What makes Amherst look like Amherst

Anonymous CVPR submission

Paper ID *****

Abstract

Every geographical region is unique in its own way as it contains an architectural pattern that makes it unique and differentiates it from others, and finding those unique features to identify that region is tremendously a difficult task to work on. This paper will try to unravel how a machine learning mode identifies a random image and correctly predicts whether it is from one region or another. We have collected the image data set of a geographical region (in this paper it is Amherst city) manually and have applied Convolution Neural Networks (CNN) architecture like (Visual Geometry Group) VGG16, and (Residual Neural Networks) ResNet50 to train the dataset efficiently and visualize important regions in images for CNN-based models through the (Gradient-weighted class activation mapping) Grad-CAM technique to analyze what about the images encourages them to be considered "Amherst-like".

1. Introduction

This study looks closely at Amherst, a town in Massachusetts, to understand what makes it visually unique. Amherst is an older New England town where old and new buildings mix and the colorful plants of New England blend with brick walls. We got inspired by a famous study about "What makes Paris...Paris"[1] to try to identify which architectural choices make Amherst unique.

We have been using an image data set of Amherst city (collected by us) and five more cities and then predicting it on the test dataset, it would be difficult to predict which image is from which city as both would contain streets, buildings, traffic signals, etc. We have been making and training different neural networks, and modern transfer learning techniques on our data set (which contains images of both Amherst and different cities) such that it predicts which city is which with as good accuracy as we can get by searching the best hyper-parameter during training (each image in our dataset is labeled with its city name). This will be compared with some pre-trained models fine-tuned on our data set. Then, with the best model, we used Grad-CAM on images of Amherst to see what draws the model's attention and focus when labeling an Amherst image.

Figure 1. This image is of Amherst's famous building i.e. UMass W.E.B. Du Bois Library which is unique to Amherst.

2. Literature Review

Our exploration into decoding the visual identity of Amherst draws upon a comprehensive review of literature, encompassing works in urban scene recognition, advancements in Convolutional Neural Network (CNN) architectures, transfer learning techniques, and visualization tools like Grad-CAM.

Figure 2. This image is of a house in Amherst that showcases an architectural pattern including windows, and a main entrance which are unique to a specific city and will help to predict that this image belongs to Amherst.

2.1. Foundational Works

The source of inspiration for this study was provided by Doersch et al., titled "What makes Paris look like Paris?"[1] The study delves into the inherent features that contribute to a city's visual identity, Paris in this case, using different machine learning techniques and attention maps. Based on this we decided to utilize grad-cam attention maps [4] to extract the most important features of our images as the model sees them.

2.2. CNN Architectures

To extract as high-quality features as possible from our images we decided to investigate CNNs in our literature review [3]. Based on existing results with CNNs we know they have a tendency to greatly outperform more standard neural networks[3]. Therefore, inspired by this research we decided to use TensorFlow to construct a custom CNN to train on our unique dataset and compare this result to existing image classifying CNNs like resnet50 and VGG16.

2.3. Transfer Learning: Leveraging Pre-Trained Models

Inspired by the principles outlined in "What makes Paris look like Paris?,"[1] and the success of past researchers [9] we decided to adopt transfer learning techniques to enhance our models' performance. Specifically, we utilized two pre-trained CNN architectures, VGG16 and ResNet50, to compare with the results from our custom cnn. These models, initially trained on large datasets, bring a wealth of knowledge that is transferred and fine-tuned for the specific task of decoding Amherst's visual identity.

2.4. Grad-CAM and Confusion Matrix: Visualizing and Evaluating Models

The literature review extends to visualization techniques, particularly Grad-CAM (Gradient-weighted Class Activation Mapping) and confusion matrices. Grad-CAM enables us to visualize the important features within Amherst's images, providing insights into the areas of focus for our models. [4] This visualization tool aids in understanding the decision-making process of the CNN architecture. On the evaluation front, confusion matrices become a valuable tool for assessing the performance of our models. These matrices offer a detailed breakdown of classification results, shedding light on areas of potential improvement and guiding the refinement of our approach, so we can analyze the "thinking" of as a smart model as possible [8].

