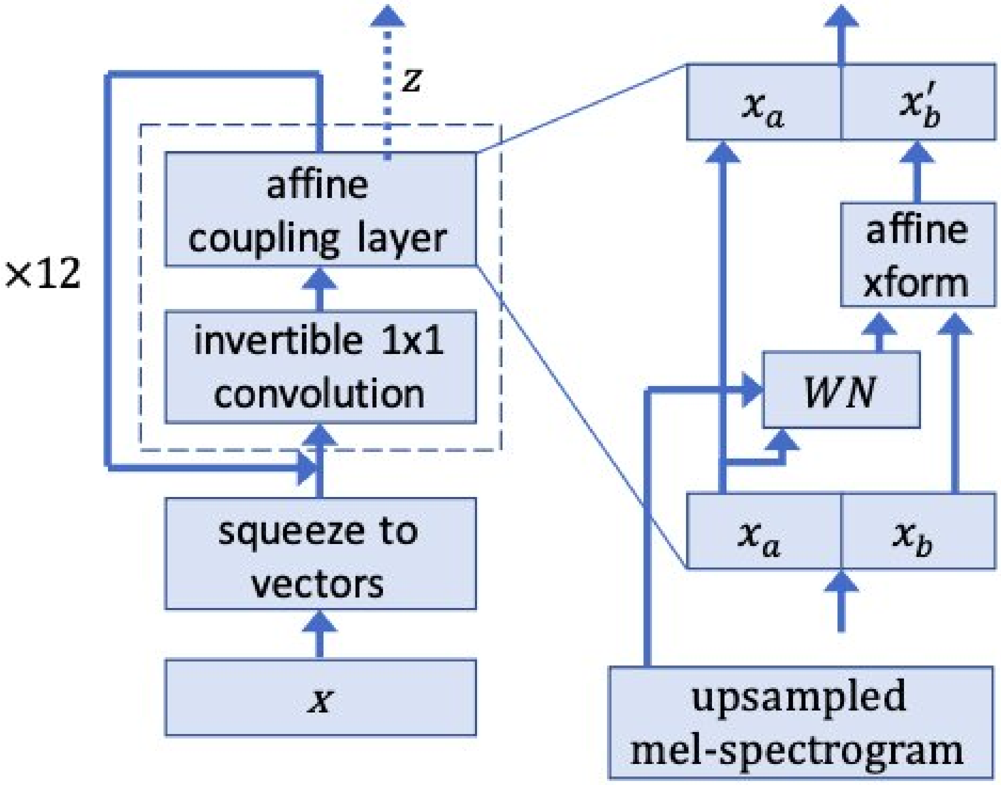
Tech Xplore

WaveGlow: A flow-based generative network to synthesize speech

November 19 2018, by Ingrid Fadelli



WaveGlow network. Credit: Prenger, Valle, and Catanzaro.

A team of researchers at NVIDIA has recently developed WaveGlow, a flow-based network that can generate high-quality speech from melspectrograms, which are acoustic time-frequency representations of sound. Their method, outlined in a paper pre-published on arXiv, uses a single network trained with a single cost function, making the training procedure easier and more stable.

"Most neural networks for synthesizing speech were too slow for us," Ryan Prenger, one of the researchers who carried out the study, told TechXplore. "They were limited in speed because they were designed to only generate one sample at a time. The exceptions were approaches from Google and Baidu that generated audio very quickly in parallel. However, these approaches used teacher networks and student networks and were too complex to replicate."

The researchers drew inspiration from Glow, a flow-based network by OpenAI that can generate high-quality images in parallel, retaining a fairly simple structure. Using an invertible 1x1 convolution, Glow achieved remarkable results, producing highly realistic images. The researchers decided to apply the same idea behind this method to speech synthesis.

"Think of the white noise that comes from a radio not set to any station," Prenger explained. That white noise is super-easy to generate. The basic idea of synthesizing speech with WaveGlow is to train a neural network to transform that white noise into speech. If you use any old neural network, training will be problematic. But if you specifically use a network that can be run backwards as well as forwards, the math becomes easy and some of the training issues go away."

The researchers ran speech clips from the training dataset backwards, training WaveGlow to produce what closely resembles white noise. Their model applies the same idea behind Glow to a WaveNet-like architecture, thus the name WaveGlow.

In a PyTorch implementation, WaveGlow produced audio samples at a rate of over 500kHz, on an NVIDIA V100 GPU. Crowd-sourced mean opinion score (MOS) tests on Amazon Mechanical Turk suggest that the approach delivers audio quality as good as the best publicly available WaveNet method.

"In the speech synthesis world, there is a need for models that generate speech more than an order of magnitude faster real time," Prenger said. "We're hoping WaveGlow can fill this need while being easier to implement and maintain than other existing models. In the deep learning world, we think that this type of approach using an invertible neural network and the resulting simple loss function is relatively under-studied. WaveGlow provides another example of how this approach can give high-quality generative results despite its relative simplicity."

WaveGlow's code is readily available online and can be accessed by others looking to try it or experiment with it. Meanwhile, the researchers are working on improving the quality of synthesized audio clips by fine tuning their model and carrying out further evaluations.

"We haven't done a lot of analysis to see how small of a network we can get away with," Prenger said. "Most of our architecture decisions were based on very early parts of training. However, smaller networks with longer training time might generate sound that is just as good. There are a lot of interesting directions this research might go in the future."

More information: WaveGlow: A flow-based generative network for speech synthesis. arXiv:1811.00002 [cs.SD]. arxiv.org/abs/1811.00002

Glow: generative flow with invertible 1x1 convolutions. arXiv:1807.03039 [stats.ML] arxiv.org/abs/1807.03039

github.com/nvidia/waveglow

Citation: WaveGlow: A flow-based generative network to synthesize speech (2018, November 19) retrieved 20 March 2024 from https://techxplore.com/news/2018-11-waveglow-flow-based-network-speech.html