WAVENET

A GENERATIVE MODEL FOR RAW AUDIO

TASSOS MANGANARIS

MARCH 2023

INTRODUCTION

- Exploring "raw audio generation techniques, inspired by … autoregressive generative models that model complex distributions"
  - PixelCNN (van den Oord et al. 2016)
  - RNNs (and 1-D convolutions) for Language Models (Jozefowicz et al. 2016)
- "Modeling joint probabilities … as products of conditional distributions"
- Can similar approaches succeed in generating wideband (>16,000hz) raw audio waveforms?

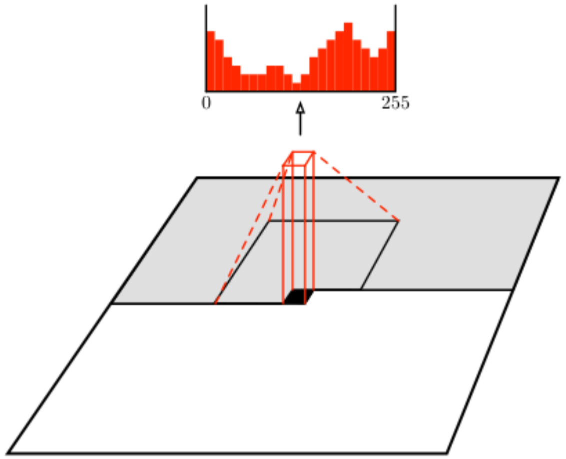
WAVENET

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

- $p\left(x_t \mid x_1, \ldots, x_{t-1}\right)$ is modelled by a stack of convolutional layers, like with PixelCNN
- "The output of the model has the same time dimensionality as the input."
- "Outputs a categorical distribution … with a softmax layer"
- Trained to optimize log-likelihood

CAUSAL CONVOLUTIONS

- Convolutions that do not violate the ordering of the model.
- In PixelCNN, implemented with masking.
- In WaveNet, implement by "shifting" output.

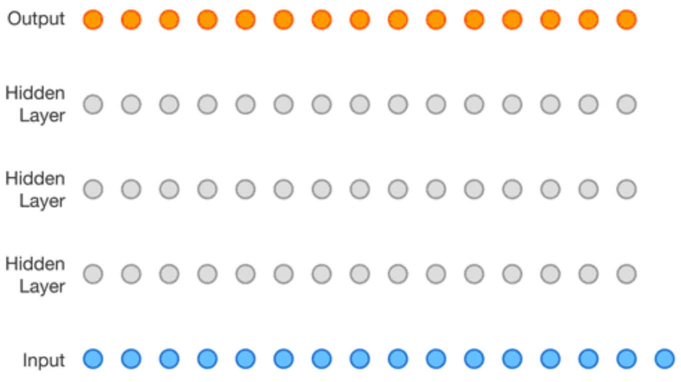


1	1	1	1	1
1	1	1	1	1
1	1	0	0	0
0	0	0	0	0
0	0	0	0	0

DILATED CAUSAL CONVOLUTIONS

- A convolution where the filter skips input values with a certain step.
- Stacked with exponential dilation factors up to a limit, then repeated.
- Receptive field grows exponentially with number of hidden layers.

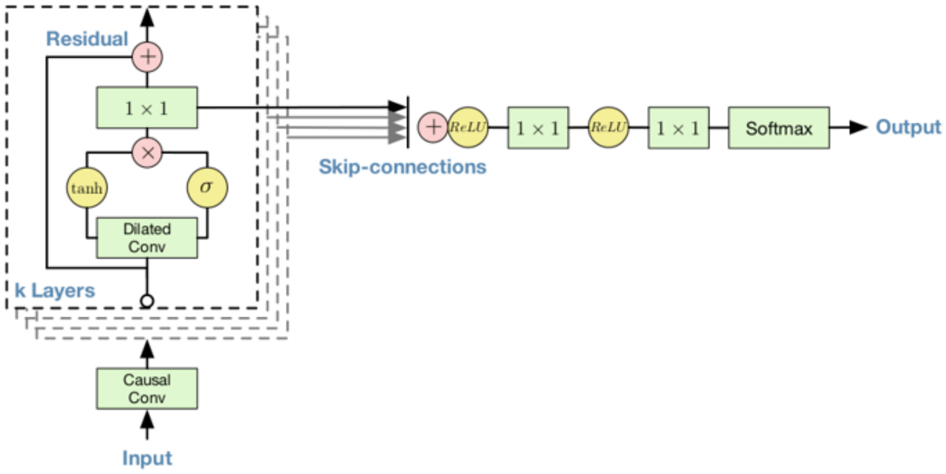
$$1, 2, 4, \ldots 512, 1, 2, 4, \ldots 512, 1, 2, 4, \ldots 512$$



SOFTMAX DISTRIBUTIONS

- PixelCNN used softmax over mixtures of Gaussians.
- A problem: raw audio samples are typically quantized with 16 bits $\Rightarrow 2^{16}$ probabilities.
- Solution: Quantize according to mu-law. Now effectively encoding the signal with 8 bits ("Mu-Law Algorithm" 2023).

