Parallel WaveNet: Fast High-Fidelity Speech Synthesis

Симкин Алексей

6 февраля 2020

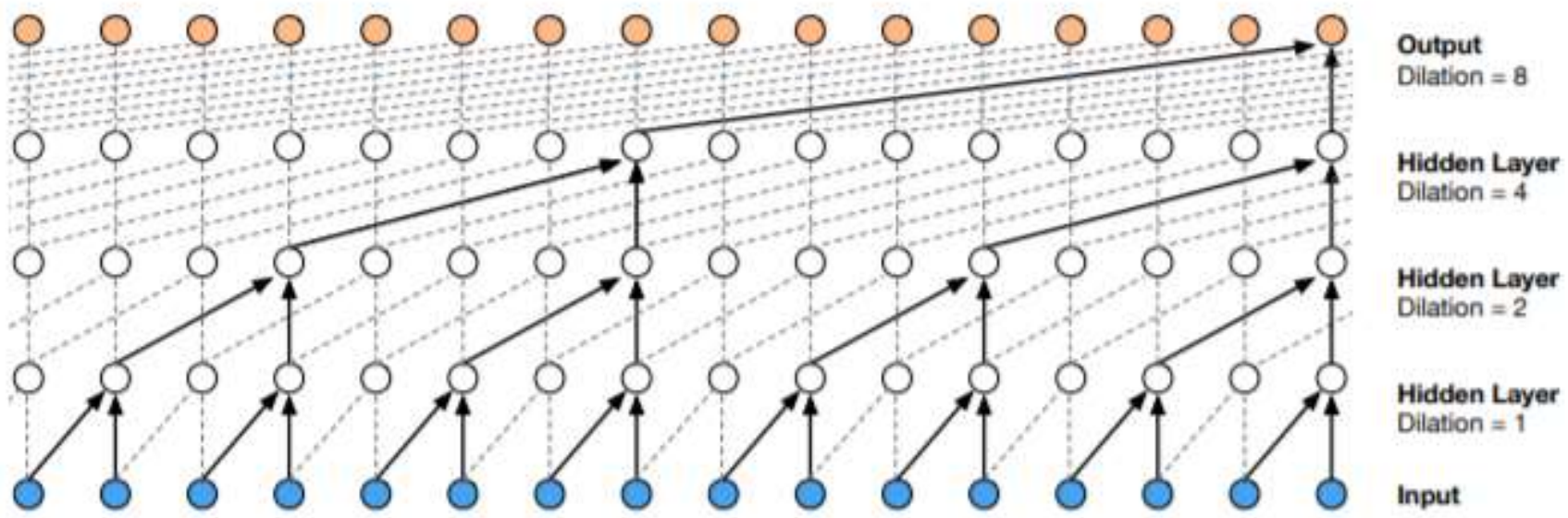
WaveNet

- Авторегресионная генеративная модель

$$p(x) = \prod_t p(x_t | x_{<t}, \theta)$$

WaveNet

- Экспоненциальное увеличение рецептивного поля достигается за счет параметра dilation в свертках.



WaveNet

+ Модель быстро обучается благодаря возможности распараллеливания

- Генерация происходит последовательно, а значит долго

Inverse-autoregressive flows

- Вместо последовательного семплирования на вход подается белый шум, который преобразуется к нужной форме

- Пусть дано простое распределение $p_z(z)$ и сложное $p_x(x)$

- С помощью IAF можно выучить преоразование $x = f(z)$

Inverse-autoregressive flows

- В таком случае

$$\log p_x(x) = \log p_z(z) - \log \left| \frac{dx}{dz} \right|$$

- Поскольку x_t зависит только от $z_{\leq t}$, то Якобиан f – это треугольная матрица
- Значит

$$\log \left| \frac{dx}{dz} \right| = \sum_t \log \frac{\partial f(z_{\leq t})}{\partial z_t}$$

Inverse-autoregressive flows

- Изначально генерируются семплы из логистического распределения $z \sim Logistic(0, I)$

- После чего применяется преобразование

$$x_t = z_t \cdot s(z_{<t}, \theta) + \mu(z_{<t}, \theta)$$

Inverse-autoregressive flows

- На выходе модели получается семпл x, при этом

$$p(x_t|z_{<t},\theta) = Logistic\big(x_t|\mu(z_{<t},\theta), s(z_{<t},\theta)\big)$$

