WAVENET: A GENERATIVE MODEL FOR RAW AUDIO

van den Oord et al., 2016

Докладчик: Иовлева Анастасия, 162

Speech Synthesis or Text-To-Speech

Speech Synthesis or Text-To-Speech (TTS)

Concatenative speech synthesis

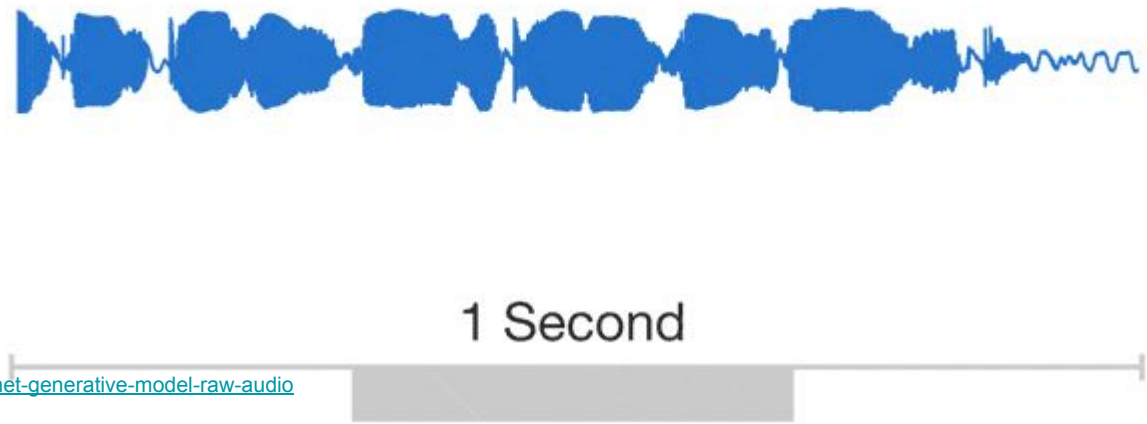
- Предварительно записывается огромный словарь звуков и фонем, из которых в дальнейшем составляется фраза.
- Так как звуки предзаписаны, практически невозможно изменить голос, эмоции или иным образом параметризировать результат.

Parametric speech synthesis

- Генеративные модели, способные изменять параметры голоса. Звуки генерируются специальными алгоритмами — vocoders.
- Сложны в настройке, результат как правило получается менее естественным, чем у Concatenative TTS.

Speech Synthesis or Text-To-Speech (TTS)

WaveNet предлагает новый метод генерации — моделирование аудио-сигнала *целиком*, *тик за тиком*, несмотря на то, что в одной секунде аудио их содержится очень много и генерировать такой массив данных сложно.



https://deepmind.com/blog/article/wavenet-generative-model-raw-audio

Speech Synthesis or Text-To-Speech (TTS)

Сравнение результатов разных методов генерации:

- **Parametric:**  
- **Concatenative:**  
- **WaveNet:**  

https://deepmind.com/blog/article/wavenet-generative-model-raw-audio

Speech Synthesis or Text-To-Speech (TTS)

WaveNet стала новым state-of-the-art.

