Weekly Schedule Schedule: Sunday, 2:00 PM

Our github repo: alckasoc/Team-Chiken-wi22: The official ACM AI Team Chiken repository. (github.com)

2/13/2022 Meeting Time: Sunday, 2:00 pm
Attendees: Nathan, Derrick, Min, Vincent

Summary of Meeting

- Digitally generated math formulas -> LaTeX
- Laid out timeline for coming 3 weeks
    - By 2/26 -> a clean dataset to work with (upload to Kaggle and make it public, include it in our repo’s README, drop it in chat too)
    - By ⅗ -> set up ML pipeline and run with a basic model (this part is just for setting up a basic script and to make sure everything is working)
    - By 3/12 -> experimenting! Try stuff out. Make changes to the pipeline and observe performance changes etc.

Action Items

1. Review how the following repo generates data: lukas-blecher/LaTeX-OCR: pix2tex: Using a ViT to convert images of equations into LaTeX code. (github.com)
2. Reproduce how they made the data (upload to Kaggle, notify everyone in the team)
    a. Everyone should get a good understanding of how the data is generated but only one person needs to ultimately upload and share the dataset
3. Set up some preprocessing! (This part will probably require you guys to meet up and talk about how it will be done)
    a. On a high level:
        i. First, we need a way to take the math formula image and turn it into a sequence of tokens (so we need some vision model to see the math symbols and convert them to a sequence of tokens, probably I could be wrong)
        ii. Now with a sequence of tokens, we can train an NLP model to produce LaTeX

Timeline

1. Getting started with deep learning and improving proficiency
2. Diving into Deep NLP (roughly 2-3 weeks with #1)
3. Start the project
    a. Debrief the problem
        i. Any related works?
        ii. Data ingestion? Where are we getting the data?
        iii. What framework?
    b. Explore the problem

        i. Given a difficult problem, get everyone up to speed on just the bare bone basics.
        ii. Once everyone is comfortable with working with the field which the project pertains to, start to explore further (e.g. have other people tackled this problem?)
    c. Start the technical process!
        i. Refer to Technical Process.

4. Mid-quarter project-progress presentation
    a. Y’all are simply presenting on your progress and what you’ve learned
    b. Use a slideshow!
5. Continue (and eventually finalize) technical process
    a. Usually the mid-quarter project-progress presentation checkpoint is early in the stages of a project
    b. Past that checkpoint, we get into the nitty gritty
6. Expand our horizons and turn this into an app (if time allows)
    a. Not sure if we have time, but if we do and we have the relevant expertise, then we can definitely beautify this project
7. ACM Project Showcase
    a. ACM will be having a project showcase (which is ages from now, but still nice to include here)
    b. Slideshow, visuals, demos, and thorough explanations of your entire project from when you started to the finished product!
    c. Talk about your difficulties, your many approaches, what you’ve done, what y’all have discussed, how you fixed issues, what you might do going forward, etc.

Timeline - Technical Process

(This is my sketch of the project)

1. Flush out the problem statement
    a. Are we using deep learning, if so what aspect of deep learning?
    b. Is this problem feasible? Will it require rare data?
    c. Where do we find the data?
    d. Is this problem suited for AI?
2. Find reliable dataset(s)
    a. Resources: Kaggle, UCI, other ones available online
    b. Is this dataset suited for our task?
    c. What is it missing? Pros and cons?
3. Wrangle the data
    a. Are there missing values? Are there missing features?
    b. Are there some underlying problems in the dataset?
4. Preprocess the data
    a. Transform the data into a format your model will like!
5. Modeling
    a. Custom model or use off-the-shelf/ready-to-use models

    b. What model should we be using? Why did you pick that model? Pros and cons?
    c. Just from learning about how the model works, how do we improve this model? What are its weaknesses?

6. Inference/Validation
    a. Inference and validation are different things!
    b. Validation is tied to the model training loop and inference is getting a few new samples of predictions from the model
    c. I put both of them here because you monitor your model’s performance through validation and also sometimes inference
    d. Let’s see some results and samples.
    e. Dive deeper into its weaknesses and perform some sort of error analysis!
7. Repeat
    a. There are many components to your ML pipeline (points 2-6 above) and also many other subcomponents (that may be excluded from the above points)
    b. Find a concrete way to diagnose the weaknesses of your *pipeline* (not model but the entire pipeline from ingesting the data to making predictions) and adjust it accordingly
        i. Maybe the model is underfitting or the hyperparameters are wack
        ii. Maybe the dataset is poor in quality
        iii. Maybe your preprocessing pipeline needs to be improved
        iv. etc