- В качестве $\mu(z_{<t},\theta)$ и $s(z_{<t},\theta)$ можно использовать любую авторегресионную модель

- Например, используемую в оригинальной WaveNet

Inverse-autoregressive flows

- Для улучшения качества может потребоваться провести несколько последовательных итераций
- Выход одной сети используется в качестве входа для следующей
- Авторы используют 4 блока

$$x^0 = z$$

$$x^i = x^{i-1} \cdot s^i + \mu^i$$

Inverse-autoregressive flows

- Параметры финального распределения $p(x_t|z_{<t}, \theta)$ равны:

$$\mu_{tot} = \sum_{i}^{N} \mu^i \left(\prod_{j>i}^{N} s^j \right)$$

$$s_{tot} = \prod_{i}^{N} s_i$$

Где N – число блоков

Обучение модели

- Сначала предобучим обычный WaveNet (teacher network)

- После чего смоделируем распределение полученной модели с помощью IAF (student network)

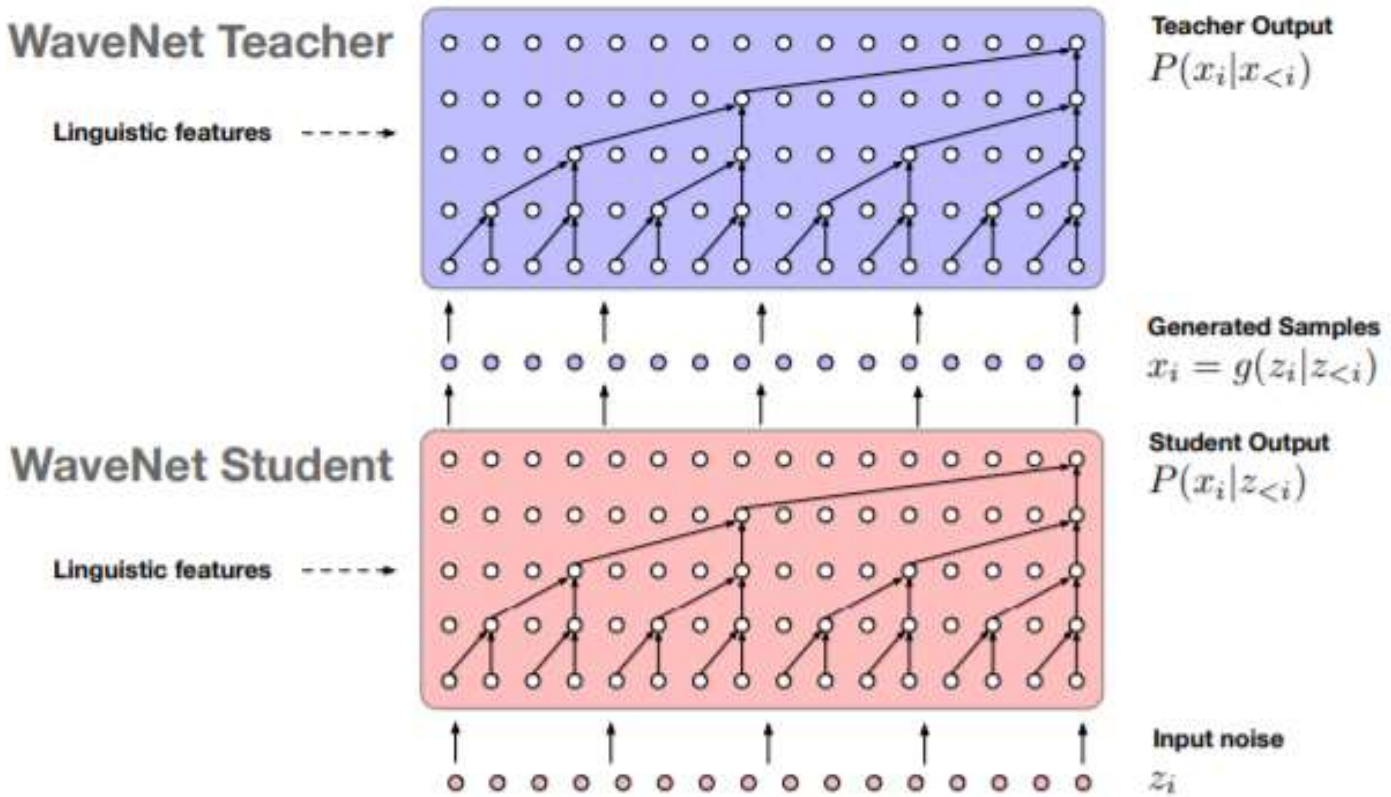
Probability Density Distillation

- Для обучения используется Probability Density Distillation loss

$$KL(P_S||P_T) = H(P_S, P_T) - H(P_S)$$

Где $P_S(x)$ – распределение обучаемой сети, $P_T(x)$ – распределение сети-учителя.

Probability Density Distillation



Probability Density Distillation

- Кросс-энтропия может быть представлена в следующем виде

$$H(P_S) = E_z\left[\sum_{t=1}^{0} \ln s\left(z_{<t}, \theta\right)\right] + 2T$$

$$H(P_s, P_T) = \sum_{t=1}^{T} E_{\rho_s(x_{<t})} H\left(p_S(x_t|x_{<t}), p_T(x_t|x_{<t})\right)$$

Probability Density Distillation

Для улучшения качества используются дополнительные функции потерь:

- Power loss
- Perceptual loss
- Contrastive loss

Power loss

$$\left\|\phi\big(g(z,c)\big)-\phi(y)\right\|^2$$

Где $\phi(x)=|STFT(x)|^2$

Помогает модели использовать различные звуковые частоты с той периодичностью, с которой они используются в речи, позволяя избежать, например, коллапсирования в шепот

Perceptual loss

$$\left\|\phi\big(g(z,c)\big)-\phi(y)\right\|^2$$

Вместо преобразования Фурье используется нейросеть

Позволяет улучшить произношение модели

