Hotel Classification to Battle Human Trafficking

Or Meiri 315920462
Eran Aizikovich........................ 316531201
Tomer Meshulam.....................207125196
Shemi Peretz........................... 208378042
Noam Munz 207042292

INTRODUCTION

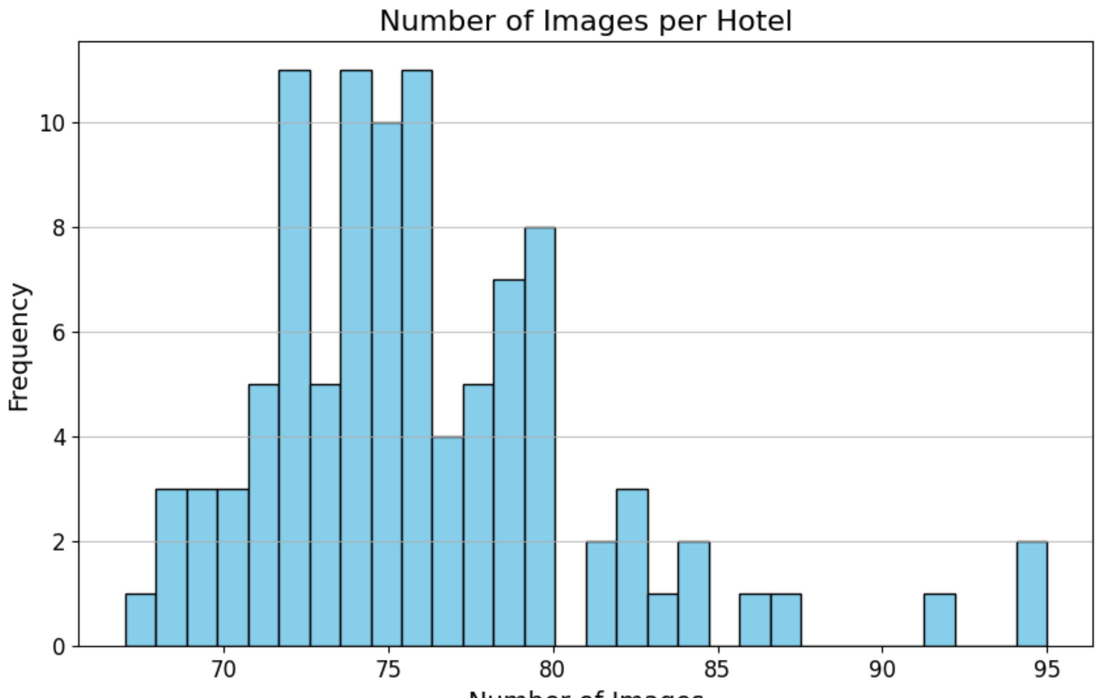
Our project is participating in the Hotel-ID to Combat Human Trafficking competition (2021), focusing on identifying hotels from images to support human trafficking investigations. This initiative collects photos of hotel rooms from everyday travelers through the TraffickCam mobile application. The goal is to build a comprehensive database of hotel room images with known hotel IDs, which can then be used to train image recognition models. These models aid investigators in identifying locations where victims of trafficking have been photographed. Our project aims to develop a model that can accurately classify hotel room images into their respective hotel IDs, using the TraffickCam dataset as a reference.

DATA EXPLORATION

The dataset consists of 91 hotel chains and 7,770 different hotels, totaling 97,554 images. Of these, 77,295 images are from hotel chains, while 20,259 images are from non-chain hotels. Due to computational limitations, we modified the dataset by excluding non-chain hotels and focusing on the 100 hotels with the highest number of images. This subset includes 21 different chains, 100 different hotels, and 7,605 images. In this dataset, each sample is a hotel image, and each label is a unique number identifying a specific hotel location.

