Hierarchical Cross-Modal Talking Face Generation with Dynamic Pixel-Wise Loss

Summary and review by - Purvanshi Mehta

September 23, 2019

1 Summary

The paper proposes an ATVG-net(AT-net and VG-net) which forms high level representations of audio signals to generate video frames. The Audio Transformative Network (AT-net) is an encoder decoder structure on the audio features combined with the landmarks of the original image frame. The PCA reconstruction of the decoder output gives the *audio video correlated landmarks.* The VG-net takes the assumption that the distance between generated and original landmarks represent the current and example image frame, the difference between them concatenated with the original image gives the modified current image frame. This input is fed into the Multi-Modal Convolutional-RNN(MMCRNN) which gives the current image feature for video frame generation. Both the networks have been trained in a decoupled manner. To disentangle the motion part from the audio visual non correlated regions an attention mask is applied to the combined output from the MMCRNN, original landmarks, original image and generated landmarks. The attention weights generated are used as the weightage to see how much does each pixel contributes to the loss. This is addressed as the *Attention based Dynamic Pixel wise loss* in the paper.

The *discriminator regresses landmarks* paired with input frames (from original or generated) along with giving a discriminative score for the sequence. The landmarks help to evaluate the input image in a frame wise manner. The GAN loss is a combination of the mean squared error between predicted and ground truth landmarks and the discriminative score. The full loss can be seen as a combination of the dynamic pixelwise and GAN loss through a relative importance parameter.

The paper presents experimental results on two datasets, user studies and ablation studies to justify their approach. Experimental evaluations have been performed on the LRW and GRID dataset. The ATVGnet shows clear state of the art results over previous methods. For measuring the performance of the generated video and not only the image frames the authors conduct a user based study where each user is asked to rate a video (shuffled with other method videos) based on whether it is realistic and whether the audio is synced with the lip movement.

2 Review

- The author claims - 'there are additional novel examples of synthesized facial movements of human cartoon characters that are not a part of any dataset', the paper does not clearly justify the differences in quantifiable terms. The term 'robustness' associated with such experiments has not been justified properly. According to the Figure 5 the cartoon images look similar to any other dataset used - GRID and LRW.

- The fusion between audio and landmark features has been done by a simple concatenation. Have other fusion techniques - like tensor product between two modalities been explored?

- The spelling of cartoon in Figure 5 has been misspelled as caetoon.

- The paper does not provide a description of handling of head movements in the AT-net. The open ended question which remains in my mind is - How is the audio combined with the landmark able to generate the small head movements.

- The reason for training the networks (AT-net and VG-net) in a decoupled manner has not been stated in the paper.