Liaison CS Meeting Agenda – Thursday 10/3/24

1. Technical updates - 20-30min
   1. Random forests model (**Zach**, Lillian, Devanshi)
      1. Background
      2. Preprocessing
      3. Results
      4. Issues
      5. Next steps
   2. Temporal convolutional network (**Milo**, Mehrezat)
2. Upcoming deadlines
   1. Fall Presentation (internal), October 29
   2. Midyear Update, December 6
3. Liaison items

Liaison CS Meeting Minutes, Thursday 3 October 2024

Present: Everyone

Random Forests (Zach)

* Worked to replicate random forests model from Willet et al.
* Random forests are randomly generated decision trees each of which is trained on different features of the data set
* Divide waveform into 1 second chunks, extract various features from each chunk
  + Most common label in the chunk: what we’re trying to predict
  + voltage mean and standard deviation
  + top 6 frequency components, using Fourier transform (ignoring DC component)
  + used pre-rectified voltage
* Results:
  + Confusion matrix shows actual label against predicted label
  + M and L labels get mixed up sometimes
  + Rarely predicts J, K, N, Z because there are few examples in the data set
  + “Barcoding” issue: incorrect labels in short segments within section dominated by another label
  + J and K ignored because durations are short
  + Model is “shortsighted,” looking at 1 second at a time
  + Could possibly add postprocessing based on label orders we know must happen
  + Poor performance with repeated probes
  + Model confidence usually drops when labels are incorrect
* Questions
  + Dr. Reif
    - How easy is it to integrate rules to prevent illegal transitions?
    - Can we provide rules based on input resistance?
  + Zach
    - More flexible if we can avoid hardcoding rules, and instead include the desired data in the input to the random forest model
    - Input resistance was not found to be useful in improving predictive performance for this model. Will certainly consider using it in future models though.
  + Dr. Cooper: model struggles with L-M and M-N transition, these are also the places she struggles with labeling
  + Dr. Backus: with higher Ri level, most plant-feeding waveforms are negative-going as well as positive-going — so why not consider the DC offset?
  + Zach: we are still considering the DC component through the mean of the chunk

Convolutional Neural Networks (Milo)

* CNNs (convolutional neural network) commonly used for images
* convolutional layers extract features, create an abstract representation of the input data
* Work by combining layers of filters and ultimately getting a classification for a waveform
* This is a form of deep of learning, in that there are lots of layers (i.e. it has depth) and the model is allowed to decide on what features are important
* CNNs are a fairly advanced technique and it is not entirely important to understand exactly why they make the decisions that they do
* CNNs are at 80% accuracy, although this isn’t the most important way to characterize performance.
* CNN training is currently limited by our computational resources, Milo is working on making his code more efficient so he can implement methods that have a better idea of time dependance. One problem with this approach could be that it will require more data than we have available, but we are investigating methods of dealing with this potential problem.
* Questions
  + Dr. Backus: Do we envision blending these two models in any way? Or do we see them as alternatives where we go with one or the other?
  + Milo: It remains to be seen. If they both do really well, we should combine the models if possible. Otherwise, it may make sense to only go with one method compared to the other.
  + Dr. Reif: Do we know what waveforms the CNN had poor performance on? Are they similar to the places that Random Forests had issues?
  + Milo: It is very likely that it had issues in the same places, and this is an issue we are trying to fix with further models.
  + Dr. Reif: How are we doing in terms of computational power? Is everything too complex?
  + Milo: Computational power is mainly a limitation because the CNN code is very inefficient. Milo emphasizes that complexity isn’t necessarily a thing that limits us, it mostly just speaks to the flexibility of the model
  + Dr. Reif: Could the model’s transitions be compared to human labels?
  + Milo: They could, although we would need to be sure our model is very trustworthy.
  + Dr. Cooper: She is optimistic that the addition of context will improve performance a lot
  + Dr. Backus: She thinks it would be useful for models to be able to act in advisory role on new data. This could allow for better standardization across naming standards.
  + Dr. Reif: There should be a way to edit outputs when a researcher and the model disagree. Also researchers should be able to set their own names for waveforms.
  + General agreement that being able to edit names and define semantics would be a nice thing for our tool to have.
  + Dr. Cooper: Excited to see what these models can do on Midge data since they don’t have very well defined waveform boundaries in terms of voltage drop.

Conclusions

* Clinic Team will make sure to send out slides in a four per page format with our minutes
* Most important upcoming deadline is the Midyear Report Due on December 6
* Clinic Team needs to discuss whether it would be useful for us to visit Parlier sometime in the early second semester and communicate with Dr. Backus about that before December 6 or so.