WAVEGLOW: A FLOW-BASED GENERATIVE NETWORK FOR SPEECH SYNTHESIS

Ryan Prenger, Rafael Valle, Bryan Catanzaro

NVIDIA Corporation

Reviewer: Mark Baushenko

Glow + WaveNet = WaveGlow

INTRODUCTION

PROBLEM

WAVEGLOW

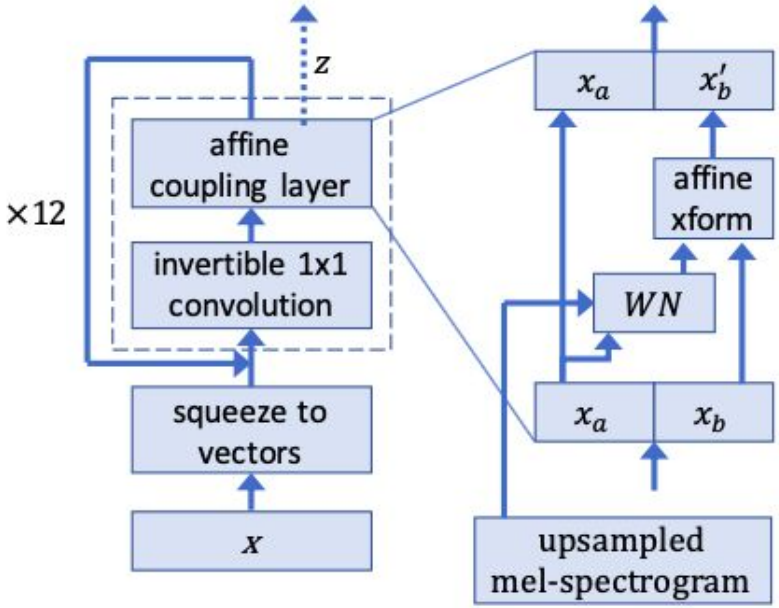

Fig. 1: WaveGlow network

Affine Coupling Layer

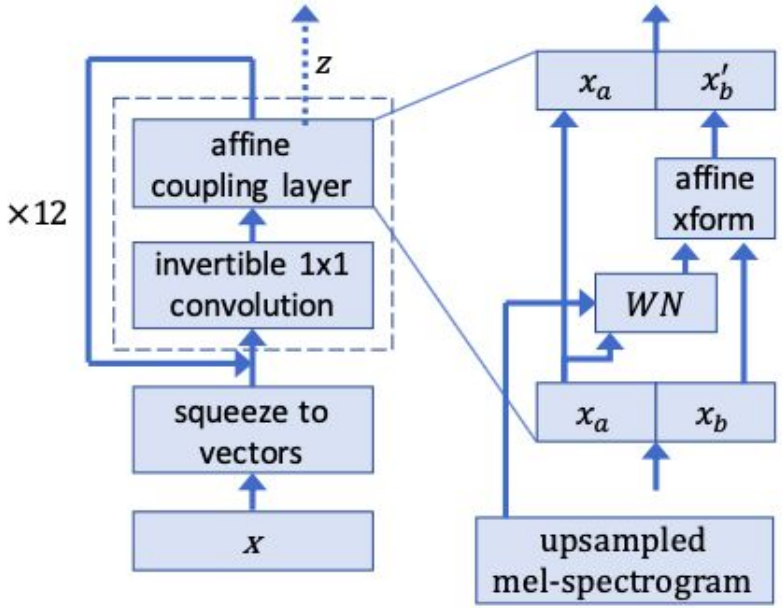

Fig. 1: WaveGlow network

$$\boldsymbol{x}_a, \boldsymbol{x}_b = split(\boldsymbol{x})$$

$$(\log \boldsymbol{s}, \boldsymbol{t}) = WN(\boldsymbol{x}_a, mel\text{-}spectrogram)$$

$$\boldsymbol{x}_b\prime = \boldsymbol{s} \odot \boldsymbol{x}_b + \boldsymbol{t}$$

$$\boldsymbol{f}^{-1}_{coupling}(\boldsymbol{x}) = concat(\boldsymbol{x}_a, \boldsymbol{x}_b\prime)$$

