Argmax Flows and Multinomial Diffusion: Learning Categorical Distributions

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling

NeurIPS 2021

Presented by Minho Park

https://arxiv.org/abs/2102.05379

Contributions

- Propose Argmax Flows
  - Remove autoregressive sampling in discrete domain.
- Propose Multinomial Diffusion models
  - Propose Categorical noise and experiment in text and segmentation map domain.

Motivation

- Abundant categorical data:
    - Text, Semantic map, Molecules, Proteins, DNA …
- Autoregressive models are slow.
    - Fast training but slow sampling.

Normalizing Flows

- A lot of flow-based model in continuous domain (e.g., image and audio).
- Forward:
  - $z \sim p(z)$ sampling from a (typically simple) tractable density.
  - $x = f_\theta(z)$
  - Then, we can achieve $p(x) = p(z) \cdot \left|\det \frac{\mathrm{d}z}{\mathrm{d}v}\right|$
- Optimization:
  - $\mathcal{L}(\mathcal{D}) = \frac{1}{N}\sum_{i=1}^{N} - \log p_\theta(x)$
  - $\log p(x) = \log p(z) + \log \left|\det \underbrace{\frac{\mathrm{d}z}{\mathrm{d}x}}\right|$ $\left.(f_\theta^{-1})'\right|_{x=x^{(i)}}$
  - $= \log p(z) + \sum_{i=1}^{K} \log \left|\det \frac{\mathrm{d}h_i}{\mathrm{d}h_{i-1}}\right|$

$$\text{minimize KL}\big(p(x) \parallel p_\theta(x)\big) = \int p(x) \log \frac{p(x)}{p_\theta(x)} \mathrm{d}x$$

$$\Leftrightarrow \text{minimize } \frac{1}{N}\sum_{i=1}^{N} - \log p_\theta(x)$$

Figure 1: Synthetic celebrities sampled from our model; see Section 3 for architecture and method, and Section 5 for more results.

Kingma, Durk P., and Prafulla Dhariwal. "Glow: Generative flow with invertible 1x1 convolutions." *Advances in neural information processing systems* 31 (2018).

Discrete Data + Flows

	Ordinal	Categorical
Discrete Flows	Integer Discrete Flows (Hoogeboom et al. 2019)	Discrete Flows (Tran et al. 2019)
Surjective Flows	Dequantization (Uria et al. 2013)	Argmax Flows (Hoogeboom et al. 2021)

Discrete Flows

- **Sampling from Discrete density $\boldsymbol{P(z)}$.** (Assume $P_Z(z) = P_X(x)$)
- Mapping with Discrete function f.

For **Ordinal** data:

Integer Discrete Flows

(Hoogeboom et al. 2019)

$$z_d = x_d + \mu_d$$ (Not sure…)

Use Straight-Through estimator:

Forward: $\mu_d = [\theta_d]$

Backward: θ_d (ignore [ ])

For **Categorical** Data:

Discrete Flows

(Tran et al. 2019)

$$z_d = \mu_d + \sigma_d x_d \pmod K$$

Use Straight-Through estimator:

Forward: $\mu_d = \text{ont\_hot}\big(\text{argmax}(\theta_d)\big)$

Backward: $\text{softmax}(\theta_d/\tau)$

Hoogeboom, Emiel, et al. "Integer discrete flows and lossless compression." *Advances in Neural Information Processing Systems* 32 (2019).
Tran, Dustin, et al. "Discrete flows: Invertible generative models of discrete data." *Advances in Neural Information Processing Systems* 32 (2019).

Drawbacks of Discrete Flows

- Limited flexibility: Can only permute probability mass.
  - They suppose $P(x) = P(z)$ and function f only permute the $P(z)$.
- Gradient bias: introduced by the straight-through estimator.

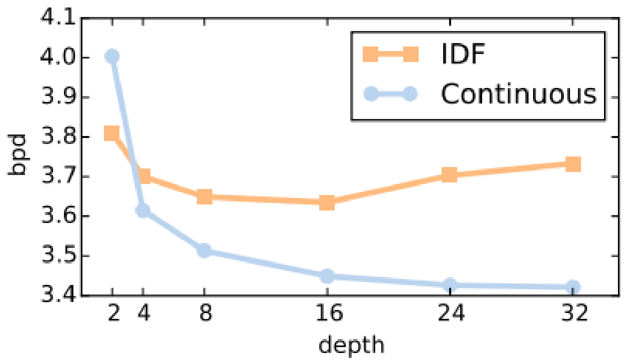


Figure 5: Performance of flow models for different depths (i.e. coupling layers per level). The networks in the coupling layers contain 3 convolution layers. Although performance increases with depth for continuous flows, this is not the case for discrete flows.

