.org/pdf/2210.08933.pdf