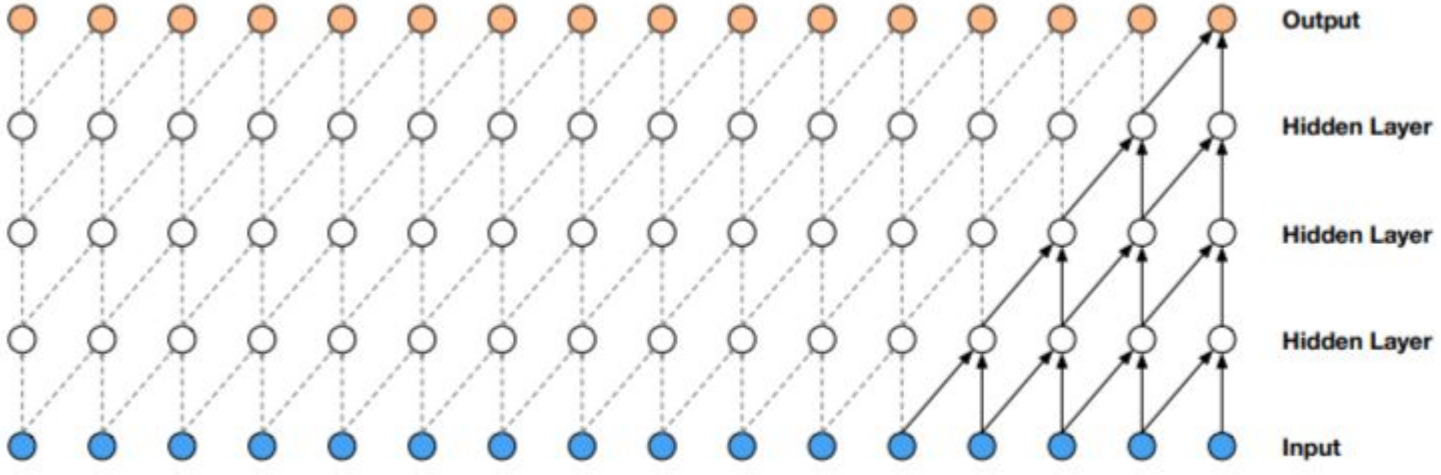
- По оценке добровольцев, качество генерации получается значительно лучше предыдущих state-of-the-art моделей.
- Одна и та же модель может генерировать разные голоса.
- Может использоваться для генерации любых звуков, а не только речи.
- С небольшими изменениями, модель также показывает неплохие результаты в распознавании речи, что говорит о потенциале в этой области.

Как WaveNet этого достигла?

Архитектура

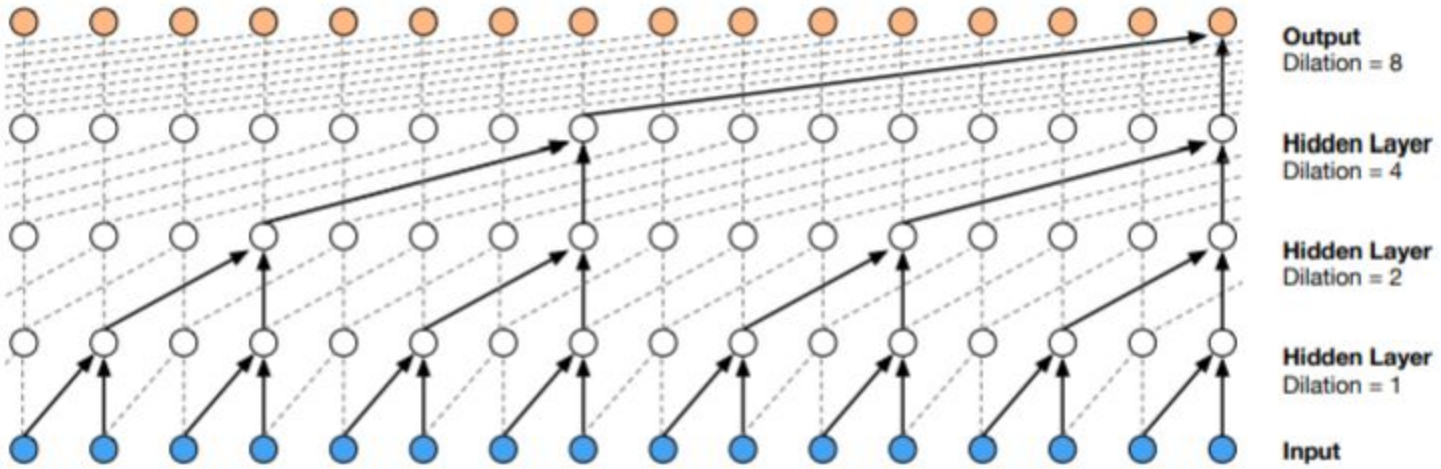
Causal convolutions

- Позволяет модели использовать для предсказания только предыдущие тики аудио, то есть не нарушается последовательность данных.
- Но — небольшой receptive field, растущий линейно с количеством параметров. Это очень плохо подходит для анализа аудиоданных.