GATED ACTIVATION UNITS + RESIDUAL AND SKIP CONNECTIONS



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

CONDITIONAL WAVENETS

- After training, we can generate likely, but incoherent waves.

speaker-1.wav

- Modify the model to include an extra vector for conditioning.

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

GLOBAL CONDITIONING

- "A single latent representation h that influences output distribution across all time steps."

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

- Result from $V_{*,k}^T \mathbf{h}$ is broadcast across time dimension, and V is like a vector with length `(n_aux)`.

LOCAL CONDITIONING

- h_t, a time series of linguistic features. Therefore, WaveNet plays the role of the acoustic model + vocoder.
- Up sample with transposed CNN, so that length of the final time series matches with $\mathbf{x}$

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

- $V_{*,k}$ is a 1x1 convolution for each layer, that take `n_aux` channels and outputs `n_quant` channels.

IN CODE

- From ESPNet:

```
def _residual_forward(
        self,
        x, # series of quantized, one-hot-ed waveform points (B, T, 256).

        h, # upsampled conditioning tensor (B, n_aux, T)
        dil_sigmoid,
        dil_tanh,
        aux_1x1_sigmoid,
        aux_1x1_tanh,
        skip_1x1,
        res_1x1,
):
    output_sigmoid = dil_sigmoid(x)
    output_tanh = dil_tanh(x)
    aux_output_sigmoid = aux_1x1_sigmoid(h)
    aux_output_tanh = aux_1x1_tanh(h)
    output = torch.sigmoid(output_sigmoid + aux_output_sigmoid) * torch.tanh(
        output_tanh + aux_output_tanh
    )
    skip = skip_1x1(output)
    output = res_1x1(output)
    output = output + x
    return output, skip
```

EXPERIMENTS

MULTI-SPEAKER SPEECH GENERATION

- Global conditioning for speaker identity (a one-hot vector).
- "able to model speech from any of the [109] speakers"
- "internal representation was shared among multiple speakers"
- "it also mimicked the acoustics and recording quality, as well as the breathing and mouth movements of the speakers."

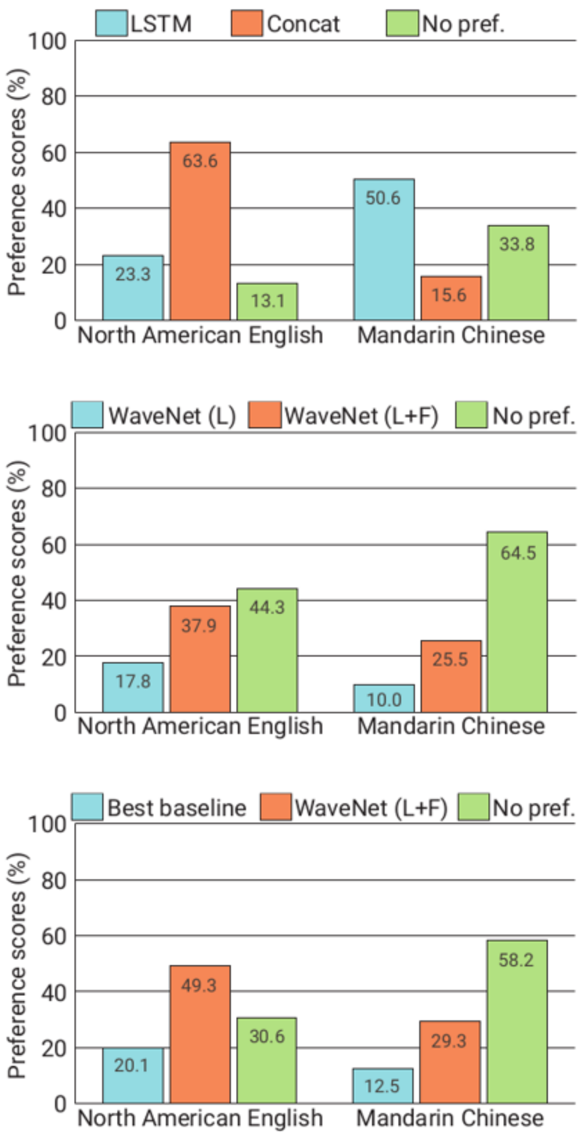
speaker-1.wav speaker-2.wav speaker-3.wav speaker-4.wav speaker-5.wav speaker-6.wav

(“WaveNet: A Generative Model for Raw Audio,” n.d.)

