FloWaveNet: A Generative Flow for Raw Audio

Sungwon Kim¹, Sang-gil Lee¹, Jongyoon Song¹, Jaehyeon Kim², Sungron Yoon¹,³

¹Seoul National University, ²Kakao Corporation, ³ASRI, INMC, Institute of Engineering Research, Seoul National University
WaveNet

\[
\log p_X(x_{1:T}) = \sum_{t=1}^{T} \log p_X(x_t|x_{<t})
\]

https://deepmind.com/blog/wavenet-generative-model-raw-audio/
\[
\log p_X(x_{1:T}) = \sum_{t=1}^{T} \log p_X(x_t | x_{<t})
\]
Previous parallel speech synthesis models

- Pre-trained WaveNet
- Inverse Autoregressive Flows (IAFs)

**Probability Density Distillation**

\[
KL \left( P_S(x) \middle\| P_T(x) \right)
\]

Previous parallel speech synthesis models

Pre-trained WaveNet

Parallel sampling

Inverse Autoregressive Flows (IAFs)

Probability Density Distillation

$$KL(P_S(x) || P_T(x))$$

Previous parallel speech synthesis models

Pre-trained WaveNet

Inverse Autoregressive Flows (IAFs)

Parallel sampling

Probability Density Distillation

\[ KL(P_S(x) || P_T(x)) \] +

Power Loss
Perceptual Loss
Contrastive Loss
Frame Loss

Our Objectives

• Simplify the training procedure for parallel sampling

• Maintain the quality of speech samples
Our Objectives

• Simplify the training procedure for parallel sampling

• Maintain the quality of speech samples

Flow-based generative models for raw audio!
FloWaveNet

Training phase: \[
\log p_X(x_{1:T}) = \log p_Z(f_\theta(x_{1:T})) + \log \det \left( \frac{\partial f_\theta(x)}{\partial x} \right)
\]
FloWaveNet

Training phase
\[ \log p_X(x_{1:T}) = \log p_Z(f_\theta(x_{1:T})) + \log \det \left( \frac{\partial f_\theta(x)}{\partial x} \right) \]

Sampling phase
\[ z = z_{1:T} \sim P_Z(z) = N(0, I), \quad \hat{x} = f_\theta^{-1}(z) \]
FloWaveNet

Training phase

\[
\log p_X(x_{1:T}) = \log p_Z(f_\theta(x_{1:T})) + \log \det \left( \frac{\partial f_\theta(x)}{\partial x} \right)
\]

Sampling phase

\[
z = z_{1:T} \sim P_Z(z) = N(0, I), \quad \hat{x} = f_\theta^{-1}(z)
\]

Both the transformation \(f_\theta\) and \(f_\theta^{-1}\) are designed to be computed efficiently → Efficient training & Parallel sampling
\[
\log p_X(x_{1:T}) = \log p_Z(f(x_{1:T})) + \sum_n \log \det \left( \frac{\partial (f^n_{\text{AC}} \cdot f^n_{\text{AN}})(x)}{\partial x} \right)
\]
## Mean Opinion Scores

<table>
<thead>
<tr>
<th>METHODS</th>
<th>5-scale MOS</th>
<th>TEST CLL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ground Truth</td>
<td>4.67± 0.076</td>
<td></td>
</tr>
<tr>
<td>MoL WaveNet</td>
<td>4.30± 0.110</td>
<td>4.6546</td>
</tr>
<tr>
<td>Gaussian WaveNet</td>
<td>4.46± 0.100</td>
<td>4.6526</td>
</tr>
<tr>
<td>Gaussian IAF</td>
<td>3.75± 0.159</td>
<td></td>
</tr>
<tr>
<td>FloWaveNet</td>
<td>3.95± 0.154</td>
<td>4.5457</td>
</tr>
</tbody>
</table>

FloWaveNet $\geq$ Gaussian IAF
# Sampling speed

<table>
<thead>
<tr>
<th>METHODS</th>
<th>ITER/SEC</th>
<th>SAMPLES/SEC</th>
</tr>
</thead>
<tbody>
<tr>
<td>WAVENet</td>
<td>N/A</td>
<td>172</td>
</tr>
<tr>
<td>PARALLEL WAVENet</td>
<td>N/A</td>
<td>500K</td>
</tr>
<tr>
<td>GAUSSIAN WAVENet</td>
<td>1.329</td>
<td>44</td>
</tr>
<tr>
<td>GAUSSIAN IAF</td>
<td>0.636</td>
<td>470K</td>
</tr>
<tr>
<td>FLOWAVENet</td>
<td>0.714</td>
<td>420K</td>
</tr>
</tbody>
</table>

*FloWaveNet ≈ Gaussian IAF ≈ Parallel WaveNet  >>  Autoregressive WaveNet*

1000s times faster
Conclusion

• FloWaveNet produces **high quality audio samples** as well as previous distilled models.

• FloWaveNet synthesizes audio samples **in parallel**
  – w/o well pre-trained WaveNet (No distillation!)
  – w/o auxiliary loss terms

[QR Code for Demo page]
[QR Code for Code]

Poster 6/12 6:30 PM @Pacific Ballroom #2
Thank You!