WaveNet

- A generative model for raw audio
  - probabilistic
  - autoregressive
    - each audio sample is conditioned on all previous ones
  - can be efficiently trained
  - state-of-the-art performance on TTS
    - English and Mandarin
  - can model different speakers
    - speaker identity
  - other usage
    - model music
    - phoneme recognition

WaveNet

- Why?
  - 得益于：用神经网络建模复杂分布，例如联合概率
    - 可以轻松建模具有上千个随机变量的概率分布
    - Pixel CNN, Pixel RNN
    - 在图像和文本领域取得了很好的结果
  - 是否也可以用来生成语音波形?
    - 声音信号具有很高的临时分辨率
    - 每秒 16000 个采样点
    - 基于 Pixel CNN 的结构，提出 WaveNet

WaveNet

- Contributions
  - 能合成非常自然的语音
    - 自然度 state-of-the-art，明显超过参数和拼接方法
  - 设计了新的网络结构
    - dilated causal convolution
    - 具有超大感受野，能处理长程依赖
  - 单个模型输出多种声音
    - 通过 speaker identity 控制
  - 可应用到其它领域
    - 音素识别
    - 音乐合成

WaveNet

- 音频波形的联合概率
  - 用条件概率的乘积来表示联合概率

$$p(\mathbf{x}) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1})$$

  - each audio sample is conditioned on all previous ones
  - 通过堆叠卷积层来建模条件概率
    - no pooling layers
    - 模型的输出和输入有相同的时间维度
    - 模型输出当前采样点的 softmax 概率分布

WaveNet

- Dilated causal convolutions
  - 因果卷积， WaveNet 中的干货
    - 确保条件概率在物理上是因果的，即时序是正确的
    - 在图像领域， causal conv 等价于 masked conv
      - 先用 mask 和 conv kernel 进行 elementwise 乘法
      - 再用 masked conv kernel 执行卷积操作
    - 在 1 维数据领域，如音频，对卷积结果进行移位即可
  - 训练阶段
    - 可以并行地计算各个时刻的条件概率（所有真值都已知）
  - 预测阶段
    - 只能串行计算（计算下一个采样点时需要当前采样点的值）

WaveNet

- Dilated causal convolutions
  - 由于因果卷积没有循环连接，所以，在长序列上训练速度比 RNN 快很多
  - 可通过多层堆叠或者扩大 filter 来增大感受野
  - 使用 dilated conv 可显著增大感受野，且计算量没有明显提升
  - 下图感受野为 5 （ #layers+filter length-1 ）

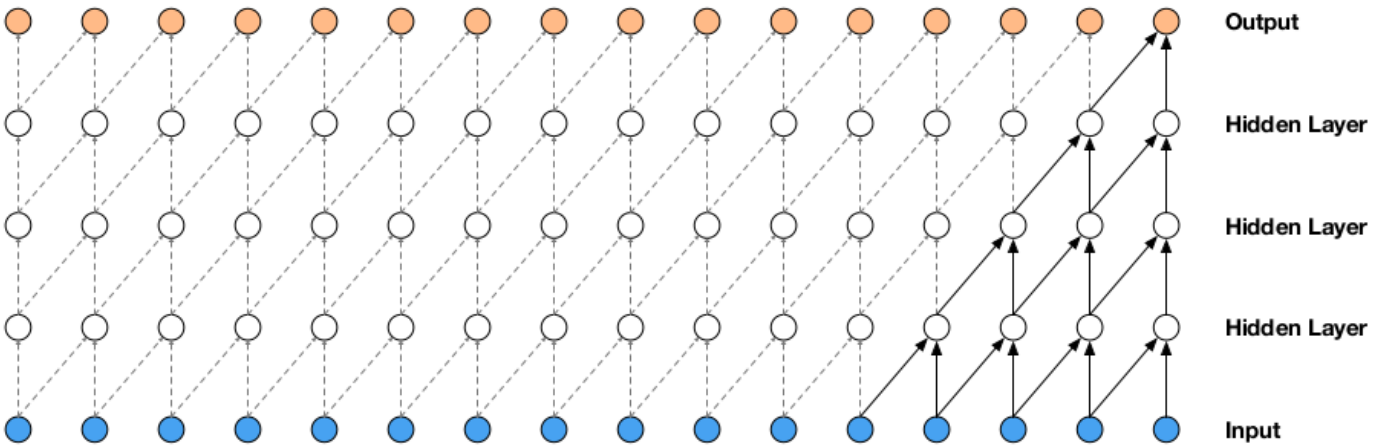

Figure 2: Visualization of a stack of causal convolutional layers.