Hoogeboom, Emiel, et al. "Integer discrete flows and lossless compression." *Advances in Neural Information Processing Systems* 32 (2019).

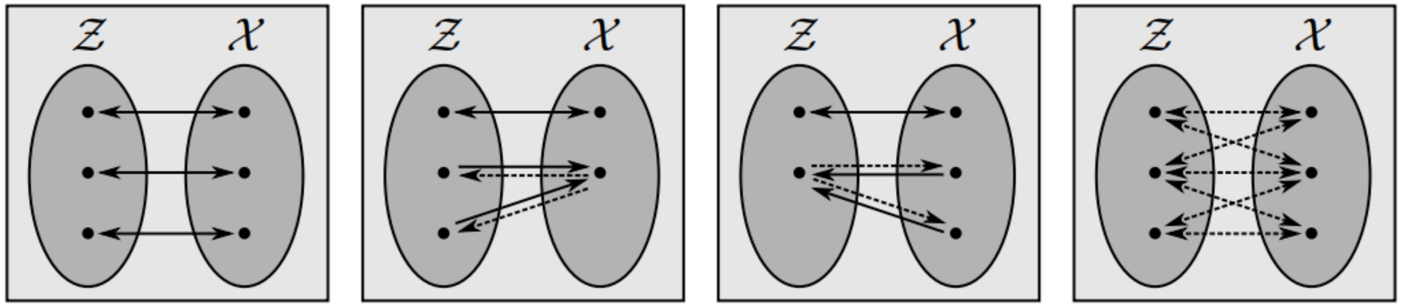
Surjective Flows

- **Sampling from Continuous density $p(z)$.**
- Mapping with continuous function f.

NF / DF

Our Focus

VAE



(a) Bijective (b) Surjective (Gen.) (c) Surjective (Inf.) (d) Stochastic

Figure 1: Classes of SurVAE transformations $\mathcal{Z} \to \mathcal{X}$ and their inverses $\mathcal{X} \to \mathcal{Z}$. Solid lines indicate deterministic transformations, while dashed lines indicate stochastic transformations.

Transformation	**Forward** $\boldsymbol{x} \leftarrow \boldsymbol{z}$	**Inverse** $\boldsymbol{z} \leftarrow \boldsymbol{x}$
Bijective	$\boldsymbol{x} = f(\boldsymbol{z})$	$\boldsymbol{z} = f^{-1}(\boldsymbol{x})$
Stochastic	$\boldsymbol{x} \sim p(\boldsymbol{x}\|\boldsymbol{z})$	$\boldsymbol{z} \sim q(\boldsymbol{z}\|\boldsymbol{x})$
Surjective (Gen.)	$\boldsymbol{x} = f(\boldsymbol{z})$	$\boldsymbol{z} \sim q(\boldsymbol{z}\|\boldsymbol{x})$
Surjective (Inf.)	$\boldsymbol{x} \sim p(\boldsymbol{x}\|\boldsymbol{z})$	$\boldsymbol{z} = f^{-1}(\boldsymbol{x})$

Table 1: Composable building blocks of SurVAE Flows.

Nielsen, Didrik, et al. "Survae flows: Surjections to bridge the gap between vaes and flows." *Advances in Neural Information Processing Systems* 33 (2020): 12685-12696.

Surjective Flows

- Dequantization (Uria et al. 2013)
- Forward:
  - $z \sim p(z)$ sampling from a **Continuous** simple density (e.g., spherical Gaussian)
  - $x = f_\theta(z) \Rightarrow x = \mathrm{round}(z)$
  - Then, we can achieve $p(x) = p(z) \cdot \left|\det \frac{\mathrm{d}z}{\mathrm{d}x}\right|$
- Inverse:
  - $z \sim q(z|x)$: stochastic right inverse. $\Rightarrow z = \mathrm{Unif}(z|x, x+1)$ w/ support $\mathcal{S}(x) = \{x|x = \mathrm{round}(z)\}$

Uria, Benigno, Iain Murray, and Hugo Larochelle. "RNADE: The real-valued neural autoregressive density-estimator." *Advances in Neural Information Processing Systems* 26 (2013).