Dilated causal convolutions

- Фильтр применяется на области большей, чем его размер, за счет того, что он пропускает несколько тиков данных (*dilation*).
- Если dilation растет экспоненциально от слоя к слою, то это дает *экспоненциальный рост* receptive field при линейном росте параметров.

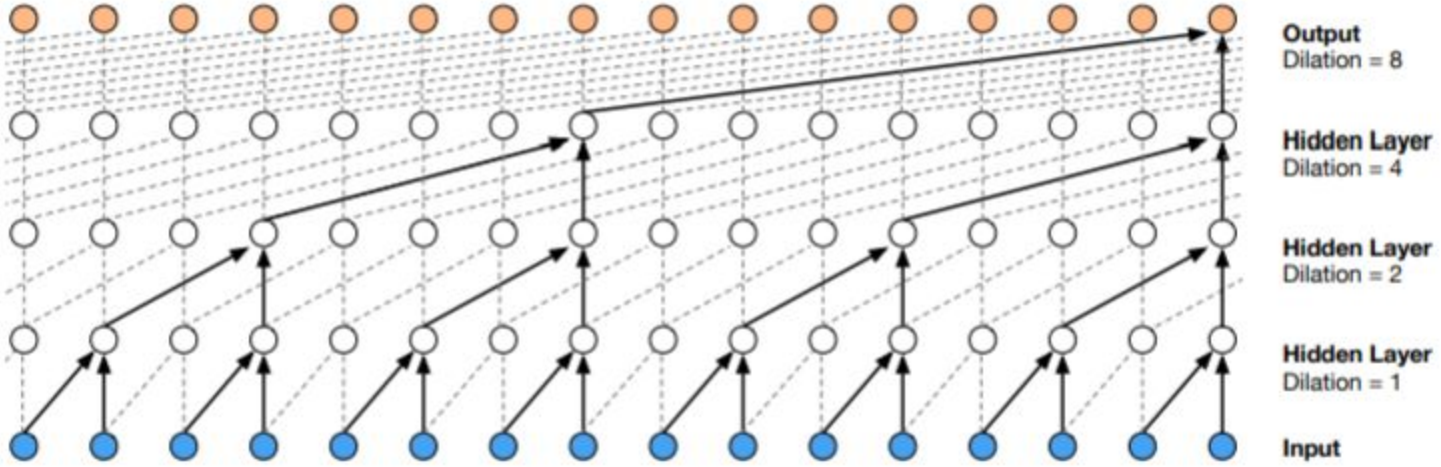


Dilated causal convolutions

- Авторы статьи используют степени двойки в качестве dilation:

  1, 2, 4, 8, … **512**, 1, 2, 4…

- Максимальный порог для размера dilation (512 в данном случае) позволяет контролировать рост receptive field и вычислительные затраты.



Softmax distribution

- Категориальные распределения на практике показывают лучший результат, чем непрерывные.
- Аудиоданные обычно хранятся в 16-битных числах, то есть имеют 65,536 возможных значений. Это слишком много, поэтому они “проецируются” в меньшее количество категорий, например, в 256.

$$f(x_t) = \text{sign}(x_t)\frac{\ln(1+\mu|x_t|)}{\ln(1+\mu)}$$

x_t — тики, нормализованные в диапазон (-1, 1); μ — количество категорий.

Активация

- $*$ — свертка
- $\odot$ — поэлементное умножение
- $\sigma()$ — сигмоида
- $W_{f,k}$, $W_{g,k}$ — веса свертки

$$\mathbf{z} = \tanh\left(W_{f,k} * \mathbf{x}\right) \odot \sigma\left(W_{g,k} * \mathbf{x}\right)$$

Conditioning

Передача дополнительных данных в модель помогает настроить характеристики генерируемой речи. Conditioning может быть глобальным и локальным.