TEXT-TO-SPEECH

Speech samples	Subjective 5-scale MOS in naturalness	
	North American English	Mandarin Chinese
LSTM-RNN parametric	3.67 $\pm$ 0.098	3.79 $\pm$ 0.084
HMM-driven concatenative	3.86 $\pm$ 0.137	3.47 $\pm$ 0.108
WaveNet (L+F)	**4.21** $\pm$ 0.081	**4.08** $\pm$ 0.085
Natural (8-bit μ-law)	4.46 $\pm$ 0.067	4.25 $\pm$ 0.082
Natural (16-bit linear PCM)	4.55 $\pm$ 0.075	4.21 $\pm$ 0.071

parametric-1.wav concatenative-1.wav wavenet-1.wav tacotron.wav


LSTM
Concat
No pref.
Preference scores (%)
100
80
60
40
20
0
63.6
50.6
33.8
23.3
15.6
13.1
North American English
Mandarin Chinese
WaveNet (L)
WaveNet (L+F)
No pref.
Preference scores (%)
100
80
60
40
20
0
64.5
44.3
37.9
25.5
17.8
10.0
North American English
Mandarin Chinese
Best baseline
WaveNet (L+F)
No pref.
Preference scores (%)
100
80
60
40
20
0
58.2
49.3
30.6
29.3
20.1
12.5
North American English
Mandarin Chinese

MUSIC

sample_1.wav sample_2.wav sample_3.wav sample_4.wav sample_5.wav sample_6.wav

> *"We found that enlarging the receptive field was crucial to obtain samples that sounded musical. Even with a receptive field of several seconds, the models did not enforce long-range consistency which resulted in second-to-second variations in genre, instrumentation, volume and sound quality. Nevertheless, the samples were often harmonic and aesthetically pleasing, even when produced by unconditional models."*

SPEECH RECOGNITION

- 18.8 PER on TIMIT – "…the best score obtained from a model trained directly on raw audio."
- Required a mean-pooling layer after the dilated convolutions, for aggregating activations to coarser frames spanning 10 milliseconds (160× downsampling).

CONCLUSION

- WaveNets produce raw speech signals with highly-rated naturalness
  - Decent Training Time, Slow Generation
- Global conditioning can produce a single model that can be used to generate different voices, different instruments, etc.
- The same architecture shows strong results when tested on a small speech recognition dataset.

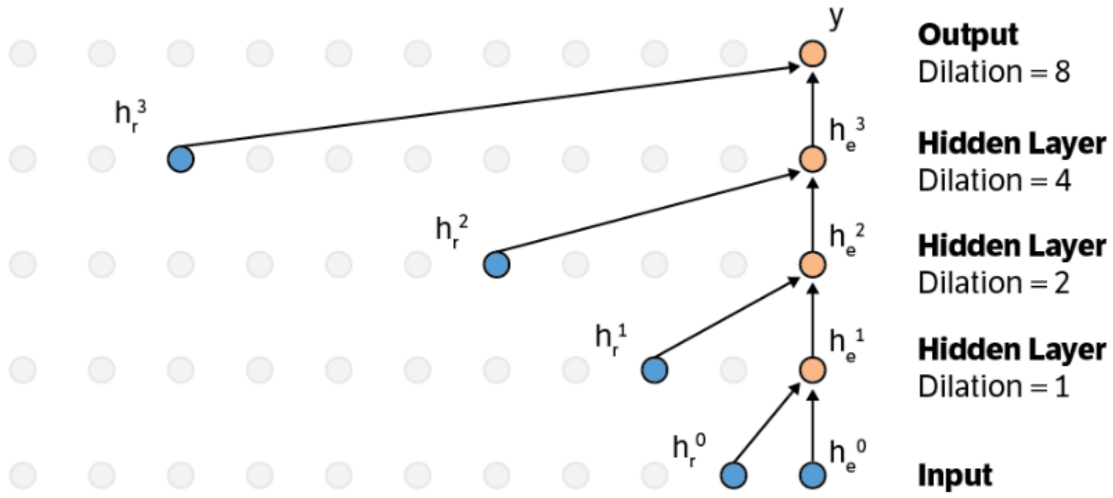
DISCUSSION

- General Thoughts…
- Moving closer to end-to-end $\rightarrow$ increasing generality.
- Pros over recurrent models?
  - When training?
  - When generating?
- Cons?

HIGH COST OF GENERATING SAMPLE BY SAMPLE

- Hours to generate just one second of audio.
- Solutions?

ELIMINATE REDUNDANT COMPUTATIONS



(Paine et al. 2016)

PROBABILITY DENSITY DISTILLATION

- Use a fully-trained WaveNet model to teach a smaller, more parallel student network ("High-Fidelity Speech Synthesis with WaveNet," n.d.).

- Train student to match teacher's distribution.

BIBLIOGRAPHY

"High-Fidelity Speech Synthesis with WaveNet." n.d. https://www.deepmind.com/blog/high-fidelity-speech-synthesis-with-wavenet.

Jozefowicz, Rafal, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. "Exploring the Limits of Language Modeling." arXiv. https://doi.org/10.48550/arXiv.1602.02410.

"Mu-Law Algorithm." 2023. *Wikipedia*, February.

Oord, Aaron van den, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. "WaveNet: A Generative Model for Raw Audio." arXiv. https://doi.org/10.48550/arXiv.1609.03499.

Paine, Tom Le, Pooya Khorrami, Shiyu Chang, Yang Zhang, Prajit Ramachandran, Mark A. Hasegawa-Johnson, and Thomas S. Huang. 2016. "Fast Wavenet Generation Algorithm." arXiv.