3. Dataset

3.1. Collection and Augmenting Diversity

Our dataset acts as a gateway to the unique visual world of Amherst, carefully composed of around 11,000 images. It was created by a combination of photography and frame extraction from videos of Amherst. However, we also needed images from other cities to train our models one. Flickr has become the preferred source of data for many applications in computer vision and graphics, especially those involving visual geo-location. [7] However, a challenge with Flickr and similar consumer photo-sharing websites in geographical tasks is the noticeable bias toward famous landmarks. To address this bias and ensure a more balanced representation of geographical locations, we rely on a dataset constructed using Google Street View [6]. This extensive database contains street-level images captured as panoramas using specially designed vehicles. Utilizing this resource allows us to extract views of building facades that are roughly front-parallel, minimizing the impact of large variations in camera viewpoints. To broaden our understanding, we extend the scope of our dataset by including an additional 50,000 street view images sourced from five major cities—New York, Washington, Detroit, Chicago, and San Francisco. The hope here is that by having a diverse pool of images from different cities we will be able to train a model to find things truly unique to Amherst. One downside of course to using this dataset is that while we will have a diverse number of perspectives of the city from the road we will not have perspectives of the cities from any other vantage point; say from a sidewalk, top of a building, or from someone's window view. Inspired by this shortcoming we decided to record some video from inside the car while driving around Amherst to try and get photos from Amherst that would be from a similar perspective as the outside dataset. That said, we still captured photos walking around Amherst and so in that sense, some of our Amherst images were captured from

perspectives not existent in our dataset of the other 5 cities.

3.2. Preprocessing

Each of the images in our dataset underwent preprocessing. Where each of those 5 cities' dataset were labeled according to their coordinates (longitude and latitudes). The geographical coordinates associated with each image are utilized to extract city labels, a critical step in organizing the dataset effectively. A custom Python script, leveraging the geolocator module, assists in this process, allowing for the seamless extraction of city names from the coordinates. Later, all images are resized to standardized 128x128 pixels to ensure each visual element receives equitable representation.

4. Methodology

Our technical approach involves a multi-faceted strategy employing diverse Convolutional Neural Network (CNN) architectures, transfer learning, and advanced visualization tools to decipher the unique visual identity of Amherst.

4.1. Custom CNN Architecture

We designed a custom CNN architecture specifically tailored for the task at hand. This sequential model consists of three convolutional layers, each followed by a max-pooling layer for spatial downsampling. The convolutional layers have increasing filter sizes, enabling the network to capture features of varying complexities. Subsequently, the model incorporates a flattened layer, enhancing feature extraction, and two dense layers for classification. The use of the softmax activation function in the output layer facilitates the categorization of images into predefined classes. The model is optimized using the Adam optimizer, and categorical cross-entropy is employed as the loss function during training.

4.2. Transfer Learning with VGG16

The adoption of the VGG16 architecture [5] in our project brings with it a wealth of pre-trained knowledge from ImageNet, a large-scale dataset encompassing a myriad of visual concepts. VGG16 is characterized by its deep structure, comprising 16 layers with a uniform kernel size of 3x3 and max-pooling layers for downsampling. By leveraging the convolutional layers of VGG16, we tap into its ability to capture hierarchical features, ranging from simple textures to complex patterns. These features, learned from a diverse array of images during pre-training, serve as a foundation for our model to discern and extract relevant visual elements specific to Amherst.

The transfer learning process involves freezing the convolutional layers, preventing them from updating during training. This decision is made to preserve the generalizable features learned by VGG16 from ImageNet. We extend the architecture by appending a custom classifier tailored to our Amherst dataset. The classifier includes a flattened layer followed by dense layers for feature extraction and classification. Fine-tuning the model with the Adam optimizer and categorical crossentropy loss ensures its adaptability to the nuanced visual characteristics of Amherst.

4.3. Transfer Learning with ResNet50

ResNet50 [2], renowned for its residual learning framework, is another cornerstone of our transfer learning strategy. The architecture's distinctive feature is the introduction of residual connections, mitigating the vanishing gradient problem and enabling the training of deeper networks. This is particularly advantageous when dealing with intricate visual patterns in Amherst's dataset.

Similar to VGG16, we adopt a transfer learning approach with ResNet50 by utilizing its pre-trained convolutional layers from ImageNet. By freezing these layers, we retain the knowledge acquired by ResNet50 in recognizing complex hierarchical features. A custom classifier is integrated, comprising a flattened layer and dense layers for feature extraction and classification specific to Amherst. The resulting model is fine-tuned using the Adam optimizer and categorical cross-entropy loss, ensuring a seamless integration of ResNet50's capabilities with the unique visual nuances of Amherst.

4.4. Grad-CAM for Interpretability

To interpret and visualize the decision-making process of our models, we implemented Grad-CAM (Gradient-weighted Class Activation Mapping) [4]. Grad-CAM highlights the regions in input images that contribute most significantly to the model's predictions. This visualization technique aids in understanding the focal points of the CNN architectures when distinguishing visual elements specific to Amherst.

