Scalable Variational Inference

Lets start with fitting $p(x)$ into a dataset

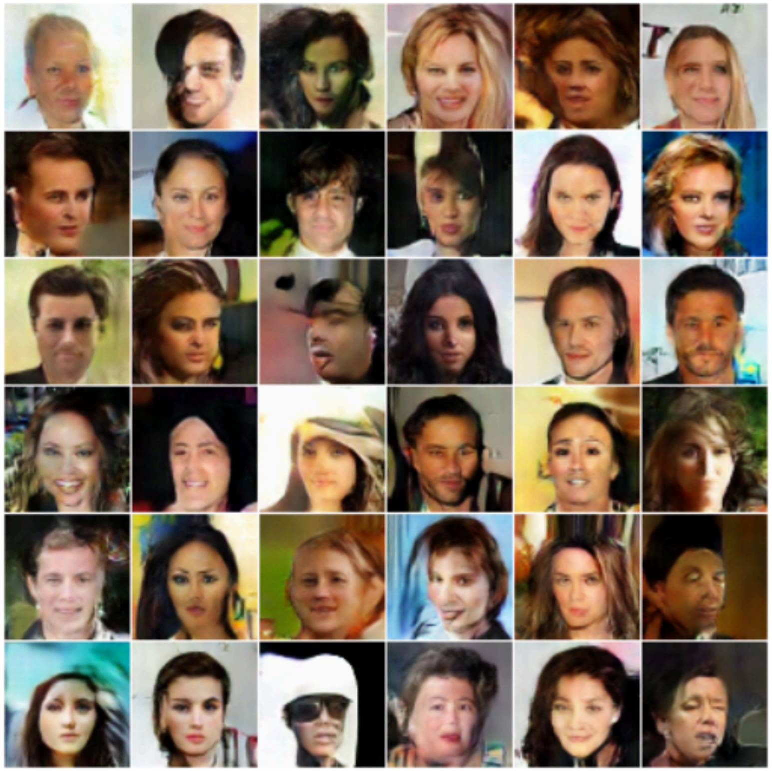
Scalable Variational Inference

Lets start with fitting $p(x)$ into a dataset

But why do we need it?

Why model $p(x)$

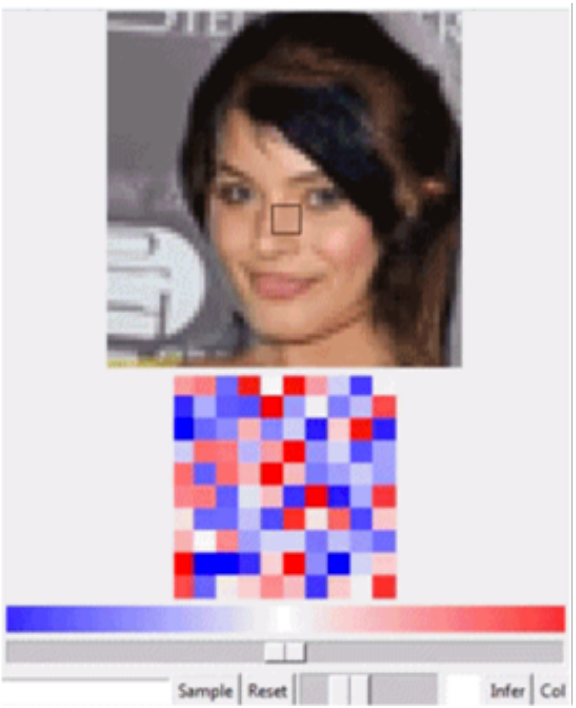
- Generate new data

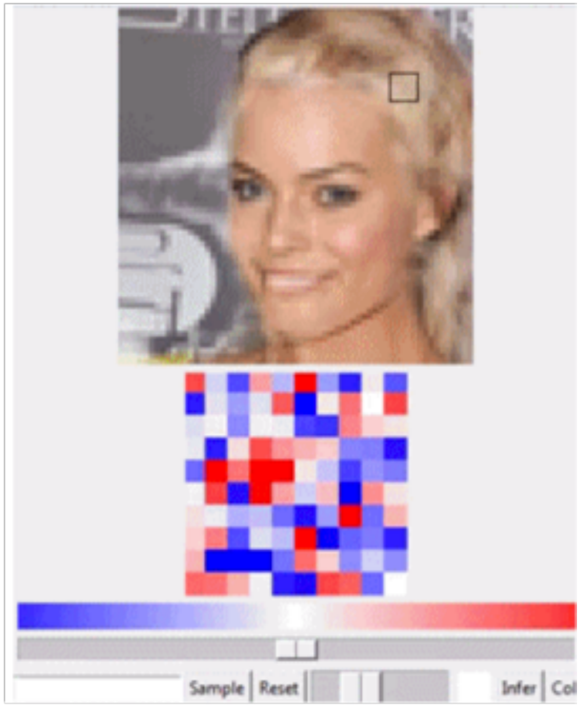


[Zhao et. al. Energy-based generative adversarial network]