Surjective Flows

- Objective function: $x = \mathrm{round}(y)$ and $y = f_\theta(z)$

$$\mathcal{L}(\mathcal{D}) = \frac{1}{N}\sum_{i=1}^{N} -\log P_\theta(x)$$

$P(x|z) = 1$ if $q(z|x)$ is enforced over
$\mathcal{S} = \{z \in \mathbb{R}^d : x = \mathrm{round}(z)\}$.

$\mathrm{Unif}(y|x, x+1)$

$$\log P_\theta(x) \geq \mathbb{E}_{y \sim \underline{q(y|x)}}[\cancel{\log P(x|y)} + \log p_\theta(y) - \log \underline{q(y|x)}]$$

$$\because \log P_\theta(x) = \log \int P(x|y) \cdot p_\theta(y) \cdot \frac{q(y|x)}{q(y|x)} \mathrm{d}y \geq \int \log\left(P(x|y) \cdot p_\theta(y) \cdot \frac{q(y|x)}{q(y|x)}\right) \mathrm{d}y$$

$$\mathrm{ELBO} = \mathbb{E}_{y \sim \mathrm{Unif}(y|x, x+1)}[\log p_\theta(y)]$$

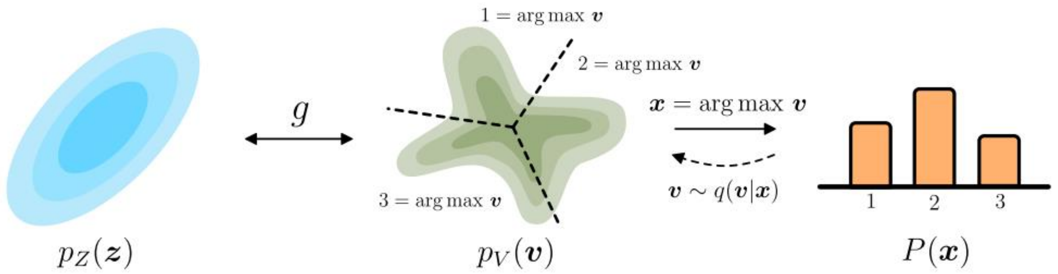
$$= \mathbb{E}_{y \sim \mathrm{Unif}(y|x, x+1)}\left[\log\left(p_\theta(z) \cdot \left|\det \frac{\mathrm{d}z}{\mathrm{d}y}\right|\right)\right]$$

$$= \mathbb{E}_{y \sim \mathrm{Unif}(y|x, x+1)}\left[\log\left(p(z) \cdot \left|\det (f_\theta^{-1})'(y)\right|\right)\right]$$

Uria, Benigno, Iain Murray, and Hugo Larochelle. "RNADE: The real-valued neural autoregressive density-estimator." *Advances in Neural Information Processing Systems* 26 (2013).

Argmax Flows

- Forward:
  - $x = \text{argmax}(z)$
- Inverse:
  - $z \sim q(z|x)$ w/ support $\mathcal{S}(x) = \{x | x = \text{argmax}(z)\}$



(a) Argmax Flow: Composition of a flow $p(\boldsymbol{v})$ and argmax transformation which gives the model $P(\boldsymbol{x})$. The flow maps from a base distribution $p(\boldsymbol{z})$ using a bijection g.

Argmax Flows

- Modeling $q_\theta(v|x)$
- Thresholding
- $u \sim q(u|x)$: Normalizing Flows or conditional Gaussian
- $v_x = u_x$ and $\boldsymbol{v}_{-x} = \text{threshold}_T(\boldsymbol{u}_{-x})$ ($-x$ means remained elements)
- $v = \text{threshold}_T(u) = T - \log(1 + e^{T-u}) \in (-\infty, T)$

Algorithm 3 Thresholding-based $q(\boldsymbol{v}|\boldsymbol{x})$

Input: $\boldsymbol{x}$, $q(\boldsymbol{u}|\boldsymbol{x})$
Output: $\boldsymbol{v}$, $\log q(\boldsymbol{v}|\boldsymbol{x})$
$\boldsymbol{u} \sim q(\boldsymbol{u}|\boldsymbol{x})$
$\boldsymbol{v}_{\boldsymbol{x}} = \boldsymbol{u}_{\boldsymbol{x}}$
$\boldsymbol{v}_{-\boldsymbol{x}} = \text{threshold}(\boldsymbol{u}_{-\boldsymbol{x}}, \boldsymbol{x})$
$\log q(\boldsymbol{v}|\boldsymbol{x}) = \log q(\boldsymbol{u}|\boldsymbol{x}) - \log|\det \mathrm{d}\boldsymbol{v}/\,\mathrm{d}\boldsymbol{u}|$

Table 4: Performance of different dequantization methods on squares and cityscapes dataset, in bits per pixel, lower is better.

