Project Proposal
CSCI 2270 Spring 2024

Title: DishDash: Dishing out Delicious Recommendations with Visual-LM Enhanced Recipe Discovery

Group members:
- Alexander Wang
- Muskaan Patel
- Vincent Zhu

Motivation and Description

- **Problem:** The abundance of online recipe content presents a challenge for users seeking personalized recipe recommendations tailored to their preferences and constraints. Existing recommendation systems often struggle to provide relevant suggestions due to limitations in understanding and analyzing unstructured recipe data. Additionally, users may have specific dietary restrictions, ingredient preferences, or time constraints that need to be accommodated for effective recommendations. For example, users might be unable to follow a cooking video due to lack of specific ingredients while looking for available replacements.
- **Solution:** Our proposed solution is to develop a recommendation system that leverages Visual-LM models to analyze cooking videos and vlogs, extracting rich textual descriptions of recipes and cooking techniques. By combining this data with structured recipe datasets, we aim to create a comprehensive database of recipes enriched with textual descriptions.
- **Proposed Methodology:**
    - We will collect diverse cooking videos and vlogs from platforms like YouTube and extract textual descriptions or transcripts using automated tools.
    - Next, we will fine-tune a Visual-LM model on the collected data to generate detailed textual descriptions of recipes, ingredients, cooking methods, and culinary experiences.
    - Following that, we will work on developing a tool to convert the textual descriptions generated by the Visual-LM model into structured tabular data, including relevant attributes such as ingredients, cooking time, difficulty level, etc.
    - Next, we design a schema to organize the structured recipe data, capturing essential information for effective recommendation, including ingredients, cooking instructions, dietary labels, and user ratings.
    - Lastly, we will implement an SQL-based engine that allows users to query the recipe database based on their preferences, dietary restrictions, ingredient availability, and time constraints. Users can specify their requirements through SQL queries to receive personalized recipe suggestions.
    - In order to evaluate the effectiveness of the system, we will be utilizing user testing and feedback.

Resources

- Cluster with GPU for training.
- Large datasets containing videos. There are a good number of such datasets available from sources all throughout the Internet, such as datasets containing many YouTube videos, and more specialized datasets like YouCook2 and MPII-Cooking, which specifically contain cooking videos on YouTube.
- Some kind of large Video Language Model as a base. We are considering either building on top of the existing Zelda model or starting with CLIP as a base and fine-tuning it for our task.
- Potentially, a vector database. We may want to model ingredients or other forms of unstructured data as vectors, in which case having the functionality to compare these vectors out-of-the-box could be very useful. We could start with something like pgvector since it works with Postgres, and if necessary, we may modify the vector comparison methods to better suit our purposes.

Deliverables

- A demo application that can output related cooking video clips based on user inquiry (natural language/SQL)
- A 5-6 page research report or documentation explaining design decisions and implementation, as well as

experimental results from comparing our output with different baselines
- Our own dataset that we create using both an existing dataset as well as populating that using information acquired from the video datasets.

Timeline

- Determine basic foundational ideas (ie. metrics to measure performance, concretely settle on datasets, begin setting up codebase) (Feb. 29)
- Complete dataset, setup database, begin writing code for project. (Mar. 3)
- Complete semi-functional version of project, ie allow for database population for generic videos, not necessarily cooking yet (Mar. 12)
- Complete designing proper database design and model tuning to focus around cooking as opposed to generic videos (Mar. 21)
- Begin designing and implementing basic tests to determine model performance (Mar. 28) -
Begin running tests (Apr. 4)
- Finalize project and put last base optimizations/improvements (Apr. 15)
- Begin work on paper (Apr. 24)
- Finish paper by Apr. 30 (or whenever end is)

Expected division of labor and responsibilities to individual team members.

- Muskaan will handle finding datasets, database design, and consider optimizations to improve model performance. Alex will focus on fine-tuning the model and designing tests. Vincent will design metrics and test the model against other baselines with differing hyperparameters.
- All members are expected to contribute equally to codebase and towards general design of the overall project

Consider potential risks and mitigation strategies.

- *Specific definition of many aspects are still somewhat nebulous, like how we would determine if the relevance of a particular video clip from our model is greater than or less than the output from some other model's, such as Zelda's*
- We save testing for later so we have more time to tackle these issues and design a good metric -
Model performance is dif icult to improve to relevant levels.
- We provide a decent amount of time for testing still, so that we have some more time to tune our models. -
Data is insuf icient to prevent underfitting
- We provide ourselves a decent amount of time early on to research proper datasets to use, to hopefully avoid this situation. Otherwise, we can attempt to parse more general cooking datasets to glean more data, or use data augmentation on existing videos to increase the amount of data.

References:

- CLIP (https://openai.com/research/clip)
- Helpful Paper (https://arxiv.org/abs/2109.02707)
- Zelda (https://arxiv.org/abs/2305.03785)
- Potential Dataset YouCook2 http://youcook2.eecs.umich.edu/
- Potential Dataset MPII-Cooking
https://www.mpi-inf.mpg.de/departments/computer-vision-and-machine-learning/research/human-activity-recognition/mpii-cooking-activities-dataset
- Fine Tuning a Vision Model https://www.kaggle.com/code/jhoward/the-best-vision-models-for-fine-tuning -
Video Datasets https://github.com/xiaobai1217/Awesome-Video-Datasets?tab=readme-ov-file#Pose-Estimation -
Helpful Paper https://www.mdpi.com/2076-3417/13/13/7880
- Helpful Paper https://aclanthology.org/2022.acl-long.180.pdf
- Helpful Blog https://learn.mixpeek.com/ai-video-search-engine/

Possibly Useful Links:

Video 2 Text basic, summarizes video (note: super outdated): https://github.com/scopeInfinity/Video2Description
Uses LLM to generate random dialogue about what is happening on screen:
https://github.com/xISSAx/Alpha-Co-Vision