Real Estate Listings with Image Classification Neural Networks

Group 2: Xiang Liu and Ashley Mercado

WandB Links: https://api.wandb.ai/links/jennisleepin/u9ae82x9 and https://api.wandb.ai/links/jennisleepin/h8jzc3ue

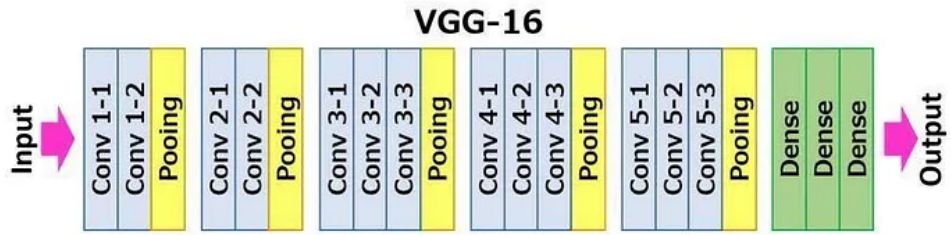
Problem Overview: The digital era has transformed the real estate market through online property listings, making searches more accessible but also more challenging due to the sheer volume of options. Our project uses advanced machine learning, specifically neural networks, to develop a classification model for indoor room images. This model categorizes property images based on room type, size, and style, improving the user experience by quickly identifying listings that match preferences, reducing time spent on manual searches.

Data Overview: Our project utilizes a subset of the MIT Indoor Scene Recognition dataset from HuggingFace, and we focus on six key categories of indoor spaces that are most relevant to real estate listings: 'bedroom', 'bathroom', 'dining room', 'living room', 'kitchen', and 'garage'. The dataset comprises 2676 images split into training (80%), validation (10%), and testing (10%) sets.

Examples from each category of our indoor-room dataset

Methodologies:

- **Support Vector Machine (SVM)** - The SVM is a traditional machine learning model well-suited for classification tasks. It works by finding the hyperplane that best divides a dataset into classes with a maximum margin. For image classification, we transformed the image data into a feature vector that SVM could process. This approach, while more simplistic compared to neural network methods, provides a strong baseline for performance due to its effectiveness in handling high-dimensional spaces.
- **Convolutional Neural Network (CNN)** - We used CNNs, optimized for image processing, to automatically detect features in our real estate dataset. Our custom CNN had three convolutional layers followed by pooling for feature reduction and fully connected layers for predictions. We also adapted VGG-16, a well-known image classification model, by adjusting its layers for our dataset's needs. See the diagram of our adapted VGG-16 structure below.



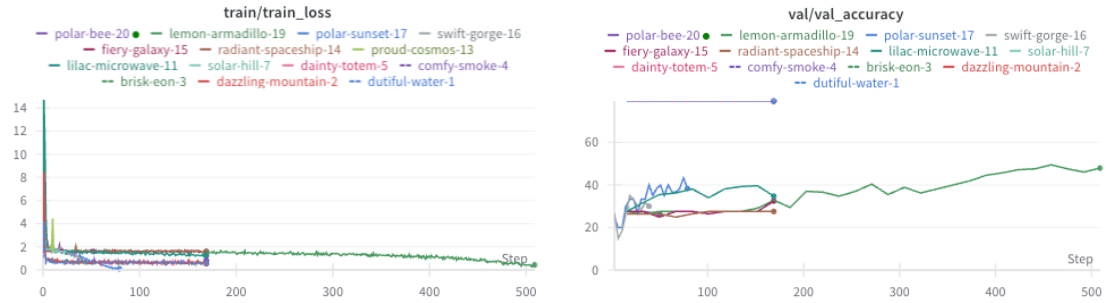
VGG-16 Structure Map [cited from Everything you need to know about VGG16]

- **Fine-tuned Vision Transformer (ViT)** - Vision Transformers represent a newer approach in the field of deep learning for images, utilizing mechanisms traditionally used in natural language processing. Unlike CNNs, ViT processes an image as a sequence of patches and uses self-attention mechanisms to weigh the importance of different patches relative to each other. We fine-tuned a pre-trained ViT model available from the Hugging Face Model Hub. This method is advantageous as it leverages a model pre-trained on a vast amount of data, bringing a deeper understanding of image features that can significantly enhance performance on our specific task.

