Parallel WaveNet: Fast High-Fidelity Speech Synthesis

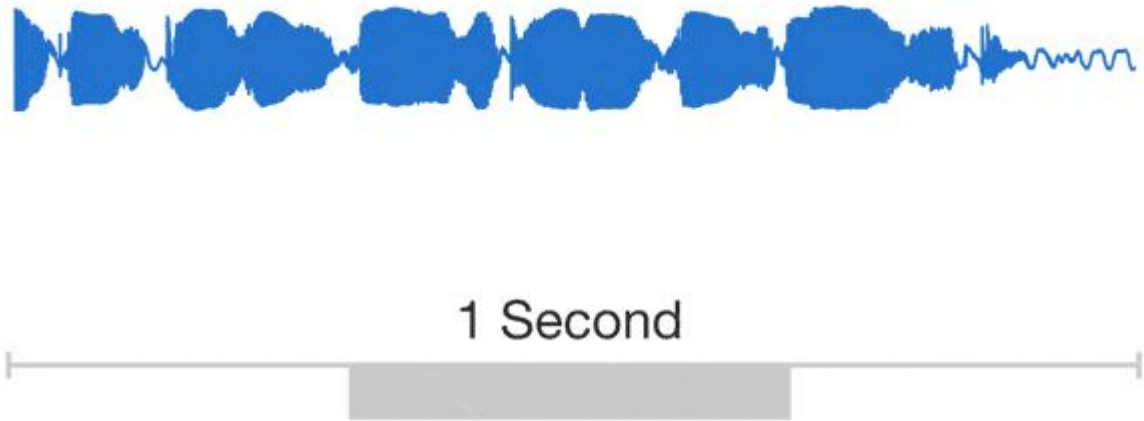
Федоров Павел
20.01.2021

Синтез речи (Text-To-Speech)

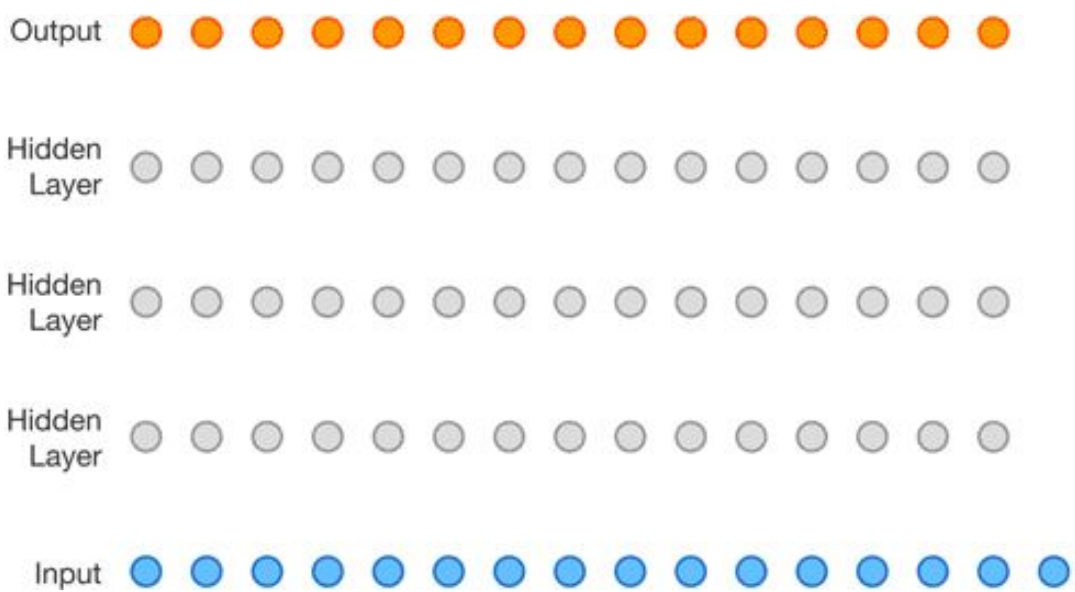
- **Concatenative TTS** - предварительно записывается огромный словарь звуков и фонем, из которых в дальнейшем составляется фраза. Так как звуки предзаписаны, практически невозможно изменить голос, эмоции или иным образом параметризировать результат
- **Parametric TTS** - Генеративные модели, способные изменять параметры голоса. Звуки генерируются специальными алгоритмами — vocoders.

WaveNet

WaveNet предлагает новый метод генерации — моделирование аудиосигнала целиком, до 16000 семплов в секунду или более, строго определенной формы в любых временных масштабах.



