WaveNet

01. abstract

- raw audio waveforms 생성하기 위한 Deep Neural Network
- 각 오디오 샘플에 대한 예측 분포는 이전의 모든 샘플에 따라 조정
- 동일한 fidelity로 다양한 화자의 특성 캡처

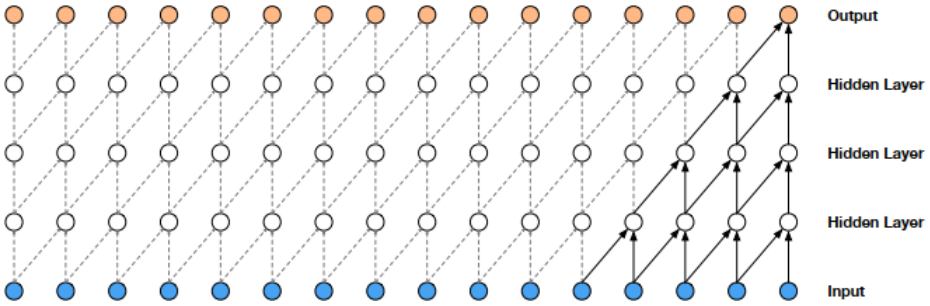
02. introduction

- WaveNet, an audio generative model based on the PixelCNN
- raw audio 생성에 필요한 장거리 종속성을 처리하기 위해 매우 큰 receptive field(RF)를 나타내는 dilated causal convolution을 기반으로 하는 새로운 구조
- 화자 정체성을 조건으로 할 때 모델을 사용하여 다른 목소리 생성

- **WAVENET**
  - waveform의 결합(joint) 확률 $x = x_1, ..., x_t$는 조건부 확률의 곱으로 분해

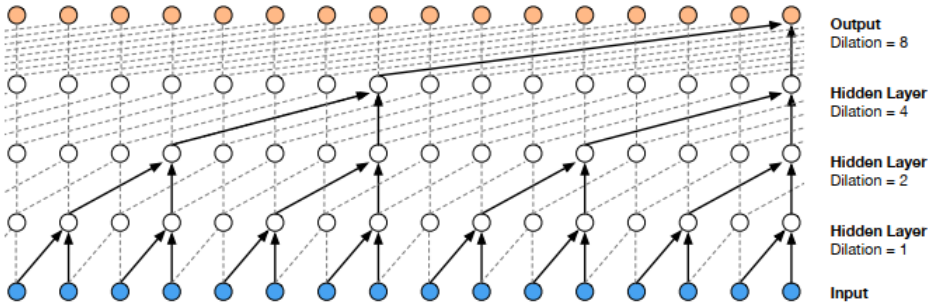
$$p(\mathbf{x}) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1})$$

  - conditional probability 분포는 CNN 스택으로 모델링 (Similarly to PixelCNNs)
    - No pooling
    - Input shape == Output shape
    - output softmax
    - data의 log-likelihood를 최대화하는 방향으로 최적화
      - 과적합, 과소적합을 쉽게 측정하기 위해서
  - Causal Convolution(CC)
    - 모델이 데이터 모델링하는 순서를 위반하지 않음.
    - 일반 CNN의 출력을 몇 단계씩 이동하여 구현
    - 각 샘플이 예측된 후 다음 샘플을 예측하기 위해 fed
    - DCC는 반복 연결이 없어 RNN보다 학습이 빠름.
    - problems
      - RF를 늘리기 위해 많은 레이어 또는 큰 필터 필요
        - 아래 그림에서는 5(layers + filter -1)
      - 본 연구에서는 DCC를 사용



Causal Convolutions

- DCC
  - 특정 단계로 입력 값을 건너뛰어 필터가 길이보다 더 큰 영역에 적용되는 Conv
  - pooling, stride CNN과 유사하지만 wavenet은 입/출력 동일
  - 1,2,4,8에 대한 DCC 생성
  - 입력 해상도와 계산 효율성을 유지하면서 매우 큰 RF



Stacked dilated convolution

- intuition
  - dilation rate를 기하적으로 키우면, 기하적으로 RF도 증가
    - 1, 2, 3, …., 512 블록은 크기 1024의 RF
    - 1X1024 Conv보다 효율적이고 차별적(비선형)
  - 블록을 쌓으면 모델 용량과 RF 크기가 더욱 증가