WaveNet

- Dilated causal convolutions
  - dilated conv => conv with holes (zeros)
  - 与 pooling 或 strided conv 相似，但输出和输入维度相同
  - dilation=1 => standard conv

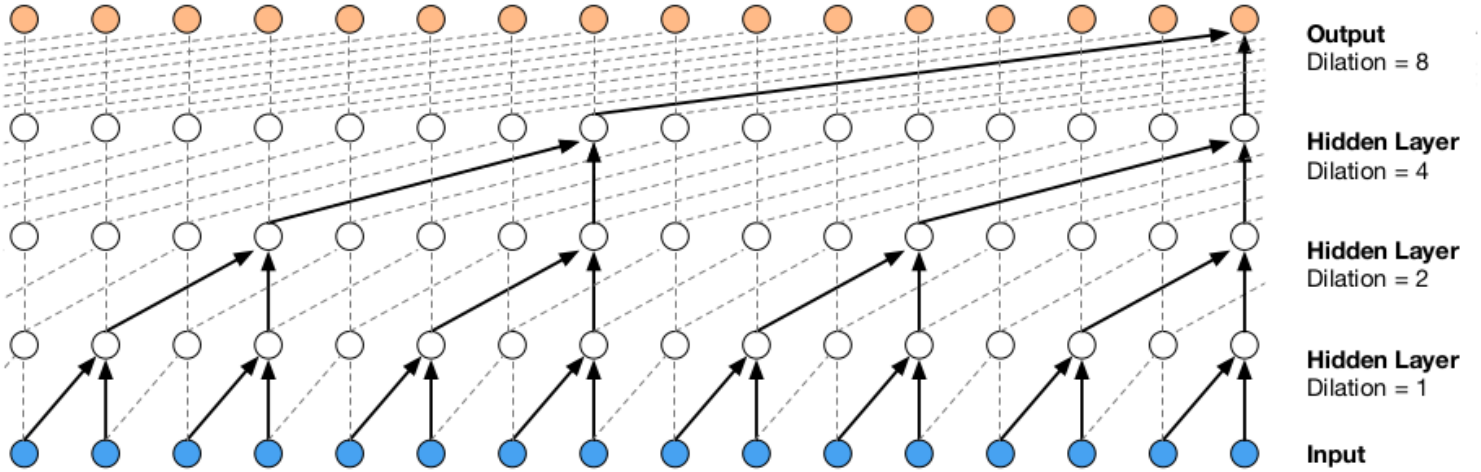

Figure 3: Visualization of a stack of *dilated* causal convolutional layers.

WaveNet

- Dilated causal convolutions
  - 堆叠 dilated convs 可以显著增大感受野
  - 论文中的 dilation 变化策略
    - 1,2,4,…,512,1,2,4,…,512,1,2,4,…,512
    - 感受野随深度呈指数级增长
    - 每一个 1,2,4,…,512 的感受野都是 1024
  - 堆叠 dilated convs 还可以增加模型的表达能力

WaveNet

- 在音频上建模条件概率分布
  - 混合模型，如混合密度网络、混合条件高斯等
  - softmax distribution
    - 因为分类分布（ categorical distribution ）更灵活
    - 更容易建模任意分布，因为它对分布的形状不作假设
  - 音频序列，每个采样点为 16 位整型，需要输出 65536 个概率
  - 先应用 mu-law 压缩变换，再量化到 256 个值的范围
    - 非线性量化比线性量化的重建效果好

$$f\left(x_t\right)=\operatorname{sign}(x_t)\frac{\ln\left(1+\mu\left|x_t\right|\right)}{\ln\left(1+\mu\right)},\quad -1<x_t<1 \text{ and } \mu=255$$

