CSC6137: Generative Models: Assignment 1

Tao Chujun 120090211

October 29, 2023

Collaborator: 223040252 Zhijun Liu

1 Affine and Elementwise Flows

You are asked to use normalizing flow to do the density estimation. Consider a base distribution as Gaussian,

$$p(\boldsymbol{u}) = \mathcal{N}(\boldsymbol{u} \mid \boldsymbol{\mu}, \sigma \boldsymbol{I})$$

where $\boldsymbol{u} \in R^d$

1. Consider an affline flow $\mathbf{A}\boldsymbol{u} + \boldsymbol{b}$ with $\mathbf{A} \in R^{d \times d}$. Please write out the loss function for the density estimation problem and clarify the constraints you need to consider.

Ans 1: For an affine flow $\mathbf{A}\boldsymbol{u} + \boldsymbol{b}$, the loss function for the density estimation problem is the negative log-likelihood of the observed data. Given the change of variables formula, the density of the transformed variable $\boldsymbol{z} = \mathbf{A}\boldsymbol{u} + \boldsymbol{b}$ is given by

$$p(\boldsymbol{z}) = p(\boldsymbol{u}) \left| \det \frac{d\boldsymbol{u}}{d\boldsymbol{z}} \right|$$

where $\frac{d\boldsymbol{u}}{d\boldsymbol{z}}$ is the Jacobian of the inverse transformation. For an affine transformation, this is simply $\mathbf{A}^{-1}$, so the determinant is $|\det \mathbf{A}^{-1}| = 1/|\det \mathbf{A}|$. Therefore, the loss function is

$$\mathcal{L} = -\log p(\boldsymbol{z}) = -\log p(\boldsymbol{u}) + \log |\det \mathbf{A}|$$

The constraint we need to consider is that $\mathbf{A}$ must be invertible, i.e., $\det \mathbf{A} \neq 0$. Since in this question:

$$p_Z(z) = \mathcal{N}(z \mid \mu, \sigma I) : \mathbb{R}^d \to (0, \infty)$$

Therefore,

$$p_X(x) = \mathcal{N}(x \mid A\mu + b, \sigma A) : \mathbb{R}^d \to (0, \infty)$$

$$\mathcal{L}_(x) = -\log p_X(x) = -\log \mathcal{N}(x \mid A\mu + b, \sigma A)$$

2. Consider an elementwise flow $\sigma(u) = 1/(1 + \exp(-u))$. Please write out the loss function for the density estimation problem and clarify the constraints you need to consider.

Ans 2: For an elementwise flow $\sigma(u) = 1/(1+\exp(-u))$, the Jacobian of the inverse transformation is diagonal, with entries given by the derivative of the inverse sigmoid function. Therefore, the loss function is

$$\mathcal{L} = -\log p(\boldsymbol{z}) = -\log p(\boldsymbol{u}) + \sum_i \log\left|\frac{d\sigma^{-1}(z_i)}{dz_i}\right| = -\log p(\boldsymbol{u}) + \sum_i \log|\sigma(z_i)\cdot(1-\sigma(z_i)|$$

the sigmoid function is invertible.

3. We consider a family of transformations of the form:

$$f(\mathbf{u}) = \mathbf{u} + \mathbf{r}h\left(\mathbf{w}^\top\mathbf{u} + b\right)$$

where $\lambda = \left\{\mathbf{w} \in \mathbb{R}^d, \mathbf{r} \in \mathbb{R}^d, b \in \mathbb{R}\right\}$ are free parameters and $h(\cdot)$ is a smooth elementwise non-linearity, with derivative $h'(\cdot)$. Consider a density $q_K(\mathbf{u})$ obtained by transforming an arbitrary initial density $q_0(\mathbf{u}) := p(\mathbf{u})$ through the sequence of maps f_k of the form defined as above:

$$\mathbf{u}_K = f_K \circ f_{K-1} \circ \ldots \circ f_1(\mathbf{u})$$

(a) Please write out the loss function for the density estimation problem and clarify the constraints you need to consider.