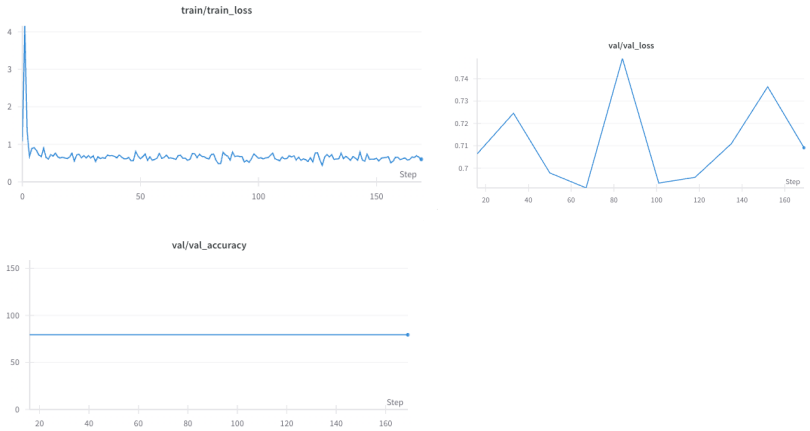
Results: The SVM baseline had the lowest accuracy at 0.2686, struggling with high-dimensional image data. Our custom CNN model improved accuracy to 0.3517 by capturing spatial hierarchies with convolutional layers. The adapted VGG16 model reached 0.3956 with its deep architecture. However, the standout was the fine-tuned Vision Transformer (ViT) at 0.8424 accuracy, showcasing ViT's attention mechanisms and pre-trained knowledge. ViT significantly outperformed traditional methods, proving its value for complex image recognition in real-world datasets.

Model	Accuracy
SVM (baseline)	0.2686
Customed CNN	0.3517
VGG16	0.3956
Finetuned ViT	0.8424

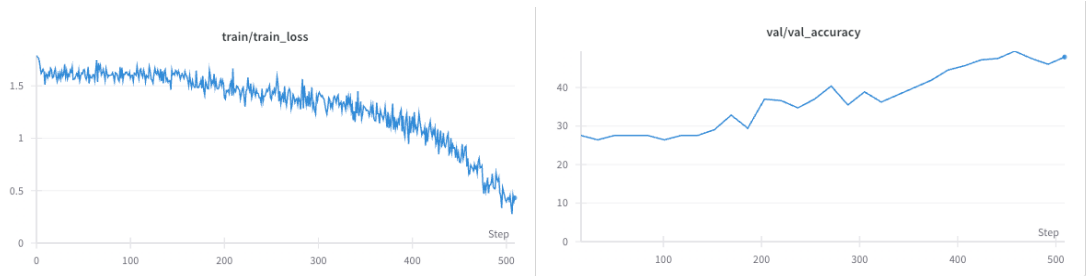
Hyperparameter Tuning: We optimized our CNN and VGG-16 models through extensive testing of learning rates and training durations. Initially, validation accuracy fluctuated significantly, suggesting suboptimal settings. Increasing training epochs stabilized improvements in validation accuracy, highlighting the importance of tuning epochs and learning rates. Our best CNN model, similar to VGG-16, was trained for 40 epochs on a GPU, emphasizing the need for careful parameter tuning in complex image classification tasks.



Testing learning rates from 0.1 to 0.0001 had a big impact. A higher rate like 0.1 caused the model to get stuck at local optima, leading to unstable validation metrics and no improvement in accuracy. Graphs illustrate the issues with high learning rates.



Testing with a very low learning rate improved validation accuracy but caused slow convergence of training loss, leading to inefficient computation. To balance efficiency and performance, we chose a learning rate of 0.001, optimizing model training for practical applications.



Conclusion: We developed and optimized various image classification models, with the fine-tuned Vision Transformer (ViT) achieving the highest accuracy, surpassing traditional SVMs and CNNs. This advanced model, with optimized hyperparameters, can revolutionize online real estate listing navigation by automatically categorizing property images, streamlining the search process and enhancing user satisfaction. Our work underscores the transformative impact of modern neural networks in real-world applications, providing a more intuitive and efficient experience for online real estate platforms.

Contribution: Xiang Liu - 50% and Ashley Mercado 50%