Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions

date 16.02.2018, paper on [arxiv (https://arxiv.org/pdf/1712.05884.pdf)](https://arxiv.org/pdf/1712.05884.pdf)

It is recommended to check tacotron review first, since some concepts were explained there.

Overview

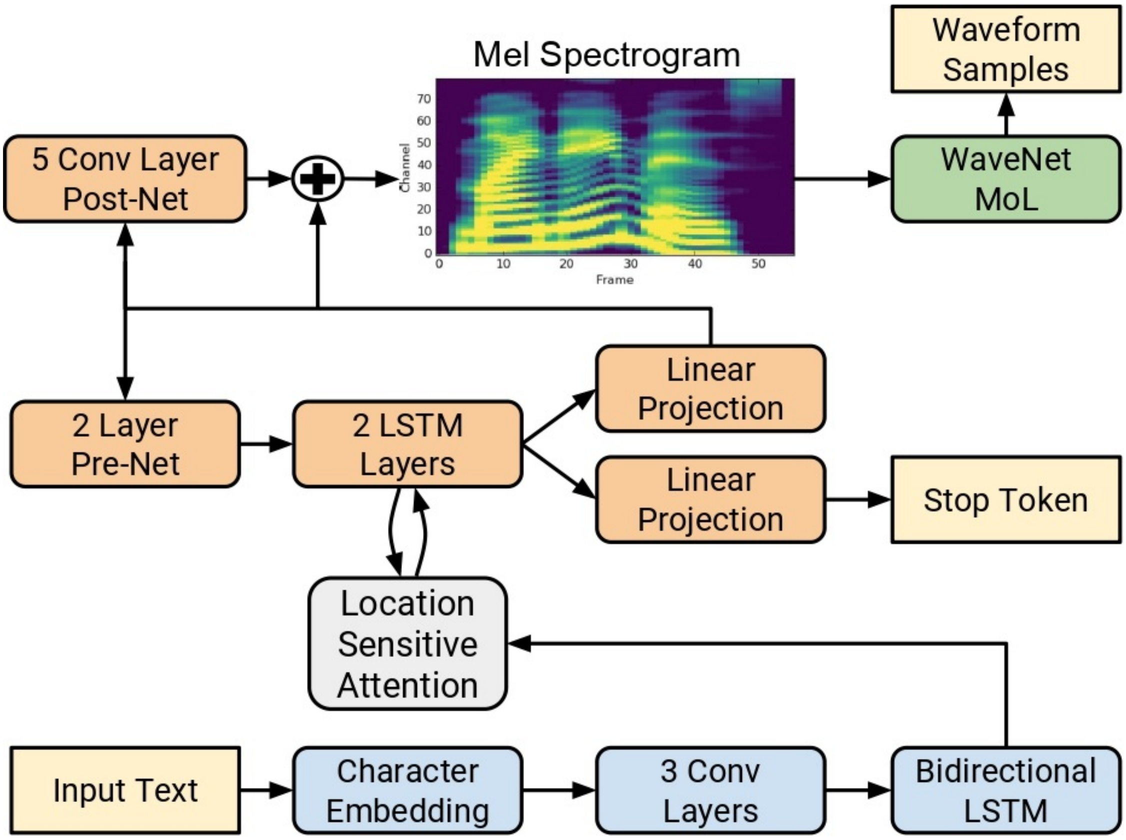
Authors proposed modifying tacotron architecture and using generative neural network (say, wavenet) instead of Griffin-Lim algorithm for synthesis. **In this report only tacotron modification are covered**, and generative part is not, since in github authors uses WaveGlow model for generating audio, and it is separate article.

Model can be splitted into several parts:

- Encoder;
- Decoder with attention;
- Post-Net;

Model takes as input one hot encoded characters and outputs limited sequence of 80 band mel spectrograms (so length of generated text is limited, but this is not an issue since current hidden vectors could be used for next sentence generation, but most common way - to generate not so long sentences separately, ignoring dependencies between them). Model also outputs probability of stop token - indicator that model outputted all required mel spectrograms for speech generation.

Overall architecture is shown below:



Encoder

Encoder takes as input sequence of learnable character embeddings (same as tacotron encoder does) and first perform convolutions on time axis preserving spatial dimension (same as CBHG convolutions in tacotron). After convolution is done working dimension (dimension for matrix multiplication in linear layers) is embedding dimension and convolutions outputs is passed through bidirectional LSTM (same as GRU in tacotron).

Decoder with attention

Decoder takes as input previous generated frame and passes it through pre-net(two linear layers with relu and dropout). The encoder output is consumed by an attention network which summarizes the full encoded sequence as a fixed-length context vector for each decoder output step. Context vector for each time step is constructed as following (let α_i be the alignment from previous step, $(h_1, \ldots, h_L)$ - encoder outputs, s_i - attention rnn hidden vector (same as tacotron)):

1. First, we extract k vectors $f_{i,j} \in \mathrm{R}^k$ for every position j of the previous alignment α_{i-1} by convolving it with a matrix $F \in \mathrm{R}^{k \times r}$:

$$f_i = F * \alpha_{i-1}.$$

2. These additional vectors $f_{i,j}$ are then used by the scoring mechanism $e_{i,j}$:

$$e_{i,j} = w^{\top} \tanh(Ws_i + Vh_j + Uf_{i,j} + b)$$

3. \begin{align} \alpha_{i,j} =

```
    \exp(e_{i,j}) \left/
    \sum\limits_{j=1}^L \exp(e_{i,j}) \right..
```

   \end{align}
4. Attention context on step i is calculated via:

$$g_i = \sum_{j=1}^{L} \alpha_{i,j} h_j$$

$\boldsymbol{F, U, V, W, b, w}$ **are trainable parameters** After that (g_i, s_i) are concatenated and passed through decoder RNN (2 layer LSTM). Its output transformed via two linear transformations, first is used for stop token prediction, latter - is used at not refined mel spectrogram at current step (refining is done in Post-Net).

Post-Net

Small 1-d convolutional neural network with dropout (unusual for CNN to use dropout) and non-linearities preserving spatial domain, convolutions is done on time axis domain, using band axis as channel axis (same as postnet in tacotron). Post-net takes as input decoder outputs after all sequential steps (same as tacotron postnet).

Training

During training groundtruth mel-spectrogram of previous frame used as input to decoder (for the first frame 0 spectrogram is used). L2 loss was used on mel spectrograms produced at each decoder timestep and postnet outputs. Binary cross entropy loss was used for stop token.

Testing

During testing previous generated by decoder mel-spectrogram is used as input. Generation is done until probability of end token is below 0.5 or until max decoder step is reached (this is uncommon situation and must be the case only for very very very long sentences).

Experiments and pros and cons

Experiments and pros and cons aren't covered in this review since they are highly depended on generative model, which is conditioned on generated by tacotron2 mel spectrograms.