**Team Name:** Alphabet Soup

**Project Title:** Recognizing image-recipe pairs using cross-modal learning

**Project summary (4-5+ sentences). Fill in your problem and background/motivation (why do you want to solve it? Why is it interesting?). This should provide some detail (don’t just say “I’ll be working on object detection”)**

We’re investigating how a deep learning model can take a food image, parse the underlying components, and produce a set of outputs (ingredients, recipes) useful to an end user (im2recipe); a process that can also be done in reverse (recipe2im). This is an interesting problem as it shows applications of deep learning models for practical use; the recreation of a food one saw online, a calorie counter for one on a diet, or understanding of food ingredients for allergy considerations.

Specifically, we’re interested in learning how to build these models, understanding and potentially improving on the used architectures. Our dataset [1] contains cross-modal data (both image and textual info). As such, our approach enables us to investigate both CNN and LSTM architectures - an exciting prospect as it melds multiple aspects of the course together.

**What you will do (Approach, 4-5+ sentences) - Be specific about what you will implement and what existing code you will use. Describe what you actually plan to implement or the experiments you might try, etc. Again, provide sufficient information describing exactly what you’ll do. One of the key things to note is that just downloading code and running it on a dataset is not sufficient for a description or a project! Some thorough implementation, analysis, theory, etc. have to be done for the project.**

Our primary goal is to replicate the architecture shown in [1]. The general approach used is to analyze the image (via CNN), analyze the ingredients (via LSTM) and analyze the recipe steps (via secondary LSTM) independently, then combine the resulting classes into a joint embedding space for loss function development. This approach is common in much of the literature [1, 2, 3, 4].

As a baseline investigation, we want to analyze ingredients in an image using only a subsection of the architecture found in [1] (ingredient-encoder + image-encoder). The decreased complexity of this subset architecture allows us to verify this subcomponent is working as expected before proceeding.

Our next goal will be to extend our architecture to the full joint-embedding architecture found in [1] and see how our results match up to that of the paper.

As a stretch goal, we would investigate the benefits of modifying the loss function using semantic regularization [1] or the AdaMine-based loss function presented in [3].

**Resources / Related Work & Papers (4-5+ sentences). What is the state of art for this problem? Note that it is perfectly fine for this project to implement approaches that already exist. This part should show you’ve done some research about what approaches exist.**

[1] J. Marín et al., "Recipe1M+: A Dataset for Learning Cross-Modal Embeddings for Cooking Recipes and Food Images," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 1, pp. 187-203, 1 Jan. 2021, doi: 10.1109/TPAMI.2019.2927476.

[2] J.-J. Chen, C.-W. Ngo, F.-L. Feng, and T.-S. Chua, “Deep understanding of cooking procedure for cross-modal recipe retrieval,” in Proc. 26th ACM Int. Conf. Multimedia, 2018, pp. 1020–1028. [Online]. Available: <http://doi.acm.org/10.1145/3240508.3240627>

[3] M. Carvalho, R. Cadene, D. Picard, L. Soulier, N. Thome, and M. Cord, “Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings,” in Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2018, pp. 35–44.

[4] M. Kumari and T. Singh, "Food Image to Cooking Instructions Conversion Through Compressed Embeddings Using Deep Learning," 2019 IEEE 35th International Conference on Data Engineering Workshops (ICDEW), 2019, pp. 81-84, doi: 10.1109/ICDEW.2019.00-31.

[5] H. Rawlani, J. Saita, V. Zambre and R. L. Priya, "Deep Learning based approach to suggest recipes," 2018 International Conference on Smart City and Emerging Technology (ICSCET), 2018, pp. 1-4, doi: 10.1109/ICSCET.2018.8537350.