4.5. Confusion Matrix Evaluation

For a comprehensive evaluation of our models, we employ confusion matrices [8]. These matrices provide a detailed breakdown of the classification results, indicating true positives, false positives, true negatives, and false negatives for each class. Metrics such as accuracy, precision, recall, and F1-score are derived from the confusion matrix, guiding model refinement. The insights gained through confusion matrices contribute to the iterative improvement of our models, ensuring their effectiveness in decoding the intricate visual language of Amherst. While not the main focus of our investigation, it is still important our model is as smart as possible so we can be reasonably certain its actually learning things about Amherst and not merely guessing.

Figure 3. Sample image of each city after pre-processing dataset

Figure 4. Image Dataset of Amherst City highlighting houses, buildings, and greeneries that define Amherst as Amherst

5. Experiments, Results & Discussion

To recap, we took around 11000 images of Amherst and combined them with existing street images of 5 other cities around 10000 images of each city to create our final dataset. We sent this dataset through our 3 models (custom CNN, pre-trained VGG16, and pre-trained ResNet50). Then, we split our dataset in an 80:20 ratio (80% Training set and 20% Test Set) for better and more effective training and testing. We split the training dataset again in an 85:15 ratio (85% actual training set and 15% validation set) as well. We found the ResNet50 and our custom CNN consistently produced the highest accuracy over the VGG16 (that gave around 65% test accuracy) and we produced a confusion matrix for both ResNet50 8b and the custom CNN 8a. Deciding to use both models for our analysis we trained the custom CNN to a test accuracy of approximately 74% and 75% test accuracy on ResNet50. We also analyzed the confusion matrices for both models to ensure they learned stuff about Amherst and not just the other cities. What we found was actually that it learned Amherst the best consistently getting the truest positives for Amherst across both models 8a 8b. After this point, we sent several images of Amherst that were correctly labeled to identify "what makes Amherst...Amherst" using our Grad-CAM attention mapping techniques. The results gave us a heat map of areas that tend to trigger significant activations colored in red and areas that trigger relatively insignificant activations more blue. The heat maps tended to highlight any photos of vegetation, trees in particular, as the parts of the image that contributed most to high activation regions. The results, often in images with little vegetation, tended to highlight the buildings as image regions that led to high activations in our model.5 6

Our original hypothesis was "If an image has the Amherst insignia, frequent reddish and brickbased architecture, and photos of UMass or Amherst College then it will be labeled as Amherst. This is because we determine these to be distinct qualities of the Amherst town area." We didn't notice the model really identifying Amherst insignia in any of the images and it seemed to treat photos of UMass and Amherst College much like the rest of the town. That said our hypothesis was correct in that the brick-based buildings were an "Amhersty" feature. Looking at the photos of Du Bois several layers produced by the grad-cam attention maps suggested buildings like Du Bois were important to the proper labeling of the image as Amherst 6 8b. Second, our hypothesis completely missed the importance of vegetation in the images as something "Amhersty". In hindsight we can see how this would be unique to Amherst for a couple reasons. Number one, we took the photos in fall when the leaves were changing colors which is something that doesn't even happen in some of the other cities in our dataset (e.g. San Francisco). Number two, the other cities in our dataset are true metropolitan cities with a far greater population density year round than Amherst. This makes vegetation far less common in those cities which is why the prevalence of vegetation is in some ways unique to Amherst in our data. This is also the best reason we could come up for why our models consistently found the most true positives for Amherst compared with the other cities in our confusion matrices. Since vegetation tends to be generally unique to photos of Amherst its a "simple" heuristic the model could learn to differentiate Amherst from the other cities. Something to explore in the future would be