WaveNet

- Gated activation units

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

- Residual and skip connections

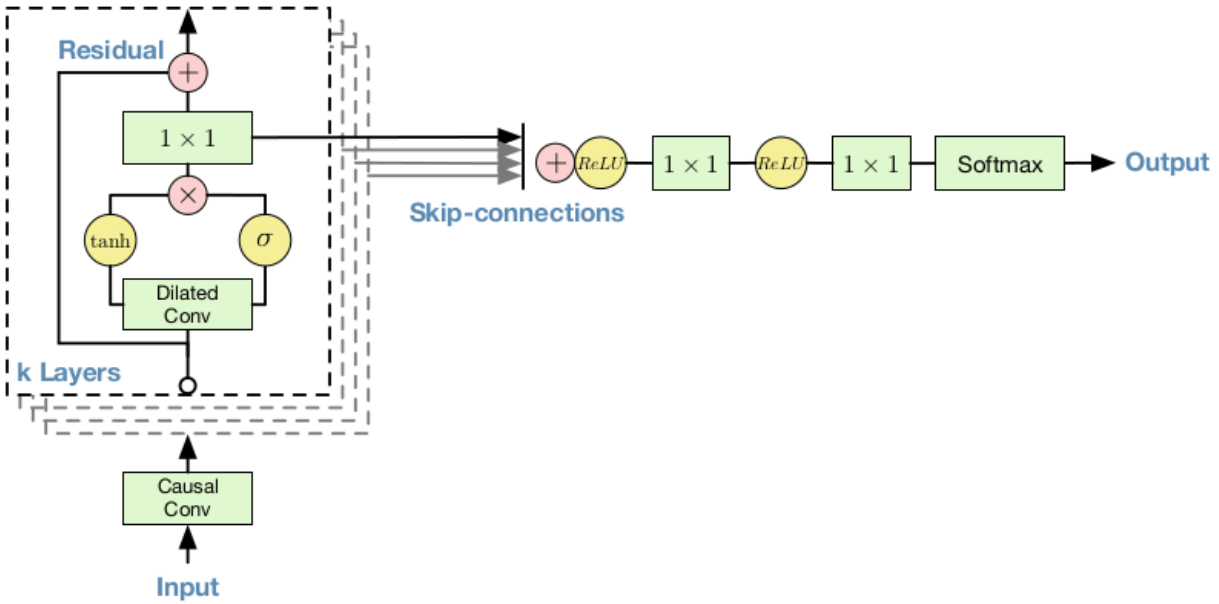

Figure 4: Overview of the residual block and the entire architecture.

WaveNet

- Conditional wavenet
  - conditional distribution

$$p(\mathbf{x} \mid \mathbf{h}) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1}, \mathbf{h})$$

  - global conditioning
    - eg. speaker identity

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

  - local conditioning
    - eg. linguistic features, need upsampling first $\mathbf{y} = f(\mathbf{h})$

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

WaveNet

- 训练数据
  - 英文 24.6 hours
  - 中文 34.8 hours
- 信息流向
  - input texts => linguistic features
  - linguistic features => logF0, phone durations
  - linguistic features, logF0 => audio wave

WaveNet

- Linguistic features
  - 音素、音节、词、短语、句子级别的特征
  - 音节重音、词中的音节数、当前音节在短语中的位置
  - 和音频、声学特征每 5ms 一对齐
  - linguistic features => logF0, phone durations

WaveNet

- TTS background
  - sequence to sequence mapping problem
    - discrete text => real-valued speech signals
  - pipeline
    - text analysis
      - NLP：分句，分词， TN ， POS tagging
      - 韵律分析： prosody prediction
      - G2P：音节转音素（ grapheme to phoneme ）
      - word sequence in, phoneme sequence out
    - speech synthesis
      - 拼接合成法
        - 从录音中选择声音单元来拼接合成
      - 参数合成法
        - 先预测声学参数，再合成音频