Cityscapes	ELBO	IWBO
Round / Unif. (Uria et al., 2013)	1.010	0.930
Round / Var. (Ho et al., 2019)	0.334	0.315
Argmax / Softplus thres. (ours)	**0.303**	**0.290**
Argmax / Gumbel dist. (ours)	0.365	0.341
Argmax / Gumbel thres. (ours)	**0.307**	**0.287**
Multinomial Diffusion (ours)	0.305	

Argmax Flows

- Modeling $q_\theta(v|x)$
- Gumbel: $P_{\text{Gumbel}}(\text{argmax}\,\boldsymbol{v} = i) = \frac{\exp\phi_i}{\sum_j \exp\phi_j}$
- Location parameter $\phi \leftarrow \text{NN}(x)$
- $v_x = \text{Gumbel}(\phi_{\max})$ where $\phi_{\max} = \log\sum_i \exp\phi_i$
- $v_{-x} = \text{TruncGumbel}(\phi_{-x}, T)$ where $T = v_x$

Algorithm 4 Gumbel-based $q(\boldsymbol{v}|\boldsymbol{x})$

Input: $\boldsymbol{x}, \boldsymbol{\phi}$
Output: $\boldsymbol{v}, \log q(\boldsymbol{v}|\boldsymbol{x})$
$\phi_{\max} = \log\sum_i \exp\phi_i$
$\boldsymbol{v}_{\boldsymbol{x}} \sim \text{Gumbel}(\phi_{\max})$
$\boldsymbol{v}_{-\boldsymbol{x}} \sim \text{TruncGumbel}(\boldsymbol{\phi}_{-\boldsymbol{x}}, \boldsymbol{v}_{\boldsymbol{x}})$
$\log q(\boldsymbol{v}|\boldsymbol{x}) = \log \text{Gumbel}(\boldsymbol{v}_{\boldsymbol{x}}|\phi_{\max})$
$\quad + \log \text{TruncGumbel}(\boldsymbol{v}_{-\boldsymbol{x}}|\boldsymbol{\phi}_{-\boldsymbol{x}}, \boldsymbol{v}_{\boldsymbol{x}})$

Table 4: Performance of different dequantization methods on squares and cityscapes dataset, in bits per pixel, lower is better.

Cityscapes	ELBO	IWBO
Round / Unif. (Uria et al., 2013)	1.010	0.930
Round / Var. (Ho et al., 2019)	0.334	0.315
Argmax / Softplus thres. (ours)	**0.303**	**0.290**
Argmax / Gumbel dist. (ours)	0.365	0.341
Argmax / Gumbel thres. (ours)	**0.307**	**0.287**
Multinomial Diffusion (ours)	0.305	

Multinomial Diffusion

Diffusion Models

- DDPMs
- Reverse process

$$p_\theta(x_{0:T}) \coloneqq p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)\,, \qquad p_\theta(x_{t-1}|x_t) \coloneqq \mathcal{N}\big(x_{t-1};\mu_\theta(x_t,t),\Sigma_\theta(x_t,t)\big)$$

- Forward process or diffusion process

$$q(x_{1:T}|x_0) \coloneqq \prod_{t=1}^{T} q(x_t|x_{t-1})\,, \qquad q(x_{t-1}|x_t) \coloneqq \mathcal{N}\Big(x_t;\sqrt{1-\beta_t}x_{t-1},\beta_t I\Big)$$