Why model $p(x)$

- Generate new data





[**DL Cade**, https://petapixel.com/2016/09/27/neural-photo-editor-like-fully-automatic-photoshop/]

Why model $p(x)$

- Generate new data

Why model $p(x)$

- Generate new data

- Detect anomalies and outliers (e.g. fraud detection)

Why model $p(x)$

- Generate new data
- Detect anomalies and outliers (e.g. fraud detection)
- Work with missing data

Why model $p(x)$

- Generate new data
- Detect anomalies and outliers (e.g. fraud detection)
- Work with missing data
- Represent your data in a nice way (e.g. model p(molecule) to search for drugs)

How to model $p(x)$

- $\log \widehat{p}(x) = \mathrm{CNN}(x)$

How to model $p(x)$

- $\log \widehat{p}(x) = \mathrm{CNN}(x)$

$$p(x) = \frac{\exp(\mathrm{CNN}(x))}{Z}$$

How to model $p(x)$

- $\log \widehat{p}(x) = \text{CNN}(x)$

Infeasible

$$p(x) = \frac{\exp(\text{CNN}(x))}{Z}$$

How to model $p(x)$

- $\log \widehat{p}(x) = \mathrm{CNN}(x)$

Infeasible

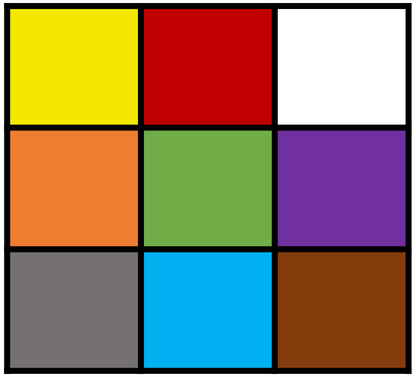
- Use the chain rule

How to model $p(x)$

- $\log \widehat{p}(x) = \mathrm{CNN}(x)$

Infeasible

- Use the chain rule

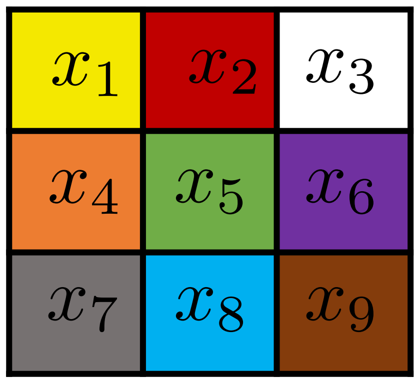


How to model $p(x)$

- $\log \widehat{p}(x) = \mathrm{CNN}(x)$

Infeasible

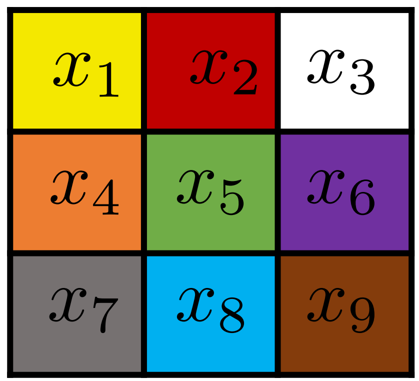
- Use the chain rule



How to model $p(x)$

- $\log \widehat{p}(x) = \mathrm{CNN}(x)$ Infeasible

- Use the chain rule



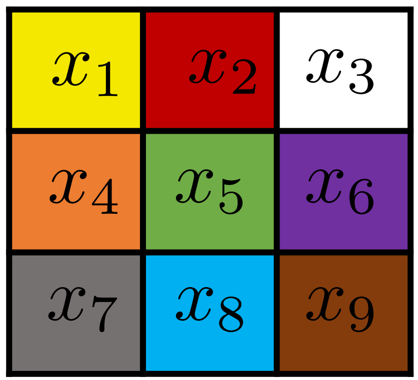
$$p(x_1, \ldots, x_d)$$
$$= p(x_1)p(x_2 \mid x_1) \ldots p(x_d \mid x_1, \ldots, x_{d-1})$$

How to model $p(x)$

- $\log \widehat{p}(x) = \mathrm{CNN}(x)$

Infeasible

- Use the chain rule



[Oord, Aaron van den, Nal Kalchbrenner, and Koray Kavukcuoglu.

