III. Model documentation and write-up

Information included in this section may be shared publicly with challenge results. You can respond to these questions in an e-mail or as an attached file. Please number your responses.

1. Who are you (mini-bio) and what do you do professionally?

   If you are on a team, please complete this block for each member of the team.

   Ming Feng received his bachelor degree from Ningxia University, Ningxia, China, on July 2018. Currently, he is a Ph.D. student at Tongji University, Shanghai, China. His current research interests include medical image processing, weakly supervised learning.

   Kele Xu received his Ph.D. degree from Université Pierre et Marie CURIE, Paris, France, on Dec. 2016. Currently, he is an assistant professor at National University of Defense Technology, Changsha, China. His current research interests include applied machine learning, silent speech interface (SSI) and multi-modal machine learning. He is also interested in the applications of machine learning for audio signal processing, time series analysis and medical image processing.

   Li Tingzhen received her bachelor degree from Xiamen University, Xiamen, China, on July 2019. Currently, she is a graduate student at Tongji University, Shanghai, China. Her current research interests include handwriting recognition and medical image classification.

2. What motivated you to compete in this challenge?
   Tissuenet competition provides very precious data. I want to explore the boundaries of my algorithm and compare the results with other top players.

3. High level summary of your approach: what did you do and why?
   First, we trained a patch-level deep learning classification model to generate probsmap from whole slide images. Second, we extract feature from probsmap. Last, we feed the feature to a mechine learning model to make wsi-level decisions.
   All our method based on above pipeline. The above pipline can achieve good results. Then we paid more attention to data understanding. It turns out that the understanding of the data is very important.

4. Copy and paste the 3 most impactful parts of your code and explain what each does and how it helped your model.
   1. Extract class-0 patches from labeld 0 wsi. It can greatly improve 01 discrimination. It can enhance the ability to distinguish between 0 and 1.
   2. Extract tissue mask from wsi. The traditional tissue area extraction method(ostu) does not work well on some images, so we have made improvements. At the same time, we use appropriate levels of wsi to extract features to ensure that all tissue regions can be extracted.

3. Efficientnet. If we use Efficientnet, the classification results will be better. Because of actual environmental constraints, we actually still use densenet201.

5. Please provide the machine specs and time you used to run your model.
   - CPU (model): Intel E5-2697 v4
   - GPU (model or N/A): A RTX 2080Ti
   - Memory (GB): 11G
   - OS: Mac OS
   - Train duration: About 3h.
   - Inference duration: 1.5h per 1000 samples

6. What are some other things you tried that didn't necessarily make it into the final workflow (quick overview)?
   1. Pesudo label. I tried it, and it seems to have little effect on the result. In other words, the right number of times will help.
   2. Sampling method of category 0 patch. It seems that random sample can get a better result.

7. Did you use any tools for data preparation or exploratory data analysis that aren't listed in your code submission?
   No.

8. How did you evaluate performance of the model other than the provided metric, if at all?
   Accuracy. But our model is optimized based on the indicators provided by the competition.

9. Anything we should watch out for or be aware of in using your model (e.g. code quirks, memory requirements, numerical stability issues, etc.)?
   Basically no. You can contact me at any time if you have any questions during training.

10. Do you have any useful charts, graphs, or visualizations from the process?
    No.

11. If you were to continue working on this problem for the next year, what methods or techniques might you try in order to build on your work so far? Are there other fields or features you felt would have been very helpful to have?
    A more reasonable pseudo-label label method.
    Recent weakly supervised learning methods may also help