Глобальное

- Дополнительные данные (h) одинаково влияют на генерацию на каждом тике.
- Пример: идентификатор говорящего для выборки с несколькими дикторами.

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

Conditioning

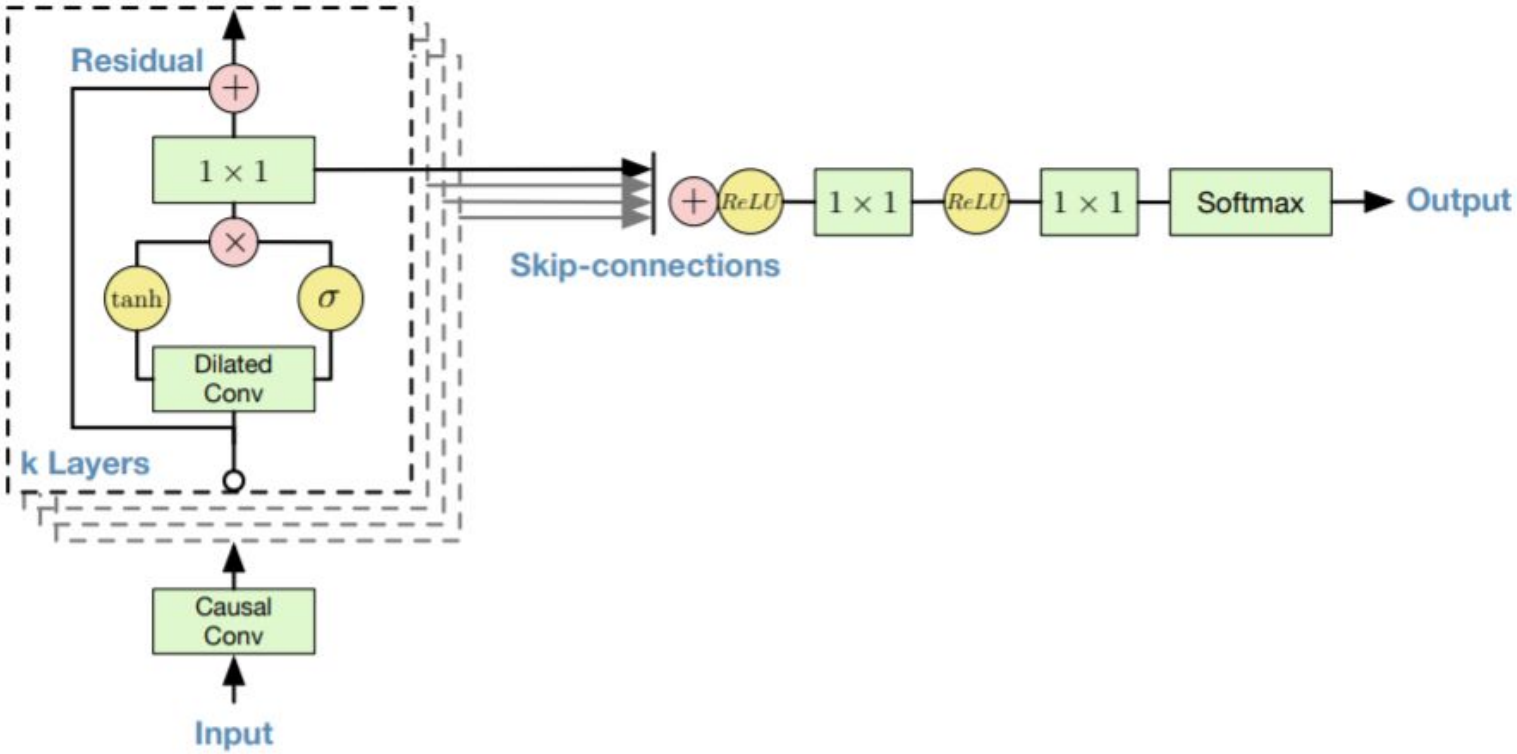
Передача дополнительных данных в модель помогает настроить характеристики генерируемой речи. Conditioning может быть глобальным и локальным.