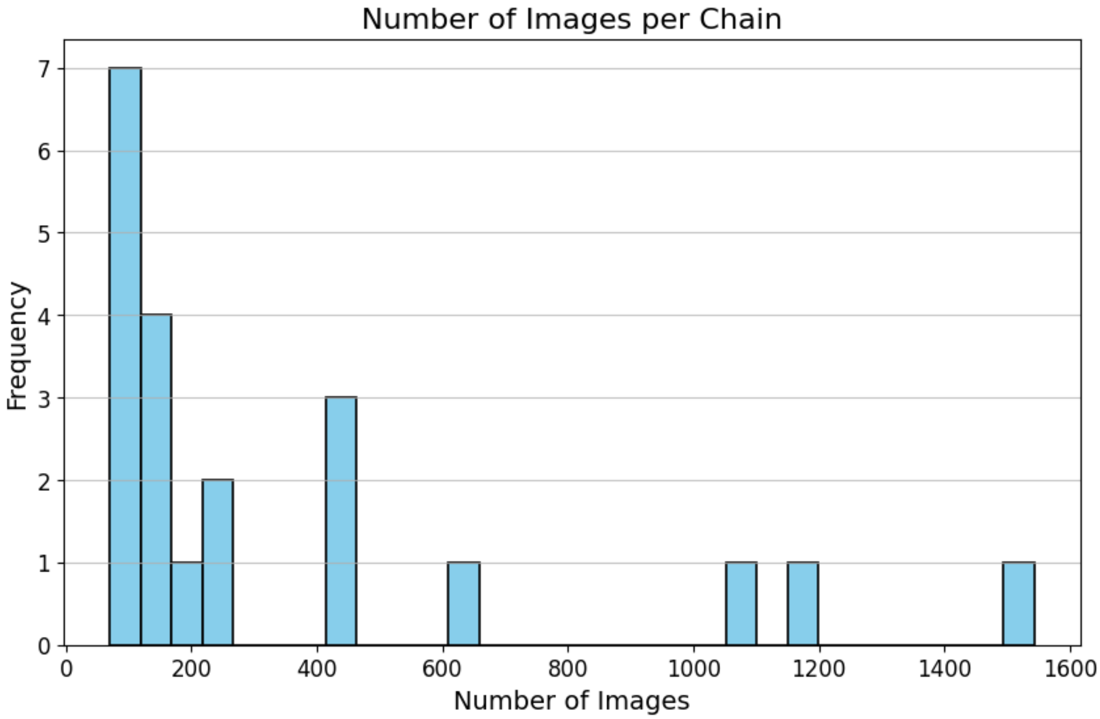
DISTRIBUTION OF IMAGES PER HOTEL

The histogram shows the distribution of the number of images per hotel in the dataset, with most hotels having between 70 and 80 images, and a few outliers having up to 95 images.



DISTRIBUTION OF IMAGES PER CHAIN

Most hotel chains have fewer than 200 images, with a few chains having significantly more, up to around 1600 images.



DISTRIBUTION OF IMAGES BY YEAR

Image collection peaked in 2017, with significant contributions in 2016 and 2018 as well. The number of images decreased from 2019 onwards, with the least in 2020, right before the competition was held in 2021.