Ans (a): The determinant of the Jacobian is given by

$$\det\left(\mathbf{I} + \mathbf{r}h'(\mathbf{w}^\top\mathbf{u} + b)\mathbf{w}^\top\right)$$

$\mathbf{I}$ is the identity matrix and $h'(\cdot)$ is the derivative of $h(\cdot)$. Therefore, the loss function is

$$\mathcal{L} = -\log p(\boldsymbol{z}) = -\log p(\boldsymbol{u}) + \sum_{k=1}^{K} \log\left|\det\left(\mathbf{I} + \mathbf{r_k}h'(\mathbf{w_k}^\top\mathbf{u_{k-1}} + b)\mathbf{w_k}^\top\right)\right|$$

The constraint we need to consider is that the transformation must be invertible, which requires the determinant of the Jacobian to be nonzero.

(b) When using $h(x) = \tanh(x)$, please verify that a sufficient condition for $f(\mathbf{u})$ to be invertible is that $\mathbf{w}^\top\mathbf{r} \geq -1$.

Ans (b): When using $h(x) = \tanh(x)$, the derivative is $h'(x) = 1 - \tanh^2(x)$. The condition for $f(\mathbf{u})$ to be invertible is that the Jacobian matrix is invertible, which requires its determinant to be nonzero. The determinant of the Jacobian is

$$\det\left(\mathbf{I} + \mathbf{r}(1 - \tanh^2(\mathbf{w}^\top\mathbf{u} + b))\mathbf{w}^\top\right)$$

With $1 \leq \tanh'(x) \leq 1$, $\det\left|\frac{\partial f(u)}{\partial u}\right| > 1 - 1 = 0$ when $w^\top u + b \neq 0$

(c) For this mapping, what is the computational complexity of computing the log-det-Jacobian term (in terms of d)?

Ans (c): The computational complexity of computing the log-det-Jacobian term is $O(d^2)$, because it involves computing the determinant of a $d \times d$ matrix. Since we have k layers, so the overall computational complexity is $O(kd^2)$

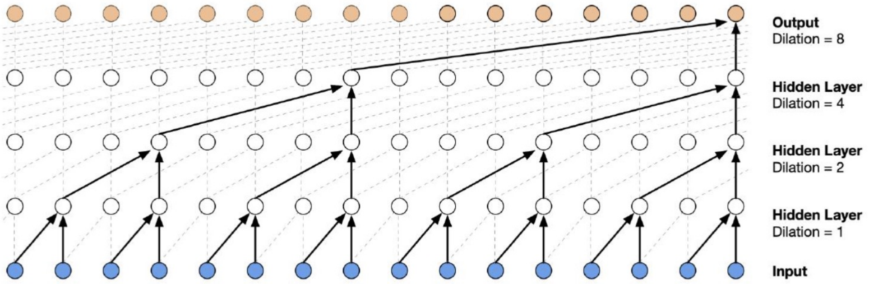


Figure 1: WaveNet

2 Parallel WaveNet

In this question, you are asked to understand the basic ideas of WaveNet and Parallel WaveNet.

1. WaveNet adopts autoregressive networks to model the joint distribution of highdimensional data as a product of conditional distributions using the probabilistic chain-rule:

$$p(\boldsymbol{x})=\prod_{t=1}^{T} p\left(x_t \mid \boldsymbol{x}_{1:t-1} ; \boldsymbol{\theta}\right)$$

where x_t is the t-th variable of $\boldsymbol{x}$ and $\boldsymbol{\theta}$ are the parameters of the autoregressive model. The conditional distributions are usually modeled with a neural network that receives $\boldsymbol{x}_{1:t-1}$ as input and outputs a distribution over possible x_t.

As illustrated in Fig. 1, WaveNet employs CNN to construct the conditional probability that receives $\boldsymbol{x}_{1:t-1}$ as input and outputs a distribution over possible x_t. Specifically, we have

$$p\left(x_t \mid \boldsymbol{x}_{1:t-1} ; \boldsymbol{\theta}\right):=\mathcal{M}\left(x_t \mid \mathcal{S}\left(\varphi\left(\sum_{\tau=1}^{t-k} \boldsymbol{w}^{\top} \boldsymbol{x}_{\tau:\tau+k}\right)\right)\right)$$

Please describe the key ideas in using CNN to compute the conditional probability, and describe the meaning of the notations $\boldsymbol{w}, \mathcal{S}(\cdot)$, and $\mathcal{M}(\cdot)$.