Локальное

- Дополнительные данные это еще одна серия тиков h_t.
- С помощью сверточной сети, h_t преобразуется в $y = f(h_t)$ с той же размерностью, что и у x_t.
- Пример: вектор характеристик текста для TTS.

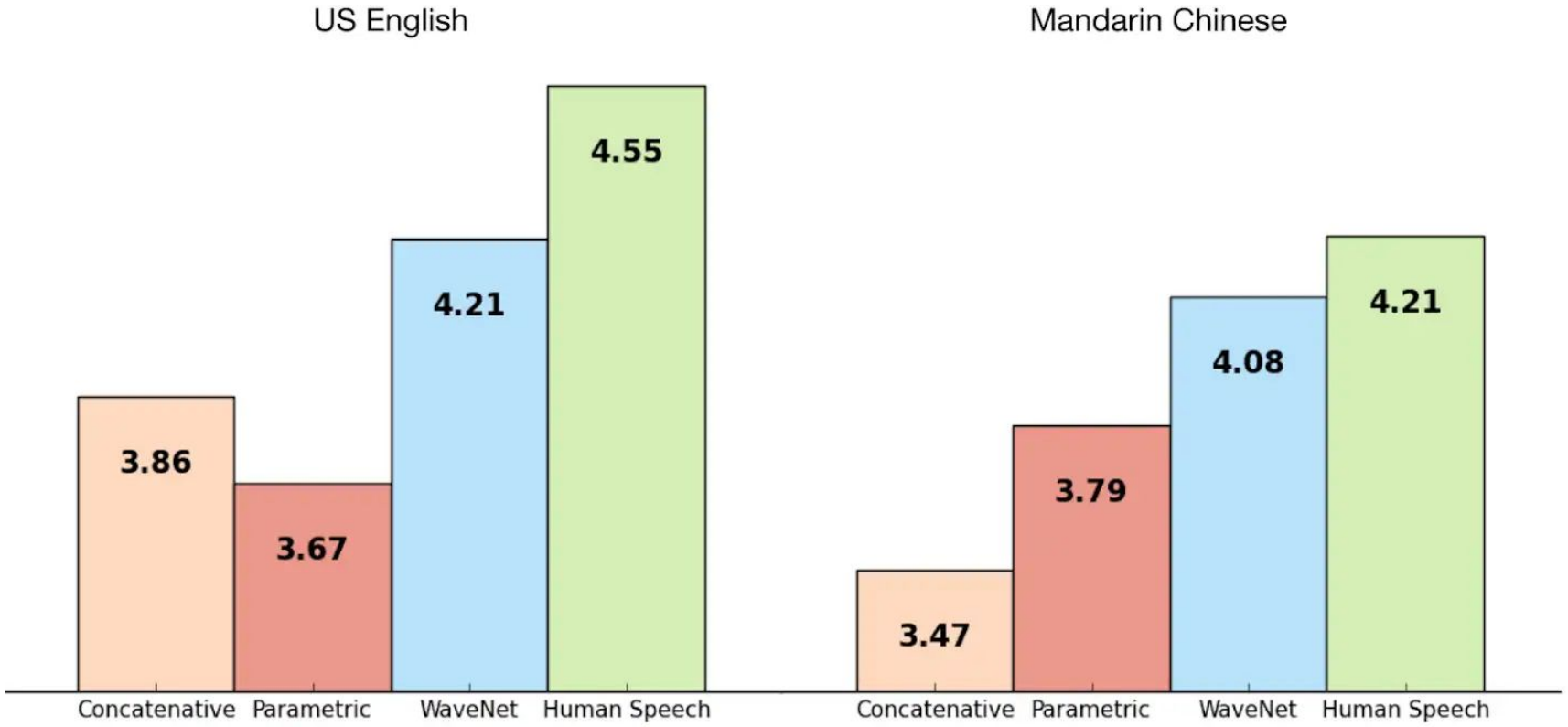
$$\mathbf{z} = \tanh\left(W_{f,k} * \mathbf{x} + V_{f,k} * \mathbf{y}\right) \odot \sigma\left(W_{g,k} * \mathbf{x} + V_{g,k} * \mathbf{y}\right)$$