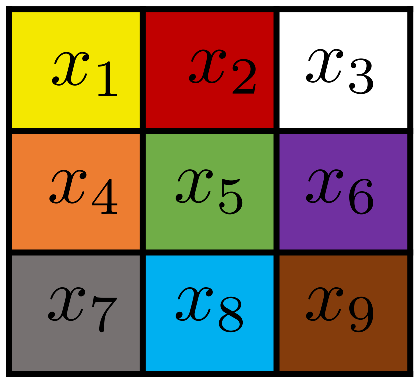
"Pixel recurrent neural networks." (2016)]

$$p(x_1, \ldots, x_d)$$
$$= p(x_1) p(x_2 \mid x_1) \ldots p(x_d \mid x_1, \ldots, x_{d-1})$$
$$p(x_k \mid x_1, \ldots, x_{k-1}) = \mathrm{RNN}(x_1, \ldots, x_{k-1})$$

How to model $p(x)$

- $\log \widehat{p}(x) = \mathrm{CNN}(x)$ — Infeasible

- Use the chain rule — Cool, but slow to generate



[Oord, Aaron van den, Nal Kalchbrenner, and Koray Kavukcuoglu.

"Pixel recurrent neural networks." (2016)]

$$p(x_1, \ldots, x_d)$$
$$= p(x_1)p(x_2 \mid x_1) \ldots p(x_d \mid x_1, \ldots, x_{d-1})$$
$$p(x_k \mid x_1, \ldots, x_{k-1}) = \mathrm{RNN}(x_1, \ldots, x_{k-1})$$

How to model $p(x)$

- $\log \widehat{p}(x) = \mathrm{CNN}(x)$ — Infeasible
- Use the chain rule — Cool, but slow to generate
- $p(x_1, \ldots, x_d) = p(x_1) \ldots p(x_d)$

How to model $p(x)$

- $\log \widehat{p}(x) = \mathrm{CNN}(x)$

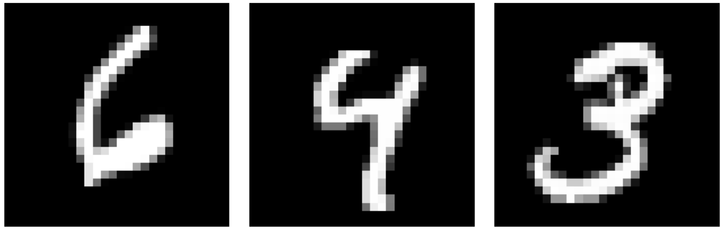
Infeasible

- Use the chain rule

Cool, but slow to generate

- $p(x_1, \ldots, x_d) = p(x_1) \ldots p(x_d)$

Data:

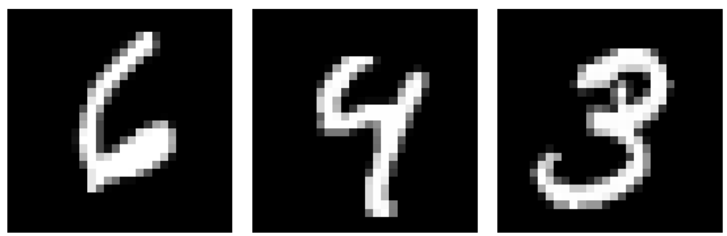


. . .

How to model $p(x)$

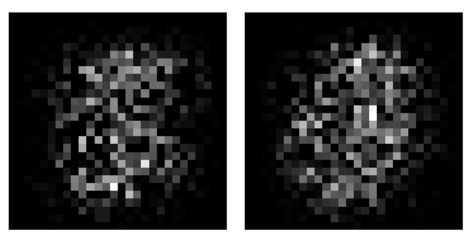
- $\log \widehat{p}(x) = \text{CNN}(x)$ Infeasible
- Use the chain rule Cool, but slow to generate
- $p(x_1, \ldots, x_d) = p(x_1) \ldots p(x_d)$ Too restrictive

Data:



. . .

Samples:



How to model $p(x)$

- $\log \widehat{p}(x) = \text{CNN}(x)$ — Infeasible
- Use the chain rule — Cool, but slow to generate
- $p(x_1, \ldots, x_d) = p(x_1) \ldots p(x_d)$ — Too restrictive
- Mixture of several Gaussians (GMM)

How to model $p(x)$

- $\log \widehat{p}(x) = \mathrm{CNN}(x)$ — Infeasible
- Use the chain rule — Cool, but slow to generate
- $p(x_1, \ldots, x_d) = p(x_1) \ldots p(x_d)$ — Too restrictive
- Mixture of several Gaussians (GMM) — Still too restrictive

How to model $p(x)$

- $\log \widehat{p}(x) = \mathrm{CNN}(x)$ — Infeasible
- Use the chain rule — Cool, but slow to generate
- $p(x_1, \ldots, x_d) = p(x_1) \ldots p(x_d)$ — Too restrictive
- Mixture of several Gaussians (GMM) — Still too restrictive
- Mixture of infinitely many Gaussians

How to model $p(x)$

- $\log \widehat{p}(x) = \mathrm{CNN}(x)$ — Infeasible
- Use the chain rule — Cool, but slow to generate
- $p(x_1, \dots, x_d) = p(x_1) \dots p(x_d)$ — Too restrictive
- Mixture of several Gaussians (GMM) — Still too restrictive
- Mixture of infinitely many Gaussians

$$p(x) = \int p(x \mid t)\, p(t) dt$$