**Datasets (Provide a link to the dataset). This is crucial! Deep learning is data-driven, so what datasets you use is crucial. One of the key things is to make sure you don’t try to create and especially annotate your own data! Otherwise, the project will be taken over by this.**

Recipe 1M+ [1]: <http://im2recipe.csail.mit.edu/>

**List your Group members.**

Justin Havas

Geoffrey Horowitz

References:

Image to recipe

- Dataset: <http://im2recipe.csail.mit.edu/>

- Corresponding paper: <https://ieeexplore.ieee.org/document/8758197>

1. CNN (ResNet-50) to classify image
2. Recipe is sent to two different LSTMs:
   1. Ingredient Bi-directional LSTM: sifts through and pulls out the pertinent ingredient(s) in each line of recipe
   2. Cooking instructions (not entirely clear on this point):
      1. Each line of instruction is sent to LSTM (only reads forward since ordering matters) (aside: can reference LSTM we make in A4)
      2. LSTM performs over skip-instruction vectors (not sure what this is)
      3. Results of LSTMs are sent to composite LSTM for entire recipe
3. Instructions, Ingredients, Image results all combined into “joint embedding” (basically just connecting the text results with the image results)
4. Semantic Regularization (fancy weight sharing during training?) - this is analyzed independently, so we can ignore it.
   1. Optimization: First fix weights of image network (already pretrained on ImageNet), learn recipe encodings. Then freeze recipe and semantic representation weights and allow image network to learn. After this initial alignment, adjust all weights at once

- Very similar: <https://ieeexplore.ieee.org/document/8099810>

- Dataset: Food 101 <https://data.vision.ee.ethz.ch/cvl/datasets_extra/food-101/>

- [corresponding paper](https://data.vision.ee.ethz.ch/cvl/datasets_extra/food-101/static/bossard_eccv14_food-101.pdf)

This is used as a benchmark for the Recipe 1M+ paper

- Adamine: <https://arxiv.org/pdf/1804.11146.pdf>

- use citation from dataset paper

- similar architecture to one proposed in Recipe1M+, seems to be state-of-the-art

- Could we consider VilBERT impl and compare?

- image branch is ResNet-50 pretrained on image net followed by FC

- ingredients use bidirectional LSTM pretrained on embeddings from word2vec

- instructions use hierarchical LSTM

- word-level pretrained usings skip-thought but not fine-tuned

- sentence-level learned from scratch

* Attention-based approach: <http://vireo.cs.cityu.edu.hk/papers/2018_p1020-chen.pdf>
  + Similar setup but instead of LSTM uses attention
  + GRU instead of LRU
  + Interestingly uses title, ingredients and instructions (others just use ingredients/instructions)

- <https://towardsdatascience.com/this-ai-is-hungry-b2a8655528be>

- <https://ieeexplore.ieee.org/document/8750911>

Matches food image to recipe/food label data

1. CNN to match image to recipe “label”
2. LSTM to match recipe label with association data (best served with, seasoned with, etc)
3. LSTM to match actual recipe steps with recipe label

- <https://ieeexplore.ieee.org/document/8537350>

Proposes recipes based on image of raw materials

1. Uses CNN (Faster-RCNN) to identify raw materials in image

Then uses look-up table to find applicable recipes that use those materials

- <https://ieeexplore.ieee.org/document/8784769>

Proposes recipes based on image of raw materials

1. Uses CNN (Resnet-101 + “Spatial Regularization”) to identify raw materials in image
2. Uses SOMETHING (NeuMF?) to predict best model using (inputs) recipes available based on raw material (lookup) and user preferences (predefined?)

- <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7169757>

- <https://link.springer.com/chapter/10.1007/978-3-319-39601-9_4#Sec7>

Uses images of food to deconstruct component parts + portion size, output caloric content of food

1. Uses CNN (modified inception modules)

- <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7837725>

Additional notes

* Recipe1M+ dataset
  + Help for problem statement: “to solve the practical and socially relevant problem of demystifying the creation of a dish that can be seen but not necessarily described”

Image to Ingredients:

https://github.com/SumithBaddam/NeuralCook/blob/master/NeuralCook.pdf