Финальная архитектура



Эксперименты

Text-To-Speech



Text-To-Speech

Сравнение результатов разных методов генерации:

- **Parametric:**  
- **Concatenative:**  
- **WaveNet:**  

https://deepmind.com/blog/article/wavenet-generative-model-raw-audio

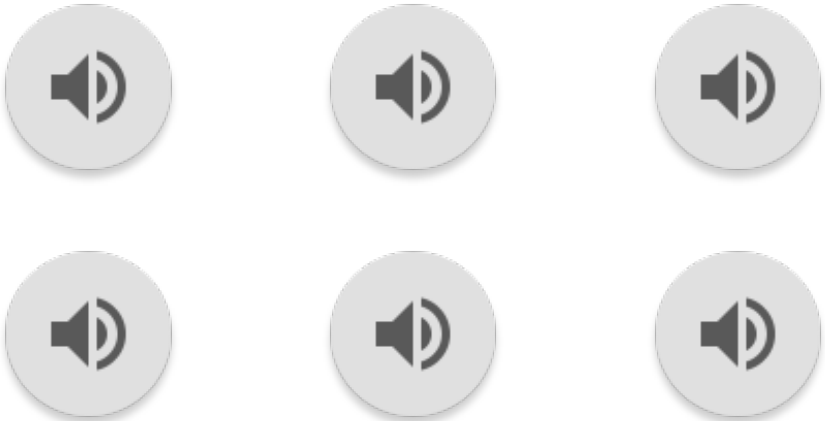
Изменение голоса

Одна и та же модель, обученная на датасете с несколькими дикторами.
Разные голоса получены с помощью глобального conditioning.

https://deepmind.com/blog/article/wavenet-generative-model-raw-audio

Генерация правдоподобной речи

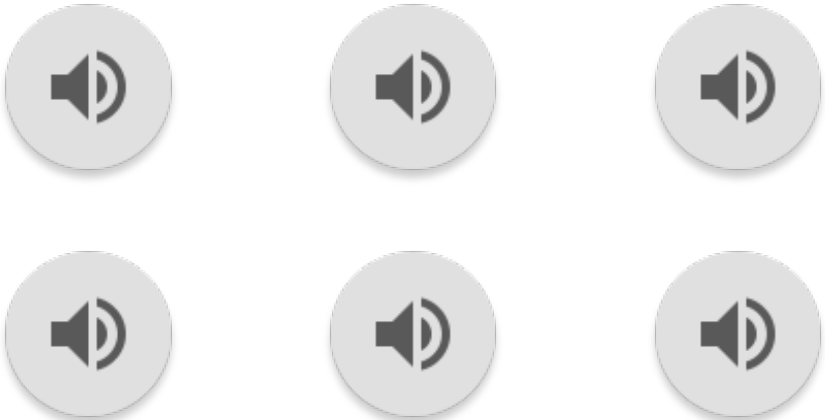
Генерация речи без передачи характеристик текста.



https://deepmind.com/blog/article/wavenet-generative-model-raw-audio

Генерация музыки

Модель обучалась на музыкальной выборке.



https://deepmind.com/blog/article/wavenet-generative-model-raw-audio

Заключение

Заключение

- WaveNet на порядок превосходит все предыдущие модели по качеству генерируемой речи.
- Одна и та же модель может использоваться для генерации разного голоса или текста.
- Модель может использоваться для генерации других звуков, например, музыки.
- Тем не менее, генерация занимает довольно много времени.
    - 17 часов для 4 секунд аудио на 1070 GPU.

Вопросы по статье

- Что такое dilated causal convolution в WaveNet? В чем отличие от causal convolution? Зачем она используется?
- В чем разница между глобальным и локальным conditioning? При решении каких задачах они используются?
- Напишите три формулы активации — без conditioning, с глобальным conditioning и с локальным conditioning.