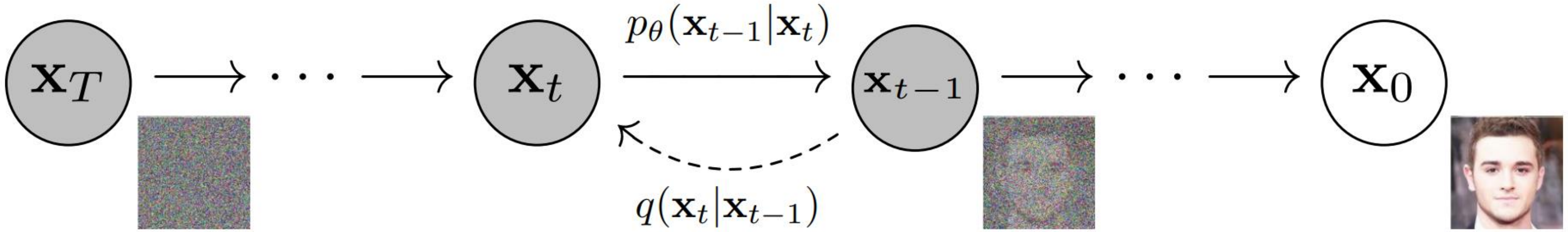


Figure 2: The directed graphical model considered in this work.

Sohl-Dickstein, Jascha, et al. "Deep unsupervised learning using nonequilibrium thermodynamics." *International Conference on Machine Learning*. PMLR, 2015.

Multinomial Diffusion

- We define the multinomial diffusion process using a categorical distribution that has a $\boldsymbol{\beta_t}$ **chance of resampling a category uniformly.**
- $q(x_t|x_{t-1}) = \mathcal{C}(x_t|(1-\beta_t)x_{t-1} + \beta_t/K)$

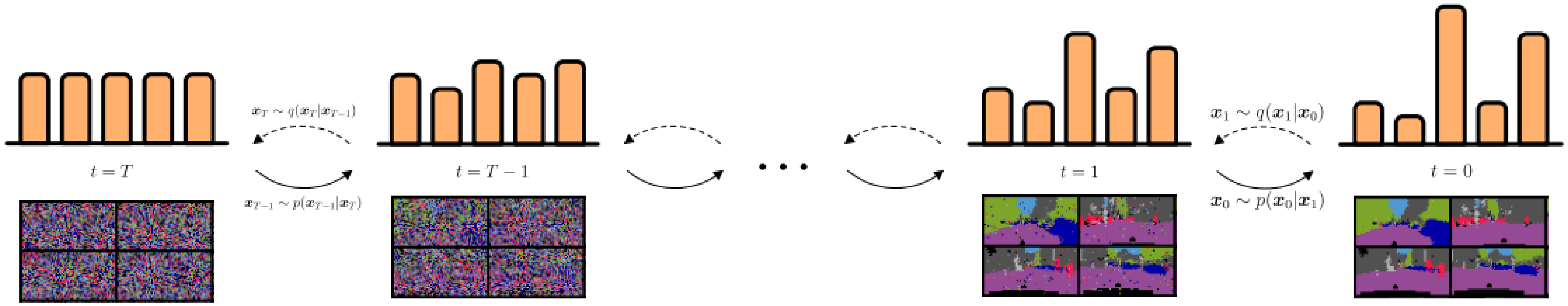


Figure 2: Overview of multinomial diffusion. A generative model $p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ learns to gradually denoise a signal from left to right. An inference diffusion process $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$ gradually adds noise form right to left.

Results

Table 3: Comparison of different methods on `text8` and `enwik8`. Results are reported in negative log-likelihood with units bits per character (bpc) for `text8` and bits per raw byte (bpb) for `enwik8`.

Model type	Model	text8 (bpc)	enwik8 (bpb)
ARM	64 Layer Transformer (Al-Rfou et al., 2019)	1.13	1.06
	TransformerXL (Dai et al., 2019)	1.08	0.99
VAE	AF/AF* (AR) (Ziegler and Rush, 2019)	1.62	1.72
	IAF / SCF* (Ziegler and Rush, 2019)	1.88	2.03
	CategoricalNF (AR) (Lippe and Gavves, 2020)	1.45	-
Generative Flow	Argmax Flow, AR (ours)	1.39	1.42
	Argmax Coupling Flow (ours)	1.82	1.93
Diffusion	Multinomial Text Diffusion (ours)	1.72	1.75

$\star$ Results obtained by running code from the official repository for the `text8` and `enwik8` datasets.

Results

```
 that the role of tellings not be required also action characters passe
d on constitution ahmad a nobilitis first be closest to the cope and dh
ur and nophosons she criticized itm specifically on august one three mo
vement and a renouncing local party of exte

nt is in this meant the replicat today through the understanding elemen
t thinks the sometimes seven five his final form of contair you are lot
ur and me es to ultimately this work on the future all all machine the
silon words thereis greatly usaged up not t
```

(a) Samples from Multinomial Text Diffusion.