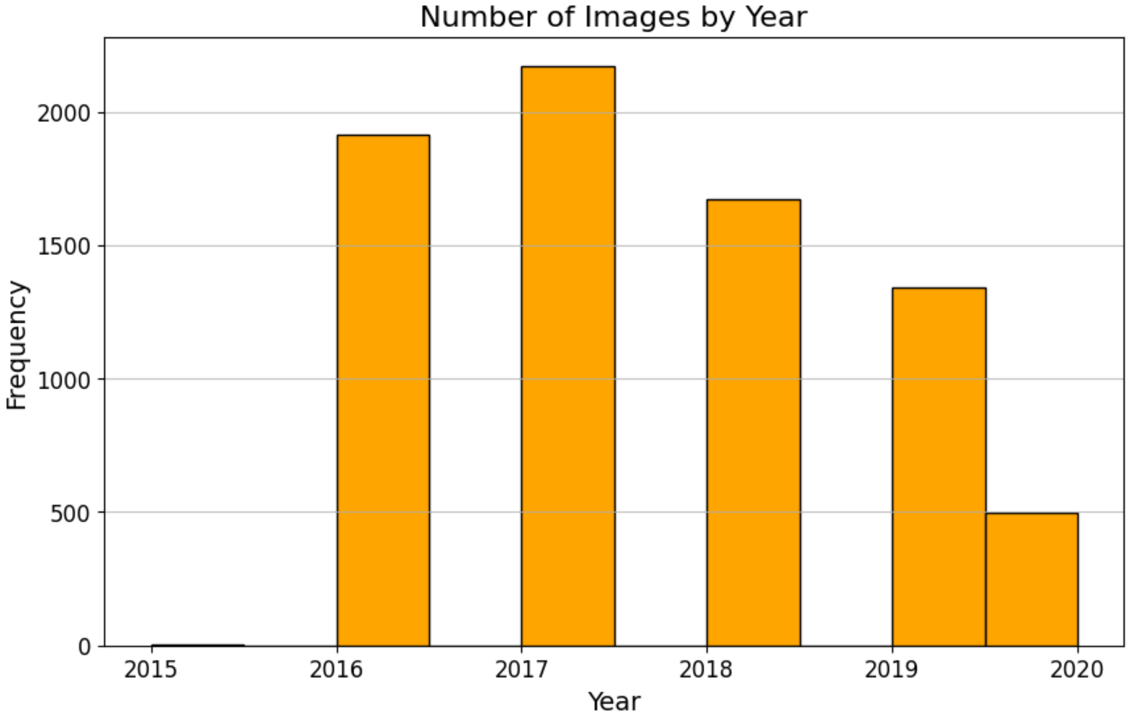


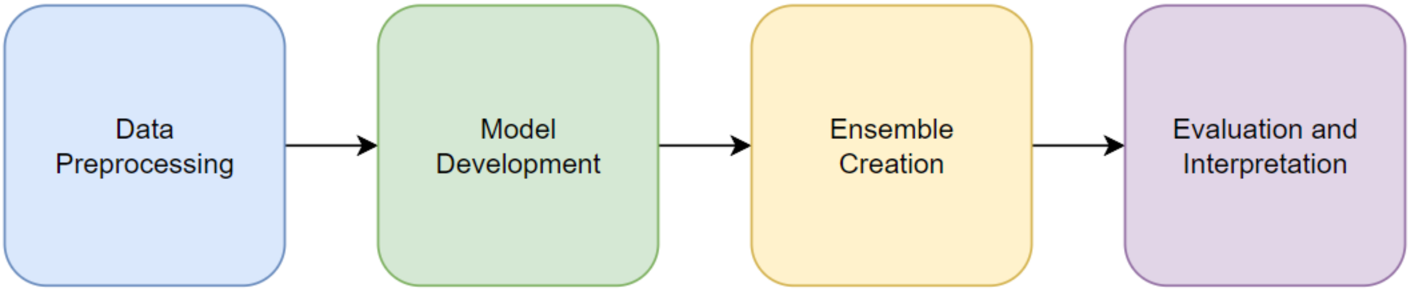
IMAGE QUALITY AND CHARACTERISTICS

The images vary in resolution and quality. Most images are taken with mobile devices, resulting in a range of image resolutions. Examples of typical images include clear, well-lit photos, while outliers may include blurry or dark images.

PIPLINE OVERVIEW

Our hotel classification pipeline consists of five key stages:

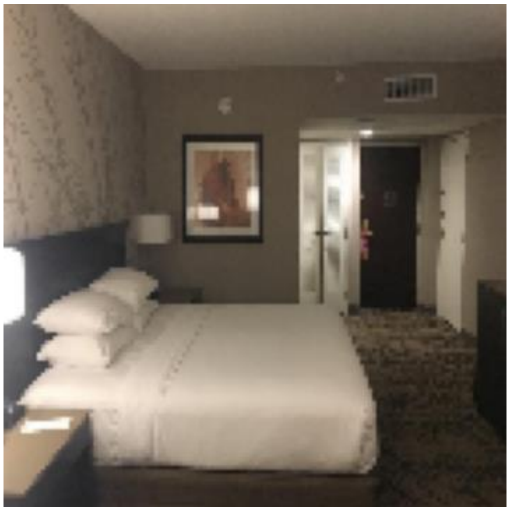
1. Data Preprocessing: Images are resized, normalized, and split into training, validation, and test sets. Data augmentation techniques are applied to increase dataset diversity.
2. Model Development: Multiple architectures (VGG16, ResNet101, Vision Transformer, and ArcFace) are trained and evaluated using various hyperparameters.
3. Ensemble Creation: Top-performing models are combined to create an ensemble model for improved accuracy.
4. Evaluation and Interpretation: Saliency and attention maps are generated to understand the models' decision-making processes.



This pipeline allows us to efficiently process hotel room images, develop accurate classification models, and gain insights into how these models identify unique hotel characteristics.

PREPROCESSING

- Resizing Images
  To ensure uniformity, reduce computational load, and meet memory requirements, all images are resized to a fixed dimension of 512x512 pixels. Resizing standardizes the input data, making it easier and more efficient for the model to process.



- Normalization

  Normalization scales pixel values to a range of [0, 1] by dividing by 255. It then standardizes the color channels using the means [0.485, 0.456, 0.406] and standard deviations [0.229, 0.224, 0.225], which are derived from the ImageNet dataset. This stabilizes the learning process and speeds up convergence by ensuring the input data has a consistent scale and distribution.

