Artistic Stylization of Images and Video
CPSC 490 Project Proposal
Proposer: Jeff Ding
Advisor: Prof. Holly Rushmeier

Summary
When analyzing an image's visual attributes, one can distinguish its content and style. In this context, content is the scene referents of the image, or in other words, physical objects such as people, trees, buildings, etc. and the spatial relationships between them, while style is image's constituent visual qualities such as texture, color, lines, patterns, etc.

The aim of this project is to create a program that can, given a source image or video, generate a video that retains the content of the original image but whose style reacts in time to audio. The program take another image as a source for stylization and applies the style to the content of the original image. This will be primarily accomplished through the application of artificial neural networks and deep learning.

Background
Convolutional neural networks (CNNs) are a class of neural networks whose arrangement of neurons are inspired by the organization of the visual cortex in animal brains. CNNs consist of layers of convolution filters that extract specific features from images. Each layer can be thought of as a feature map of the image whose specificity varies depending on the layer. The layers are arranged in a feed-forward manner so that layers further along in the processing hierarchy are representations of the image with more explicit object information. Thus, the deepest layers represent the most complex features (imagine a neuron that only activates if a tree is in the image) and so CNNs are often used for image classification tasks.

With the popularity of programs such as Google DeepDream and the mobile application Prisma, the possibility for CNNs to stylize images in creative and exciting ways is becoming apparent. The key takeaway is that a CNN’s layers are differentially specific feature maps of the image, and that sufficiently complex features are analogous to objects. In other words, a subset of layers in the CNN is effectively a representation of content in the image. Moreover, the representations of content and style in CNNs are separable and a method can be developed to create new images that mix the content and style from different sources.¹

The method is based on the idea that an image can be reconstructed from an initialization of white noise via backpropagation. To represent the notion of distance from the desired image, a loss function is formulated, whose gradient with respect to the image can be calculated. Gradient descent can then be performed to alter the noise image until the loss function is minimized (and thus, the resulting image is closest to the desired one). More concretely, the loss function can be formulated as the mean squared error between matrix representations of certain aspects of the image. Depending on which aspect, the manner of reconstruction differs.

Consider the layer activations in a CNN representation as one such aspect. Including every layer in the function will completely reconstruct the image, as every feature map at every specificity
will be matched. However, limiting the function to the deepest layers reconstructs only the highest level features (objects), thus resulting in a content representation of the desired image.

Now, instead of the direct feature map activations, consider using the correlations between feature maps for the error. This discards high level object information from the reconstruction and only considers the low level information such as lines, shapes, and patterns. The reconstruction will only retain such information that remains invariant between complex features, thus being a style representation of the desired image.

With the content and style representations separable, a new image can be constructed that mixes the former from one image source and the latter from another. This is done similarly by gradient descent on noise, however the general loss function is now a weighted sum of the direct loss function for the content source and the correlation loss function for the style source. The ratio of the weights determines the affinity of the constructed image to preserve more original content or become more stylized.

**Description**
The aim is to render the stylization of the constructed image in such a way that it reacts in time to audio. Thus, the output will be a video. For simplicity, assume that extending the construction to video only requires constructing each frame independently (note that this also allows for video inputs). Then, a naive method for audio reactivity would be to alter the ratio of weights in the loss function for each frame in the sequence proportional to some aspect of the audio that varies in time, such as amplitude. One could imagine an effect where the stylization intensity of the output video pulses in time to music.

However, as the extraction of feature maps from the source image or video provides scene-specific content information, a more visually interesting approach would be to differentially style features based on different aspects of the audio that vary in time. In order to do this, a mapping can be made between the feature maps to be styled and the audio aspects of interest. Then, a loss function can be formulated that is similar to the direct loss function for content reconstruction, only now each feature is assigned a weight of error contribution that fluctuates depending on its linked audio aspect. This function can be added to the general loss function as a weighted third term.

In other words, the general loss function has independent components for content, style, and audio reactivity in linear combination. As mentioned earlier, the ratio of the content and style weights determines the stylization intensity. Now, the ratio of the content and audio weights determines the intensity of feature differentiation for audio reactivity. Finally, the ratio of the style and audio weights determines the intensity of style fluctuation to audio.

It would be interesting to completely do away with a separate image source for style, and have the construction retain the content of the original image but derive its style entirely from audio. This creates a synesthesia effect where the image style has only auditory basis and no visual basis. One method would be to feed loss function the same source for its content and style components. Then, the reconstruction will resemble the original, only the explicitness of each feature will vary differentially with respect to audio. Another method would be to use a procedural texture as a
style source, whose generation inputs vary with audio. It is difficult to imagine what these results might look like, let alone if they will be visually appealing, but these ideas are worth exploring.

Finally, the extension of the image construction process to video is not as trivial as treating each frame as an independent image, as the loss function is non-convex and construction by gradient descent terminate at local minima. Due to the noise initialization, the final construction is not stable. Thus, there will be significant flickering if frames are independently constructed. For image sources, the solution is as simple as initializing each frame with the stylized image of the previous frame. Then, the only variations between frames will be due to audio reactivity effects. For video sources, a more complex method must be used, as any motion between frames will be initialized incorrectly by simply using the previous stylized frame. This method primarily involves calculating the optical flow between frames and penalizing deviations from point trajectories. Long term motion estimates are calculated to exempt deviations due to object occlusion. Such penalization is formulated and added as a term to the general loss function to enforce frame consistency.

**Deliverables**

- Source code for the command line program
- Video demos and original inputs

**Citations**