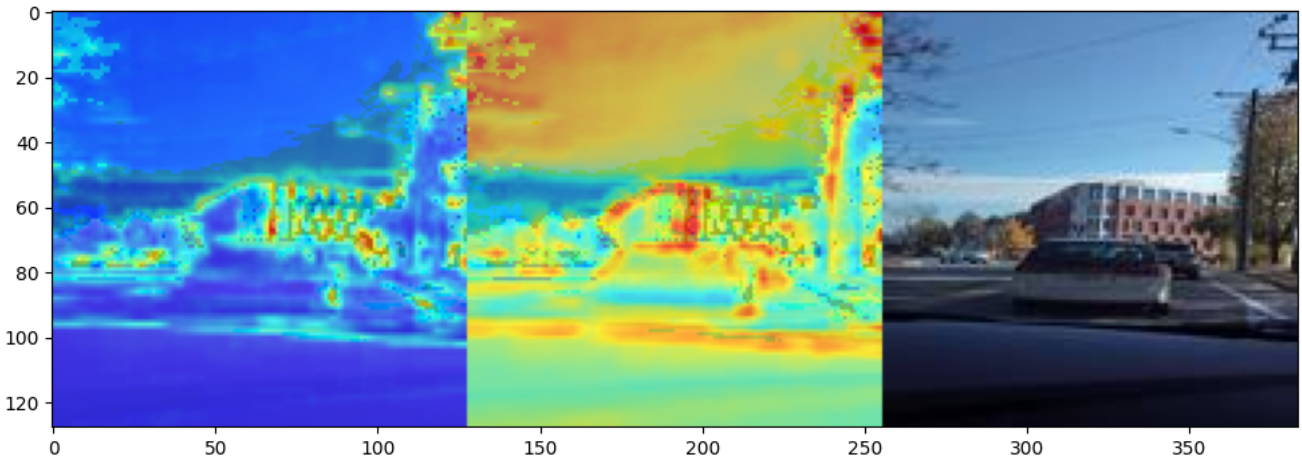


Figure 5. Amherst Grad-Cam/Attention Map

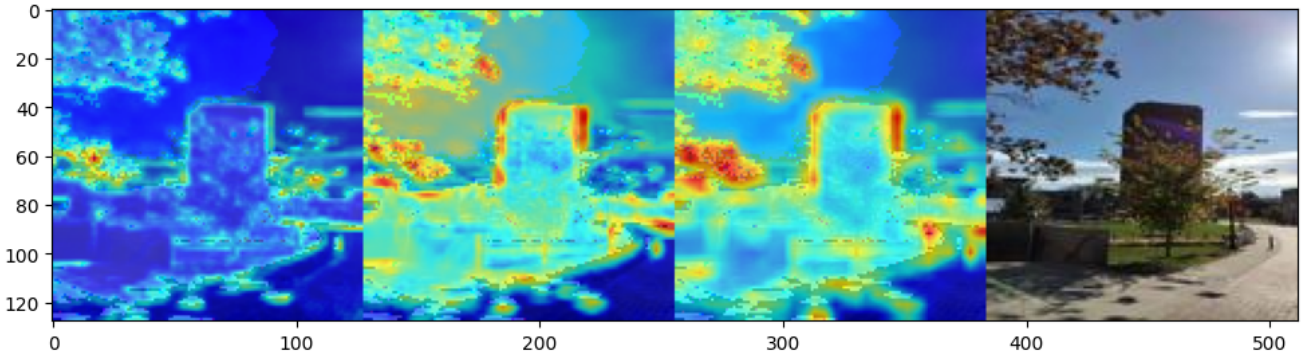


Figure 6. Amherst Grad-Cam/Attention Map

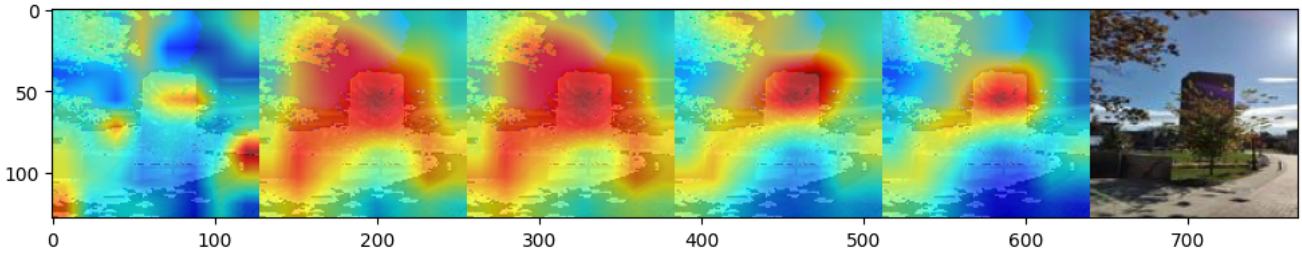
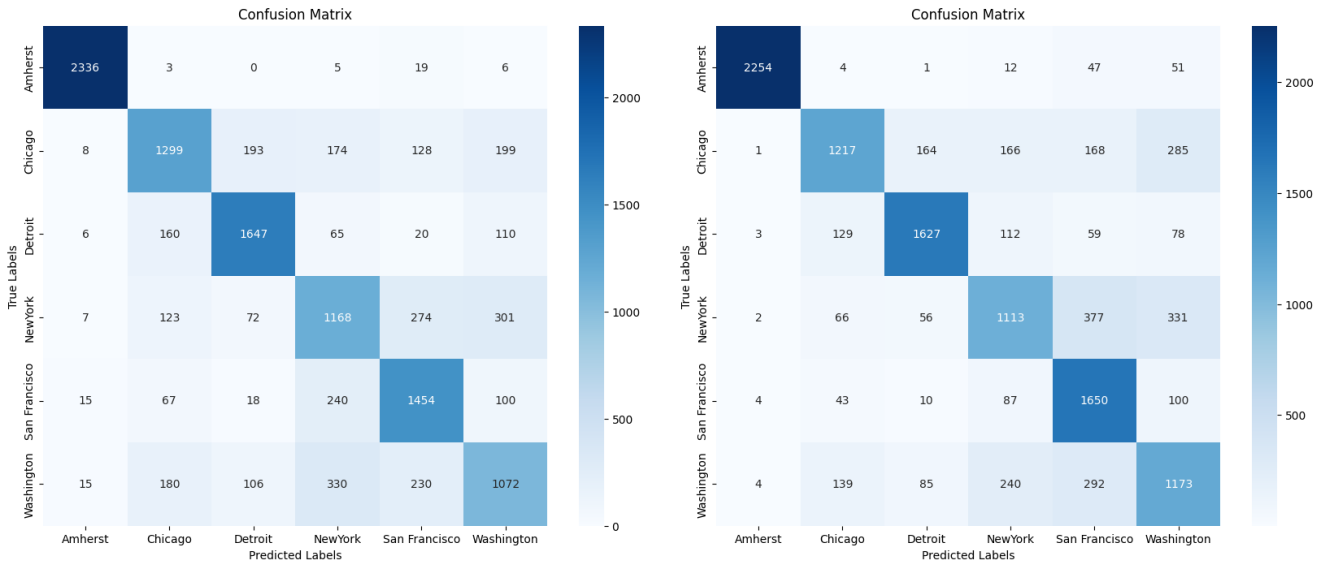