- **Softmax Distributions**
  - 개별 오디오 샘플에 대한 조건부 분포를 모델링하는 방식
    - mixture density network
    - mixture of conditional Gaussian scale mixtures(MCGSM)
  - 그러나, van den Oord et al. (2016a)는 데이터가 암시적으로 연속적일 때 softmax가 더 잘 작동한다는 것을 보여줌.
    - like image pixels, audio sample values
    - categorical distribution이 더 유연하고 모양에 대한 가정을 하지 않기 때문에 arbitrary distribution을 더 쉽게 모형화 할 수 있다.
    - μ-law compading transformation to the data
      - 이후, 256개의 가능한 값으로 양자화
      - μ = 255
      - 음성의 경우, 양자화 후 재구성된 신호가 원본가 매우 유사하게 들리는 것을 발견
- **Gated Activation Units** (same in the gated PixelCNN)

- 실험을 통해 rectified linear activation function보다 더 잘 작동함을 확인

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

- **Residual and Skip Conn**
  - 수렴이 빨라지고, 더 깊은 모델로 학습 허용
- **Conditional WaveNets**
  - given an additional input h
  - 다른 입력 변수에 대해 모델을 조정하여 필요한 특성을 가진 오디오 생성
    - ex) 다중 스피커 설정에서 speaker ID를 추가 입력으로 모델에 공급하여 speaker 선택
    - TTS의 경우 텍스트에 대한 정보를 추가 입력으로 제공해야 함.
  - global/local conditioning으로 입력에 대한 모델 조건화
    - global conditioning
      - 모든 시간 단계에 걸쳐 출력 분포에 영향을 미치는 single latent representation h
      - e.g. a speaker embedding in a TTS model.
      - as follows:
        - V : learnable linear projection
        - V^T : broadcast over the time dimension

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

    - local conditioning
      - 오디오 신호보다 샘플링 주파수가 낮을 수 있는 두 번째 시계열 h_t
      - transposed CNN(learned upsampling)
        - 오디오 신호와 동일한 해상도로 새로운 시계열 $y = f(h)$로 mapping
        - activation unit에서 사용
        - as follows:
          - $*y$: 1x1 conv, transposed CNN 대체

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

- **Context Stacks**
  - 앞서 RF를 늘리는 방법 외 보완적인 접근 방식
  - 오디오 신호의 긴 부분을 처리하고 오디오 신호의 작은 부분(끝에서 잘림)만 처리하는 더 큰 WaveNet을 로컬로 조절하는 별도의 더 작은 컨텍스트 스택을 사용
  - 다양한 길이와 hidden unit 수를 가진 여러 컨텍스트 스택 사용 가능
  - 더 큰 RF가 있는 스택은 레이어당 unit이 더 적음.
  - 계산의 요구 사항을 합리적인 수준으로 유지하고 더 긴 시간 척도에서 시간 상관 관계 모델링을 위해서 더 적은 용량 필요
  - ⇒ 그니까 요약하면 스택을 더 쌓아서 성능을 올릴 수는 있는데, 많이 쌓는다고 해서 모든 레이어가 항상 같은 횟수로 사용되지 않고, 별로 안 쓰이는 레이어가 있을 수도 있음. 그래서 합리적인 수준으로 적당히 스택을 쌓는 게 중요하다.

03. Experiments

- evaluate it on three different tasks
  - Multi-Speaker Speech Generation (not conditioned on text)
  - TTS
  - music audio modelling (이건 패스)
- **Multi-Speaker Speech Generation**
  - `CSTR voice cloning toolkit(VCTK)`의 English multi-speaker corpus를 사용
    - 109명, 44시간 동안의 데이터
  - speaker는 one-hot vector로 모델에 입력
  - 결과
    - 장거리 일관성의 결여
      - 모델의 RF의 제한된 크기에 영향
    - single WaveNet으로 모든 화자의 음성 모델링
      - 단일 모델에서 109명의 특성을 모두 캡처
      - signle speaker에 비해 speaker를 추가하고 학습하면 valid가 더 좋아졌음.
        - WaveNet의 내부 표현이 여러 speaker에게 공유가 되었음을 시사
          - ⇒ 아마 DCNN 이야기인듯 하다.
    - 오디오의 다른 특성 또한 캡처
      - speaker의 호흡과 입 움직임 뿐만 아니라 음향 및 녹음 품질도 모방
- **TTS**
  - Google's North American English and Mandarin Chinese TTS system과 같이 single-speaker speech database 사용
  - The North American English dataset은 24.6시간의 음성 데이터, Mandarin Chinese 데이터에는 34.8시간의 음성 데이터 포함
    - 모두 여성 전문 연사의 음성
  - TTS 작업
    - 입력 텍스트에서 파생된 언어 기능에 대해 로컬로 조정
    - 언어적 특징 추가
    - 앞서 언급한 모델에서 HMM(Hidden Markov Model) 기반 unit selection concatenative, LSTM-RNN 기반 statistical parametric 추가 사용
    - subjective paired comparison, mean opinion score(MOS) 테스트 수행