Ans 1: The conditional probability we want to compute is

$$p\left(x_t \mid \boldsymbol{x}_{1:t-1} ; \boldsymbol{\theta}\right)$$

This conditional probability distribution is modeled by several convolution layers. The key is to use a filter $\boldsymbol{w}$ to do convolution operation. The function $\varphi(\cdot)$ adds a nonlinearity to convolution results, and $\mathcal{S}(\cdot)$ is a softmax layer that outputs a probability distribution over possible values of the current audio sample x_t. $\mathcal{M}(\cdot)$ is a gating mechanism that controls the flow of information through the network.

$$f\left(x_t\right)=\operatorname{sign}(x_t) \frac{\ln \left(1+\mu\left|x_t\right|\right)}{\ln (1+\mu)},$$

2. Please explain why the WaveNet can perform the training by using the causal mask in parallel but can only perform the data generating in a sequential manner, which is time-consuming.

Ans 2: During the training process, we train the model using the whole audio training dataset. For each data output, we already know its previous data input. While during the inference process, the next data output need to wait until all its previous data input is generated so that it can be generated.

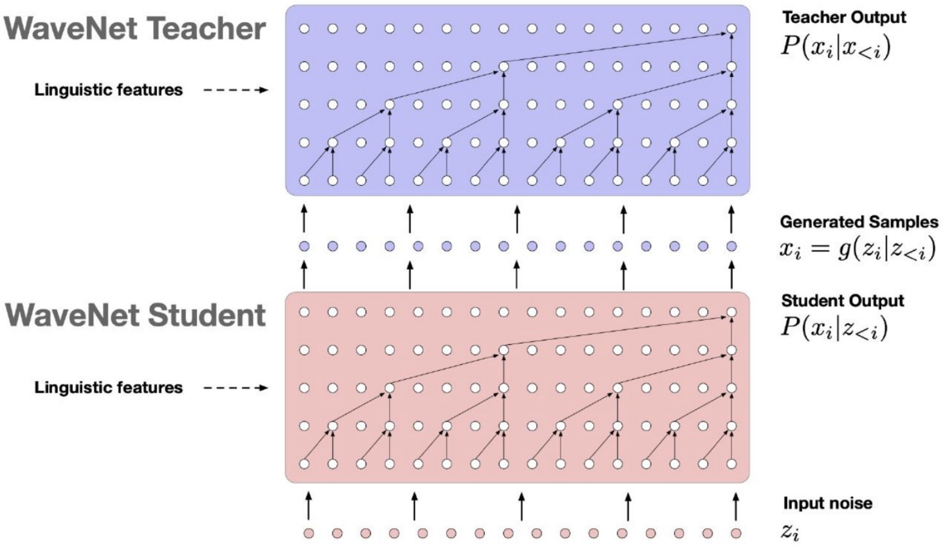


Figure 2: Parallel WaveNet

3. Parallel WaveNet is proposed to address the time efficiency issue faced by data generation as illustrated in Fig. 2. Next, we will walk you through the main ideas.

(a) Parallel WaveNet adopts a student-teacher paradigm to perform the neural network distillation. Specifically, the already ”well-trained” WaveNet is treated as a ”teacher” and an inverse autoregressive flow (IAF) based parallel WaveNet is treated as a student. For IAF, $x_t = f\left(\boldsymbol{z}_{\leq t}\right)$.

Specifically, a random sample is first drawn from $\boldsymbol{z} \sim p_Z(\boldsymbol{z})$ (a simple distribution) and is then transformed as follows:

$$x_t = z_t \cdot s\left(\boldsymbol{z}_{<t}, \boldsymbol{\theta}\right) + \mu\left(\boldsymbol{z}_{<t}, \boldsymbol{\theta}\right)$$

where μ and s are outputs by the network with parameter $\boldsymbol{\theta}$. Here, we are using the same convolutional network structure as the original WaveNet.

