WaveNet: A Generative Model for Raw Audio

Денис Золотухин

НИУ ВШЭ

15 марта 2019

О чём

- Модель звука в WaveNet
- Свёртки: простая и умная
- Архитектура сети
- Параметризация
- Эксперименты

Немного о звуке

Для хранения на компьютере звук делится на равные интервалы с выбранной частотой (дескретизации).

Рис.: Секунда речи

Иненсивность звука в каждом из них квантуется по выбранному числу уровней (глубина кодирования). Таким образом, звук можно представить как **временной ряд** $\{x_i\}_{i=1}^{T}$.

Модель распеределения в WaveNet

$X = \{x_1, \ldots, x_T\}$ - представление звука. Разбиваем $p(X)$ в произведение:

$$p(X) = \prod_{i=1}^{T} p(x_i | x_{i-1}, \ldots, x_1)$$

Стакаем слои, но только свёрточные, размерость по времени сохраняется.

Простая свёртка, на зависящая от будущего

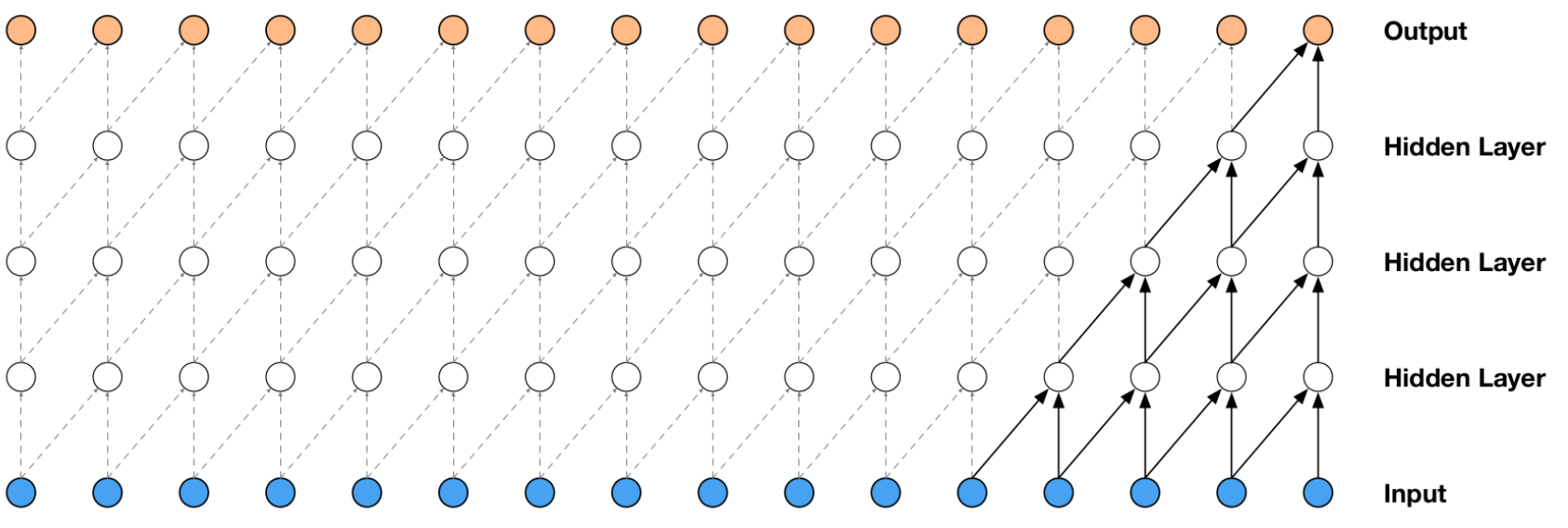

Рис.: Простая свёртка

Растянутая свёртка

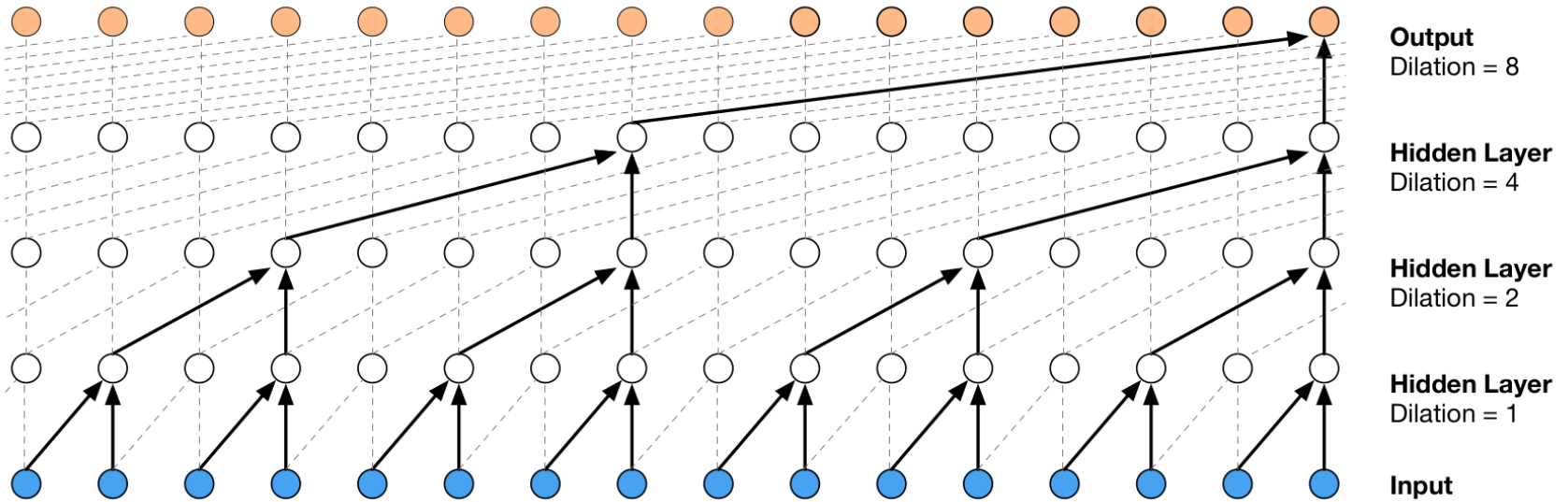

Рис.: Растянутая свёртка

Зачем?

Такая свёртка значительно расширяет область видимости, достижение которой обычными свертками непрактично.

$$\text{хотим помнить} \approx 1\text{с}$$

$$\text{частота } 44.1\text{кГц} \implies$$

$$\text{область видимости} \approx 44000$$

$$\text{будем стакать 10к слоёв?}$$

В WaveNet стакаются слои с размерами свёртки $1, 2, 4, \ldots, 512$, а потом снова начинают с размера 1.

Мю-закон

При глубине кодирования 16 бит получаем, что softmax должен предсказать 65536 вероятностей. Это число уменьшается с помощью μ-трансформации:

$$f(x_t) = \operatorname{sgn}(x_t)\frac{\ln(1+\mu|x_t|)}{1+\mu}$$

Здесь $-1 < x_t < 1$, $\mu = 256$. После этого $f(x_t)$ квантуется в одно из 256 значений.

Устройство слоя

Структура каждого слоя такова:

$$z = \tanh(W_{f,k} * X) \odot \sigma(W_{g,k} * X)$$

W - свёрточный вентиль (его мы учим), $*$ - свёртка, $\odot$ - поэлементное поизведение, k - номер слоя, z - внутреннее состояние.
f и g обозначают "filter"и "gate"вентили (привет, GRU).