1x1 Invertible Convolution

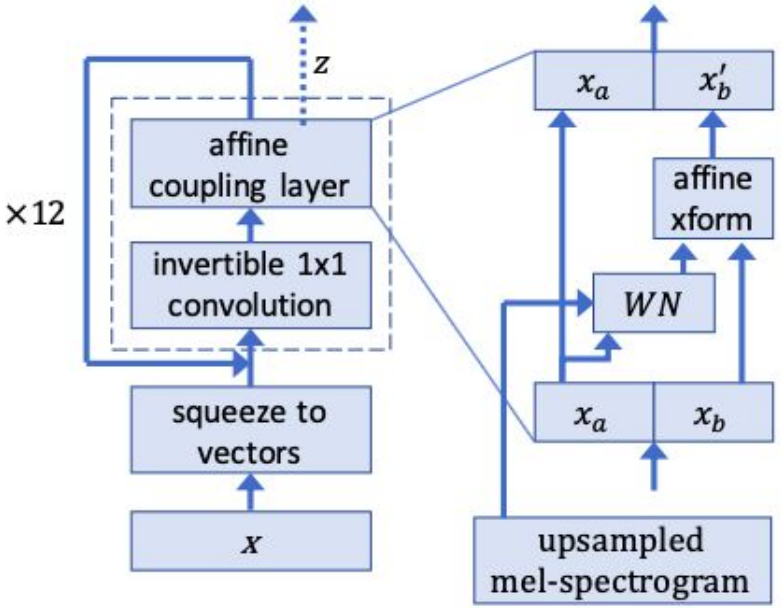

Fig. 1: WaveGlow network

$$\boldsymbol{f}_{conv}^{-1} = \boldsymbol{W}\boldsymbol{x}$$

$$\log|\det(\boldsymbol{J}(\boldsymbol{f}_{conv}^{-1}(\boldsymbol{x})))| = \log|\det \boldsymbol{W}|$$

1x1 Invertible Convolution

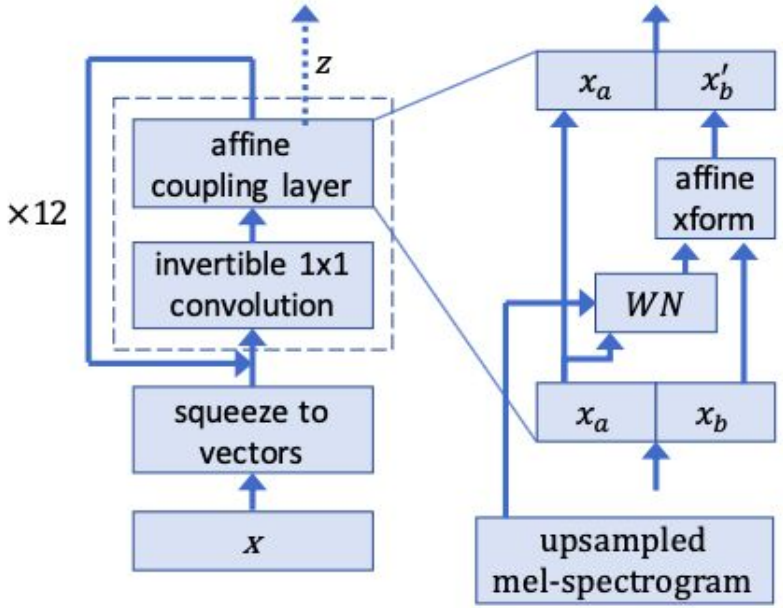

Fig. 1: WaveGlow network

$$\boldsymbol{f}_{conv}^{-1} = \boldsymbol{W}\boldsymbol{x}$$

$$\log|\det(\boldsymbol{J}(\boldsymbol{f}_{conv}^{-1}(\boldsymbol{x})))| = \log|\det \boldsymbol{W}|$$

$$\log p_\theta(\boldsymbol{x}) = -\frac{\boldsymbol{z}(\boldsymbol{x})^T\boldsymbol{z}(\boldsymbol{x})}{2\sigma^2} + \sum_{j=0}^{\#coupling} \log \boldsymbol{s}_j(\boldsymbol{x}, \textit{mel-spectrogram}) + \sum_{k=0}^{\#conv} \log\det|\boldsymbol{W}_k|$$

Audio quality comparison

Model	Mean Opinion Score (MOS)
Griffin-Lim	3.823 ± 0.1349
WaveNet	3.885 ± 0.1238
WaveGlow	3.961 ± 0.1343
Ground Truth	4.274 ± 0.1340

Thanks for attention