WaveNet



WaveNet

- Плюсы - обучение модели происходит быстро благодаря возможности распараллеливания
- Минусы - генерация проходит долго, так как происходит последовательно

Inverse-autoregressive flows

- Вместо последовательного семплирования на вход подается белый шум, который преобразуется к нужной форме
- IAF - это особый тип нормализующего потока, который моделирует многомерное распределение $p_x(x)$как явное обратимое нелинейное преобразование f простого распределения $p_z(z)$. Результирующая случайная величина x = f(z) имеет логарифмическую вероятность:

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

Inverse-autoregressive flows

Преобразование f имеет треугольную матрицу Якоби, которая делает определитель просто произведением диагональных элементов:

$$\log |\frac{dx}{dz}| = \sum \log \frac{df(z_{\leq t})}{dz_t}$$

Изначально генерируются семплы из логистического распределения

$$z \sim Logistic(0, I)$$

После чего применяется преобразование

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

Inverse-autoregressive flows

На выходе модели получается семпл x, при этом

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

В качестве $\mu(z_{<t}, \theta)$ и $s(z_{<t}, \theta)$ можно использовать любую авторегресионную модель, например, которая используется 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 статьи использовали 4 блока)

$$x^0 = z$$

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

Inverse-autoregressive flows

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

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

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

N - число блоков

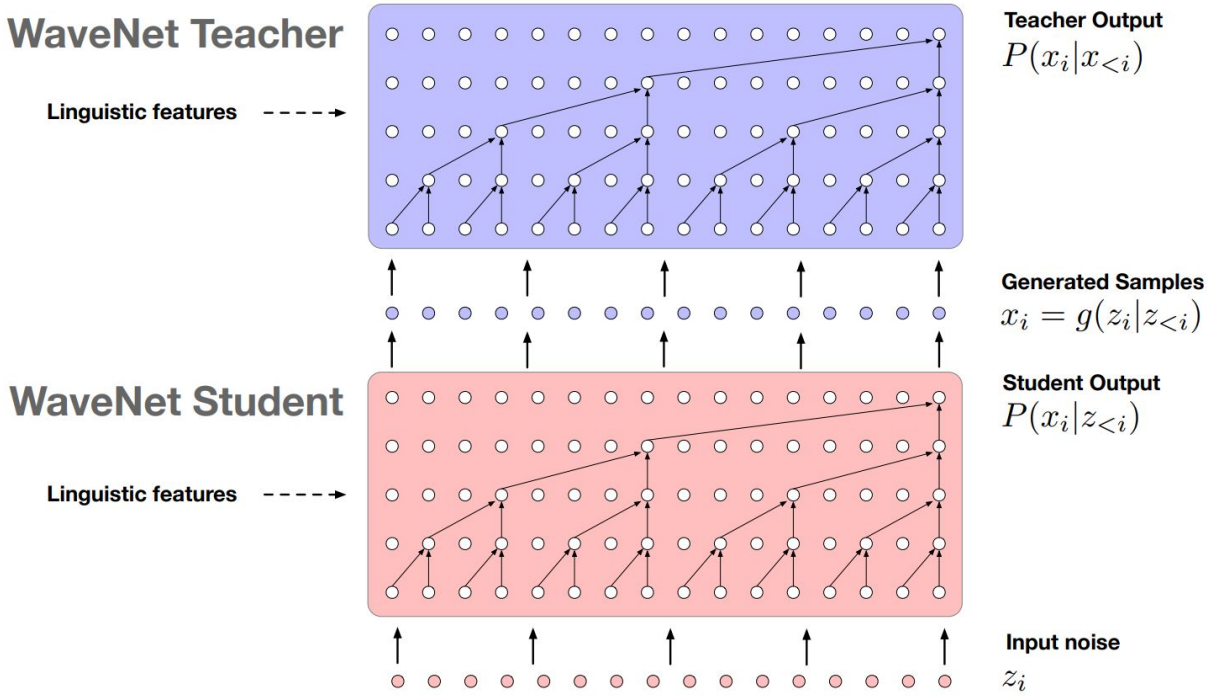
Обучение модели

- Сначала предобучается стандартный WaveNet (teacher network)
- Затем моделируется распределение полученной модели с помощью IAF (student network)
- Для обучения используется Probability Density Distillation loss

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

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

Probability Density Distillation



Probability Density Distillation

Кросс-энтропия может быть представлена в следующем виде

$$H(P_S) = E_z[\sum_{t=1}^{T} - \ln p_S(x_t|z_{<t})]$$

$$= E_z[\sum_{t=1}^{T} - \ln s(z_{<t}, \theta)] + 2T$$

$$H(P_S, P_T) = \sum_{t=1}^{T} E_{p_s(x_{<t})} H(P_S(x_t|x_{<t}), p_T(x_t|x_{<t}))$$

Обучение модели

Для улучшения качества используются дополнительные функции потерь:

- Power loss
- Perceptual loss
- Contrastive loss

Эксперименты

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

Эксперименты

	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

Эксперименты

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%

Эксперименты

- Parallel WaveNet способна генерировать выходы быстрее, чем в реальном времени
- При этом изменение качества практически отсутствует по сравнению с обычной WaveNet

Вопросы

1. Зачем для WaveNet потребовалась новая архитектура? Чем она лучше старой?
2. Что такое Probability Density Distillation и зачем это нужно?
3. В чем отличия процесса обучения Parallel WaveNet от обычного WaveNet?

Источники

- https://deepmind.com/blog/article/high-fidelity-speech-synthesis-wavenet
- https://arxiv.org/abs/1711.10433