Title: Predicting Building Energy Efficiency Using Google Aerial & Streetview, OpenStreetMap and Land Surface Temperature Data
Who: Athena Deng (akdeng), Julia Gu (jgu41), Eric Zheng (ezheng37)
Introduction: What problem are you trying to solve and why?
Paper: https://www.sciencedirect.com/science/article/pii/S0306261922017998

- This paper aims to estimate building energy efficiency using publicly available data (street & aerial view images, building footprints, and satellite-derived Land Surface Temperature). For ground truth labels, the study used the EU's EPC registry data. If results are reliable, such a model could be used to identify buildings or even general geographical areas where buildings are energy-inefficient. This would be immensely useful for sustainable building initiatives and decarbonization plans—with such a tool, one could quickly and cheaply identify potential high-yield retrofits compared with traditional methods like on-site assessments. We chose this paper not only because of its potential societal benefits, but also because it would allow us to apply many concepts we've covered in class. The paper combines many different modalities of data as well as multiple types of machine learning models, giving us the opportunity to work with all of them.
- This study is a binary classification problem: We are classifying buildings as energy efficient (A-D rating) or energy inefficient (E-G).

Related Work: Are you aware of any, or is there any prior work that you drew on to do your project?

- Please read and briefly summarize (no more than one paragraph) at least one paper/article/blog relevant to your topic beyond the paper you are re-implementing/novel idea you are researching.
    - Understanding building energy efficiency with administrative and emerging urban big data by deep learning in Glasgow
    - In this paper, the authors used a CNN trained on the ImageNet dataset to parse images and convert them into a usable vector of features. They combined the image vector with another one generated from the EPC data to create a collective vector, then fed the collective vector into a 4-layer deep neural network to output a non-binary classification. The data used in this study included Energy Performance Certificate (EPC) data containing information about each buildings' features, the UKBuildings dataset from EDINA Geomni Digimap Service which provided building footprints, and Google Street View (GSV) facade images taken from Scotland, UK. Their model achieved a final performance of 86.8% accuracy on the test set, which is pretty good considering they were determining accuracy not based on a binary classification but rather on whether they exactly matched the classification
- Public implementations:
    - None so far—this is a relatively new research paper

Data: What data are you using (if any)?

- The data used in this study are as follows:
    - Images: Google Street View (GSV) + Aerial View (GAV)
    - Coordinates: OpenStreetMap building footprints
    - Temperature: Landsat-8 Land Source Temperature (LST)
    - EPC ratings: EU Energy Performance Certificate (EPC) registry
- Though the coordinate, temperature, and EPC ratings have already been compiled in a dataset by the authors of the paper, we will have to download and clean the GSV and GAV data separately. This consists of about 80,000 images split between the two sets, representing around 40,000 buildings.
- In order to preprocess these images, we will have to use an embedder to encode them, then use k-means clustering to clean the data (eliminating unhelpful images like ones of just walls). Then, we can concatenate the embeddings to the rest of the compiled data for their corresponding buildings. After final concatenation, this will be our completed dataset.

Methodology: What is the architecture of your model?

- Model architecture: detailed in the paper
- How are you training the model?
    - We are splitting our data into three sets—train, test, and val—with sizes of approximately 23,000, 3,500, and 3,500 respectively. After we've processed the images with the pre-trained InceptionV3 model and added the LST and building area data, we will have our final feature vectors. We will store these feature vectors in a numpy array, and the labels in an array with matching indexes. Then, we will use the training data on an binary classification MLP, which will output the final prediction.
- If you are implementing an existing paper, detail what you think will be the hardest part about implementing the model here.
    - The most difficult part will be (and has been) processing the data—we've experienced significant slowdowns we weren't expecting due to the amount of time it took to first gather all the data, clean it, and rewrite and compile it in the form of numpy arrays (probably over 30 hours total). After getting the data done, though, the model should be pretty simple to create, since it's just a pretty standard MLP.

Metrics: What constitutes "success?"

- What experiments do you plan to run?
    - We will first calculate the precision, recall, and F1 scores by running the model on the test dataset. Then, if time allows, we will try to run the ablation study as well to see if we can replicate the amount each feature contributes to the final scoring of the model.

- For most of our assignments, we have looked at the accuracy of the model. Does the notion of "accuracy" apply for your project, or is some other metric more appropriate?
    - Here, accuracy will not be a good metric because the data is pretty unbalanced. Therefore, we will instead use precision, recall and F1 scores.
- If you are implementing an existing project, detail what the authors of that paper were hoping to find and how they quantified the results of their model.
    - The authors of the paper were hoping to find that the LST data contributes positively to the overall scores of the model. They did find this by quantifying their results with an ablation study.
- What are your base, target, and stretch goals?
    - Base: finish training the MLP so that everything works (i.e. can run and train smoothly without errors), but no accuracy threshold
    - Target: get within ~5% of the authors' accuracy (i.e. minimum 63% accuracy)
    - Stretch: try running some of the comparison tests with the baseline models the authors used, or conducting the ablation tests

Ethics: Choose 2 of the following bullet points to discuss; not all questions will be relevant to all projects so try to pick questions where there's interesting engagement with your project. (Remember that there's not necessarily an ethical/unethical binary; rather, we want to encourage you to think critically about your problem setup.)

- What broader societal issues are relevant to your chosen problem space?
    - This problem space lies mainly at the intersection of the climate and housing crises—mainly the topics of decarbonization, sustainable development, and retrofitting. In Rhode Island alone, residential energy consumption (e.g. building heating & cooling) makes up more than a third of the state's total energy consumption. However, it would be insensible to just try and replace every building with one we know to be energy efficient, both from an environment and housing crisis perspective—it would be a massive waste of both existing resources and existing housing. Instead, we should find ways to reduce this energy usage while still retaining old buildings, for example by replacing energy-inefficient systems with efficient ones, or retrofitting buildings to have better insulation and more energy efficient design. This leads us to our problem—how do we identify these buildings? The process of acquiring the data that will determine which buildings to prioritize is both costly and lengthy. Licensed professionals must go on-site and conduct in-depth analyses to determine each building's energy efficiency, a procedure not many building owners would willingly go through without some sort of incentive. This is the issue this paper attempts to solve, and what we are hoping to replicate.
- Who are the major "stakeholders" in this problem, and what are the consequences of mistakes made by your algorithm?
    - The major "stakeholders" in this problem would be 1) organizations or 2) homeowners that could potentially use such a model to identify buildings suitable for retrofits or renovations.

    - The organizations usually are ones promoting sustainable development or energy reduction programs, and while in theory could be both public and private, in practice are usually either nonprofits or branches of government with limited resources that are trying to incentivize homeowners to retrofit their properties. Mistakes made by our algorithm could result in inaccurate predictions, wasting the few (and diminishing) resources they have by telling them to target homes that are energy efficient, and missing the ones that aren't.
    - On the homeowner side, as the beneficiaries of these types of programs, their home usually has to qualify in order to receive financial assistance or subsidies. If they initially believe their home does qualify after identifying it as non-energy efficient using this algorithm, but it turns out that after an audit it actually is energy efficient, they will have wasted their time and money as well.

Division of labor: Briefly outline who will be responsible for which part(s) of the project.

- We will be dividing everything evenly for all parts (aka. people just do whatever is necessary whenever they have time). This has been working well for us so far and we don't foresee any issues occurring with this method.