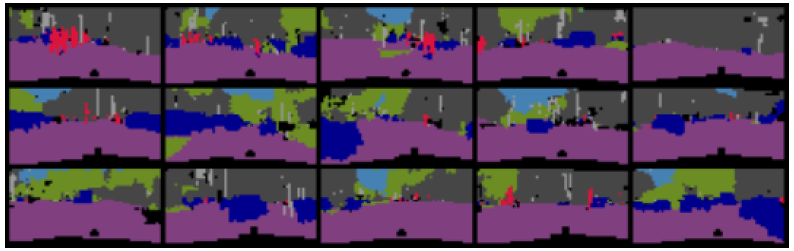
```
 heartedness frege thematically infered by the famous existence of a fu
nction f from the laplace definition we can analyze a definition of bin
ary operations with additional size so their functionality cannot be re
viewed here there is no change because its

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 Argmax AR Flow.

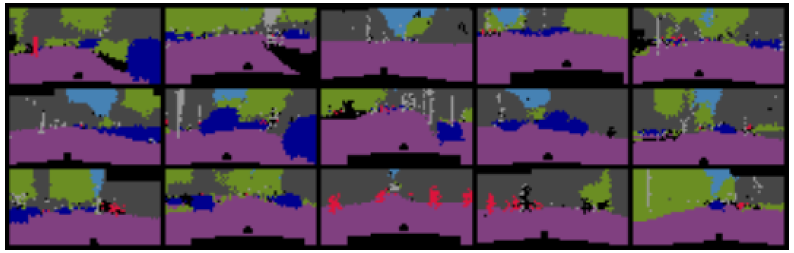
```
ns fergenur d alpha and   le heigu man notabhe leglon lm n two six a gg
opa movement as sympathetic dutch the term bilirubhah acquired the bava
rian cheeh segt thmamouinaire vhvinus lihnos ineoneartis or medical iod
ine the rave wesp published harsy varb hhgh

 danibah or manuccha but calpere that  of the moisture soods and dristi
ng attempt to cause any moderator called lk brown or totpdngs is usuall
y able to nus and hockecrits borel qbisupnias section rybancase untecce
mentation anymore the motion of plays on qr
```

(c) Samples from Argmax Coupling Flow.

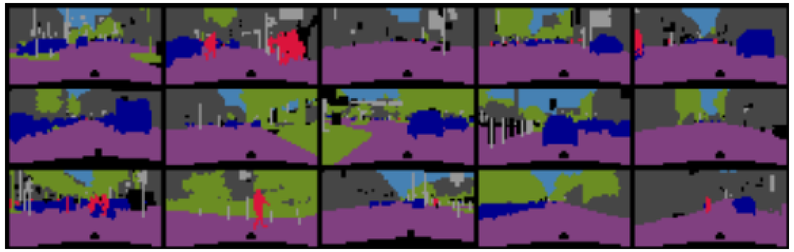
Figure 3: Samples from models, `text8`.



(a) Samples from the Argmax Flow.



(b) Samples from the Multinomial Diffusion model.



(c) Cityscapes data.

Figure 4: Samples from models, cityscapes.

```
mexico city the aztec stadium estadio azteca home of club america is on
e of the world s largest stadiums with capacity to seat approximately o
ne one zero zero zero zero fans mexico hosted the football world cup in
 one nine seven zero and one nine eight six
```

(a) Ground truth sequence from `text8`.

```
mexico citi the aztec stadium estadio azteca home of clup amerika is on
e of the world s largest stadioms with capakity to seat approsimately o
ne one zeto zero zero zero fans mexico hosted the footpall wolld cup in
 one nine zeven zero and one nyne eiggt six
```

(b) Corrupted sentence.

```
mexico city the aztec stadium estadio aztecs home of club america is on
e of the world s largest stadiums with capacity to seat approximately o
ne one zero zero zero zero fans mexico hosted the football world cup in
 one nine seven zero and one nine eight six
```

(c) Suggested, prediction by the model.

Figure 5: Spell checking with Multinomial Text Diffusion.

References

- Hoogeboom, Emiel, et al. "Argmax flows and multinomial diffusion: Learning categorical distributions." *Advances in Neural Information Processing Systems* 34 (2021): 12454-12465.
- Didrik Nielsen, Argmax Flows and Multinomial Diffusion: Learning Categorical Distributions, https://www.youtube.com/watch?v=150ceiAVDCY