Figure 7. Amherst Grad-Cam/Attention Map

to include image data from cities with more vegetation to identify if some type of vegetation is actually unique to Amherst.

6. Conclusion & Future works

Confronted with limited resources like constrained GPU access and the manual effort needed for dataset collection in a small time phase, our project pragmatically adopted less dense CNN architectures. Despite their simplicity, these models exhibited promising capabilities in capturing distinctive visual elements, showcasing adaptability to resource constraints.



(a) Confusion Matrix on Custom CNN

(b) Resnet Confusion Matrix

Figure 8. Confusion Matrices

Looking ahead, future works could explore several avenues for improvement. Techniques like dataset augmentation may enhance model generalization. While investigating alternative models may also provide better generalization. A few strategies to find alternative models would be to try out other pre-trained models or play around more with custom CNN architectures for this specific type of problem. Automating data collection processes and incorporating temporal dynamics from video frames with things like google street view and google earth may offer potential avenues for a more comprehensive understanding of Amherst's evolving visual landscape.

An area this study in particular could improve on would be choosing a dataset that included images from different perspectives other than the street. As noted, choosing only street view photos is convienent for creating a dataset but a more thorough investigation would require the photos of other cities include different perspectives. This is because it avoids the issue where all other cities are identified by some camera perspective feature instead of actual image features.

Additionally, just using more cities to populate our dataset of images would be valuable so as to avoid issues where the city in question (Amherst here) is identified uniquely by one trait not because that trait is unique to Amherst but because its unique to photos of Amherst in the dataset. For example, here we found vegetation was really only common in our Amherst photos and none of the other photos of other cities which made it an easy heuristic for the model to learn to label Amherst. While in reality vegetation is everywhere and that's not something truly unique to Amherst. Thus adding images of some city (say Seattle) with a lot more areas of vegetation would ideally reduce the usefulness of this false heuristic and force the model to learn things more unique to Amherst in order to be accurate.

Despite these areas to grow into, we were able to show that you can learn what makes cities unique by analyzing images of them with machine learning techniques in comparison to other cities. In particular, we showed attention maps can be extremely helpful towards understanding what parts of an image activate a network and tell it to label an image as one label or another and applying this to real-world cities.

References

[1] Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei A. Efros. What makes paris look like paris? *ACM Trans. Graph.*, 31(4), 2012. 1, 2

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 3

[3] Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks, 2015. 2

[4] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, 2019. 2, 3

[5] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. 3

[6] Stelath. City street view dataset, 2022. 2

[7] Viriya Taecharungroj and Boonyanit Mathayomchan. The big picture of cities: Analysing flickr photos of 222 cities worldwide. *Cities*, 102:102741, 2020. 2

[8] Sofia Visa, Brian Ramsay, Anca Ralescu, and Esther Knaap.

Confusion matrix-based feature selection. pages 120–127, 2011. 2, 3

[9] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning, 2020. 2