Please explain why compared with autoregressive flow, inverse autoregressive flow is inherently parallel in forward direction f. And also explain why its inverse direction f^{-1} is inherently sequential.

Ans (a): IAF flow is parallel in forward direction since for each $\boldsymbol{x_t}$, it depends on all previous latent variables $\boldsymbol{z_{<t}}$. All $\boldsymbol{z_t}$ are arranged in a certain way at the beginning. Therefore, $\boldsymbol{x_t}$ can be generated in parallel. IAF flow infers what it would have output at previous timesteps based on the current input. In the iverse direction, this requires the generation of the latent variables in a certain order, and the process is sequential. i.e.

$$z_t = \frac{x_t - \mu_\theta\left(z_{<t}\right)}{s_\theta\left(z_{<t}\right)}$$

(b) Please provide the conditions to guarantee that the flow is a bijection. Please describe a reasonable model for $s(\boldsymbol{z}_{<t}, \boldsymbol{\theta})$ and $\mu(\boldsymbol{z}_{<t}, \boldsymbol{\theta})$, as you like. Compute the determinant of the Jacobian.

Ans (b): Bijection means for each input x, there must be a unique output y, and for each output y, there must be a unique input x. To guarantee that the flow is a bijection, the transformation function must be invertible and have a non-zero determinant. The invertibility condition can be satisfied by ensuring that the function is monotonic. The non-zero determinant condition can be satisfied by ensuring that the Jacobian matrix of the transformation function has a non-zero determinant.

(c) Now you should understand that, parallel WaveNet is slow in training and fast in data generation, whereas WaveNet is fast in training and slow in data generation. Now let's understand that the student-teacher paradigm can train the parallel WaveNet by leveraging the benefits from the both worlds.

Given a parallel WaveNet student $p_S(\boldsymbol{x})$ and WaveNet teacher $p_T(\boldsymbol{x})$, which has been trained on a dataset of audio, we define the Probability Density Distillation loss as follows:

$$D_{\mathrm{KL}}\left(P_S \| P_T\right) = H\left(P_S, P_T\right) - H\left(P_S\right)$$

where D_{KL} is the Kullback-Leibler divergence, and $H\left(P_S, P_T\right)$ is the cross-entropy between the student P_S and teacher P_T, and $H\left(P_S\right)$ is the entropy of the student distribution.

Please using the definition of KL and entropy to verify the above expression. One should know that when the KL divergence becomes zero, the student distribution has fully recovered the teacher's distribution.

Ans (c): The KL divergence between two probability distributions P and Q is:

$$D_{\mathrm{KL}}(P \| Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

The entropy of a probability distribution P is:

$$H(P) = -\sum_x P(x) \log P(x)$$

The cross-entropy between two probability distributions P and Q is :

$$H(P, Q) = -\sum_x P(x) \log Q(x)$$

For the given equation for the Probability Density Distillation loss:

$$D_{\mathrm{KL}}(P_S \| P_T) = H(P_S, P_T) - H(P_S)$$

Substituting the definitions of KL divergence, cross-entropy, and entropy, we have:

$$\sum_x P_S(x) \log \frac{P_S(x)}{P_T(x)} = -\sum_x P_S(x) \log P_T(x) - (-\sum_x P_S(x) \log P_S(x))$$

$$\sum_x P_S(x) \log \frac{P_S(x)}{P_T(x)} = \sum_x P_S(x) \log P_S(x) - \sum_x P_S(x) \log P_T(x)$$

Obviously, the equation holds. When the KL divergence becomes zero, $P_S(x) = P_T(x)$ for every x. In other words, the student distribution is equivalent to the teacher's distribution.

(d) Please describe how to compute $H\left(P_S\right)$ based on Eq. (1). Please explain do we really need to sample $\{x_t\}$ to compute $H\left(P_S\right)$.