Contrastive loss

$$KL(P_S(\mathrm{c}_1)||P_T(\mathrm{c}_1)) - \gamma KL(P_S(\mathrm{c}_1)||P_T(\mathrm{c}_2))$$

- Минимизируем расстояния между сетями с одинаковыми дополнительными параметрами(например, голос)
- При этом увеличиваем расстояние, если параметры различны

Результаты модели

Method	Subjective 5-scale MOS
16kHz, 8-bit μ-law, 25h data:	
LSTM-RNN parametric [27]	3.67 ± 0.098
HMM-driven concatenative [27]	3.86 ± 0.137
WaveNet [27]	4.21 ± 0.081
24kHz, 16-bit linear PCM, 65h data:	
HMM-driven concatenative	4.19 ± 0.097
Autoregressive WaveNet	4.41 ± 0.069
Distilled WaveNet	4.41 ± 0.078

Table 1: Comparison of WaveNet distillation with the autoregressive teacher WaveNet, unit-selection (concatenative), and previous results from [27]. MOS stands for Mean Opinion Score.

Результаты модели

	Parametric	Concatenative	Distilled WaveNet
English speaker 1 (female - 65h data)	3.88	4.19	4.41
English speaker 2 (male - 21h data)	3.96	4.09	4.34
English speaker 3 (male - 10h data)	3.77	3.65	4.47
English speaker 4 (female - 9h data)	3.42	3.40	3.97
Japanese speaker (female - 28h data)	4.07	3.47	4.23

Table 2: Comparison of MOS scores on English and Japanese with multi-speaker distilled WaveNets. Note that some speakers sounded less appealing to people and always get lower MOS, however distilled parallel WaveNet always achieved significantly better results.

Результаты модели

Method	Preference Scores versus baseline concatenative system Win - Lose - Neutral
Losses used	
KL + Power	60% - 15% - 25%
KL + Power + Perceptual	66% - 10% - 24%
KL + Power + Perceptual + Contrastive (= default)	65% - 9% - 26%

Table 3: Performance with respect to different combinations of loss terms. We report preference comparison scores since their mean opinion scores tend to be very close and inconclusive.

Выводы

- Полученная модель способна генерировать выходы быстрее, чем realtime
- При этом изменение качества практически отсутствует по сравнению с обычной WaveNet