We explored the idea of harnessing the power efficiency of Spiking Neural Network components with the state-of-the-art accuracy that a Vision Transformer can offer by combining them. The result was a Hybrid SNN ViT model.

Datasets
- Developed hybrid architecture on MNIST Dataset
- Grayscale 28 x 28 Images
- Moved to CIFAR10 to evaluate model on color
- RGB 32 x 32 Images
- Both have 50,000 Training / 10,000 Validation Images

Approach
- No Time:
  - This is the most basic form of the Hybrid SNN ViT. It incorporates no aspect of time and trains the model based on a single rate encoding of each image.
- Time Averaged Output (TAO):
  - This version takes a time-step parameter in the training and evaluation functions. For each number of time-steps specified, the model creates a rate encoded image and gathers the outputs. It then averages those outputs and sends the result to the loss function where it is compared with the target.
- Time Averaged Loss (TAL):
  - Time Averaged Loss is similar to Averaged Output in that it takes time as a parameter. In this implementation, it calculates a full output for each rate encoding of the input image and calculates the loss for each. It then takes the average of those losses before performing back-propagation.

Model Architecture
- We started by finding a ViT model online, which used the GELU activation function, so we could establish a baseline
- Changing the final MLP Head to use SNN components greatly weakened the model, but it remained relatively stable when we converted the Transformer Block

Results
- Below are the validation accuracies collected during training:

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Base ViT</th>
<th>Hybrid No Time</th>
<th>Hybrid TAO</th>
<th>Hybrid TAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST</td>
<td>97.94%</td>
<td>97.36%</td>
<td>97.27%</td>
<td>97.29%</td>
</tr>
<tr>
<td>CIFAR10</td>
<td>85.25%</td>
<td>67.23%</td>
<td>69.04%</td>
<td>68.17%</td>
</tr>
</tbody>
</table>

Future Work
We are very excited to explore a new avenue of the model that appeared as a byproduct of our work developing the Hybrid SNN ViT
- PGD Attacks are a white-box adversarial attack that uses the known model weights to slightly alter an image such that the model incorrectly classifies it, but a human can not tell the difference.
- We hypothesize the act of Rate Encoding an image fed to the Hybrid SNN ViT is enough to completely eliminate the presence of threat from adversarial noise in an image.
- Because Rate Encoding uses pixel values as probabilities, while the slight changes may be enough to fool a regular model, they are not enough to statistically change a rate encoded image, especially over multiple time-steps.