Пропускаем слои

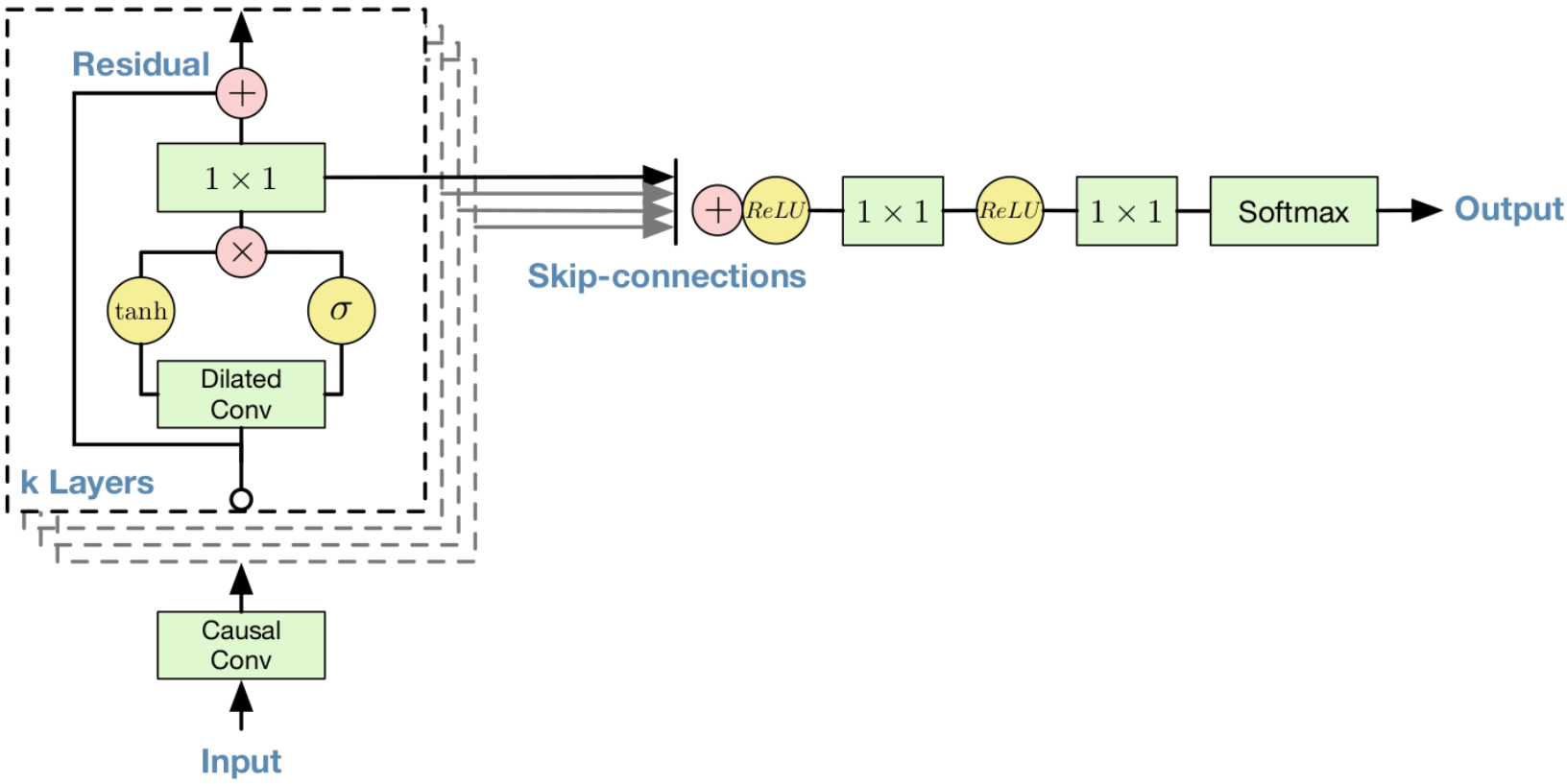

Рис.: Архитектура блока

Условная генерация

Если у нас есть дополнительный вход нашей сети, то мы можем генерировать выход, обуславливаясь на него:

$$p(x|h) = \prod_{i=1}^{T} p(x_i|x_{i-1}, \ldots, x_1, h)$$

Где h - входной параметр. Различаются два вида таких параметров: **глобальные** и **локальные**.

Глобальная параметризация

h представляет глобальное условие, напрмиер id человека. Тогда внутренние переменные считаются так:

$$z = \tanh(W_{f,k} * X + V_{f,k}^T h) \odot \sigma(W_{g,k} * X + V_{g,k}^T h)$$

V - линейные отображения h, которые мы можем выучить.

Локальная параметризация

h_t - последовательность, с более низкой частотой, чем у звука (слова текста). Тогда мы апскейлим её до нужной нам частоты $y = f(h)$ и заливаем в сеть:

$$z = \tanh(W_{f,k} * X + V_{f,k} * y) \odot \sigma(W_{g,k} * X + V_{g,k} * y)$$

Здесь $V_{f,k} * y$ это уже свёртка 1×1.