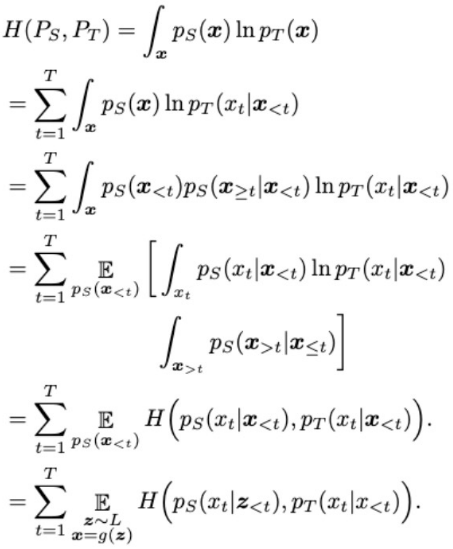
Ans (d): Sampling $\{x_t\}$ to compute $H\left(P_S\right)$ is necessary. We need to use Monte-carlo estimation to compute $H\left(P_S\right)$

(e) Please derive by your-self and show the following conclusion

$$H\left(P_S, P_T\right)=\sum_{t=1}^T \underset{\substack{z \sim L \\ x=g(z)}}{\mathbb{E}} H\left(p_S\left(x_t \mid z_{<t}\right), p_T\left(x_t \mid x_{<t}\right)\right)$$

Given the derivation, please explain how to compute $H\left(P_S, P_T\right)$ using the generated samples from the student. Please explain that why both ${}_S\left(x_t \mid z_{<t}\right)$ and $p_T\left(x_t \mid x_{<t}\right)$ can be computed in parallel.

Ans (e): the derivation is shown below:



$$\begin{aligned}
H(P_S, P_T) &= \int_{\boldsymbol{x}} p_S(\boldsymbol{x}) \ln p_T(\boldsymbol{x}) \\
&= \sum_{t=1}^T \int_{\boldsymbol{x}} p_S(\boldsymbol{x}) \ln p_T(x_t|\boldsymbol{x}_{<t}) \\
&= \sum_{t=1}^T \int_{\boldsymbol{x}} p_S(\boldsymbol{x}_{<t}) p_S(\boldsymbol{x}_{\geq t}|\boldsymbol{x}_{<t}) \ln p_T(x_t|\boldsymbol{x}_{<t}) \\
&= \sum_{t=1}^T \underset{p_S(\boldsymbol{x}_{<t})}{\mathbb{E}} \Big[\int_{x_t} p_S(x_t|\boldsymbol{x}_{<t}) \ln p_T(x_t|\boldsymbol{x}_{<t}) \\
&\qquad \int_{\boldsymbol{x}_{>t}} p_S(\boldsymbol{x}_{>t}|\boldsymbol{x}_{\leq t}) \Big] \\
&= \sum_{t=1}^T \underset{p_S(\boldsymbol{x}_{<t})}{\mathbb{E}} H\Big(p_S(x_t|\boldsymbol{x}_{<t}), p_T(x_t|\boldsymbol{x}_{<t})\Big). \\
&= \sum_{t=1}^T \underset{\substack{\boldsymbol{z} \sim L \\ \boldsymbol{x}=g(\boldsymbol{z})}}{\mathbb{E}} H\Big(p_S(x_t|\boldsymbol{z}_{<t}), p_T(x_t|x_{<t})\Big).
\end{aligned}$$

To compute $H\left(P_S, P_T\right)$ using generated samples from the student, we can first draw multiple different samples of x_t from $p_S(x_t|z_{<t})$ for each timestep t. Then, for each sample of x_t, we can compute the corresponding conditional distribution $p_T(x_t|x_{<t})$ using the teacher model. Finally, we can compute the cross-entropy between the student and teacher distributions for each sample of x_t, and take the average over all samples and timesteps to obtain the estimate of $H\left(P_S, P_T\right)$.

Both $p_S\left(x_t \mid z_{<t}\right)$ and $p_T\left(x_t \mid x_{<t}\right)$ can be computed in parallel. From (a) we know $\boldsymbol{x_t}$ is generated in parallel. So the conditional distribution of x_t given $x_{<t}$ using the teacher model can be computed in parallel. We have $\boldsymbol{z_t}$ at the beginning. Therefore, we can compute the conditional distribution of x_t given $z_{<t}$ using the student model also in parallel.