Зависимость только от говорящего

- 44 часа речи 109 людей
- Без текста: правдободоные звуки, но без смысла

- Музыка
- Ясно слышно фортепиано, но какая-либо мелодия отсутствует

Синтез речи на основе текста

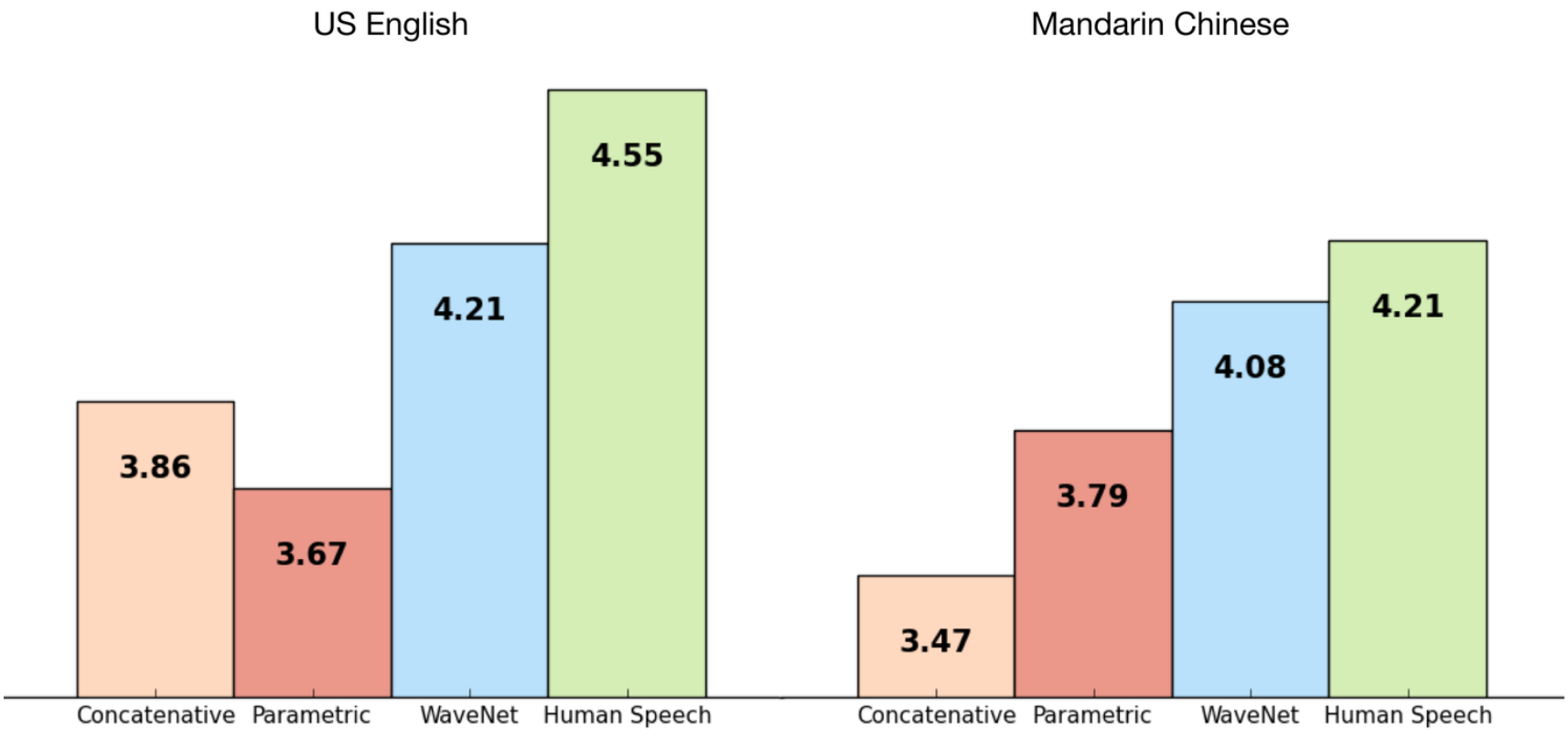

Рис.: Сравнение WaveNet с другими подходами

Выводы

- WaveNet - сеть для генерации звука
- Для увеличения области видимости применяются "дырявые"свёртки
- Используется схему вентилей, похожая на ту, которую можно увидеть в GRU
- WaveNet можно легко параметризовать: текст для синтеза, тембр голоса, качество записи
- WaveNet на данный момент одна из лучших сетей для генерации звука, особенно речи

Ссылки

[1]. https://arxiv.org/abs/1609.03499