- Splitting the Dataset

  The dataset is now split into training, validation, and test sets with a ratio of 70:15:15. This approach ensures the model is trained on the majority of the data while reserving a separate validation set for hyperparameter tuning and model selection. The test set, kept independent from training and validation, will be used exclusively for final performance evaluation.

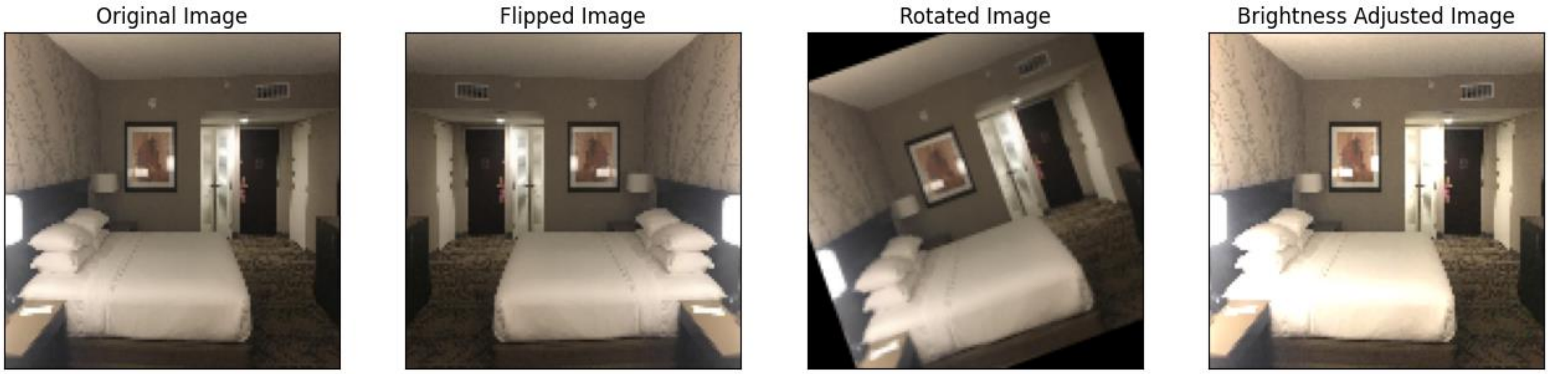
- Data Augmentation

  Data augmentation techniques are applied to artificially increase the diversity of the training dataset. This helps the model generalize better by simulating various conditions that the model might encounter. The plan involves first training the model without any augmentation to establish a baseline performance. After this, data augmentation will be applied, and the improvement in model performance will be measured. We chose the following augmentation methods:

  - Horizontal Flip: Flipping images horizontally to mimic different viewing angles.

  - Rotation: Rotating images randomly within a specified degree range to account for variations in camera angles.

  - Brightness Adjustment: Randomly adjusting the brightness of images to simulate different lighting conditions.



Original Image

Flipped Image

Rotated Image

Brightness Adjusted Image

ARCHITECTURES

For comprehensive evaluation, we will utilize pretrained models (ImageNet dataset) and employ transfer learning with three distinct architectures to understand their performance on the hotel classification task. The chosen models, VGG16, ResNet101, Vit and ArcFace, range from simple to complex and leverage different techniques to handle image data. This approach allows us to explore various aspects of image classification, from traditional convolutional neural networks to advanced transformer-based models.

VGG16

VGG16 is a deep convolutional neural network with 16 layers, primarily using 3x3 convolutional filters and max-pooling layers. Its straightforward design focuses on small filter sizes, allowing for detailed feature extraction. It's widely used for image classification due to its effectiveness and simplicity.

RESNET101

ResNet101 is a deep neural network with 101 layers, using residual learning through skip connections to address the vanishing gradient problem. This allows the network to be deeper and more accurate. It's known for its strong performance in image recognition tasks.

VISION TRANSFORMER (VIT)

Vision Transformer (ViT) uses a transformer architecture for image classification. It divides images into patches and processes them like tokens in language models, capturing long-range dependencies and contextual information effectively. ViT has shown competitive results in image classification benchmarks.

ARCFACE WITH RESNET101 BACKBONE

ArcFace is a face recognition model that uses additive angular margin loss to improve feature discrimination. It ensures larger angular distances between different classes, enhancing recognition accuracy. This model uses ResNet101 as its backbone for robust feature extraction. We use this combination due to its proven success in competitions.

TRAINING

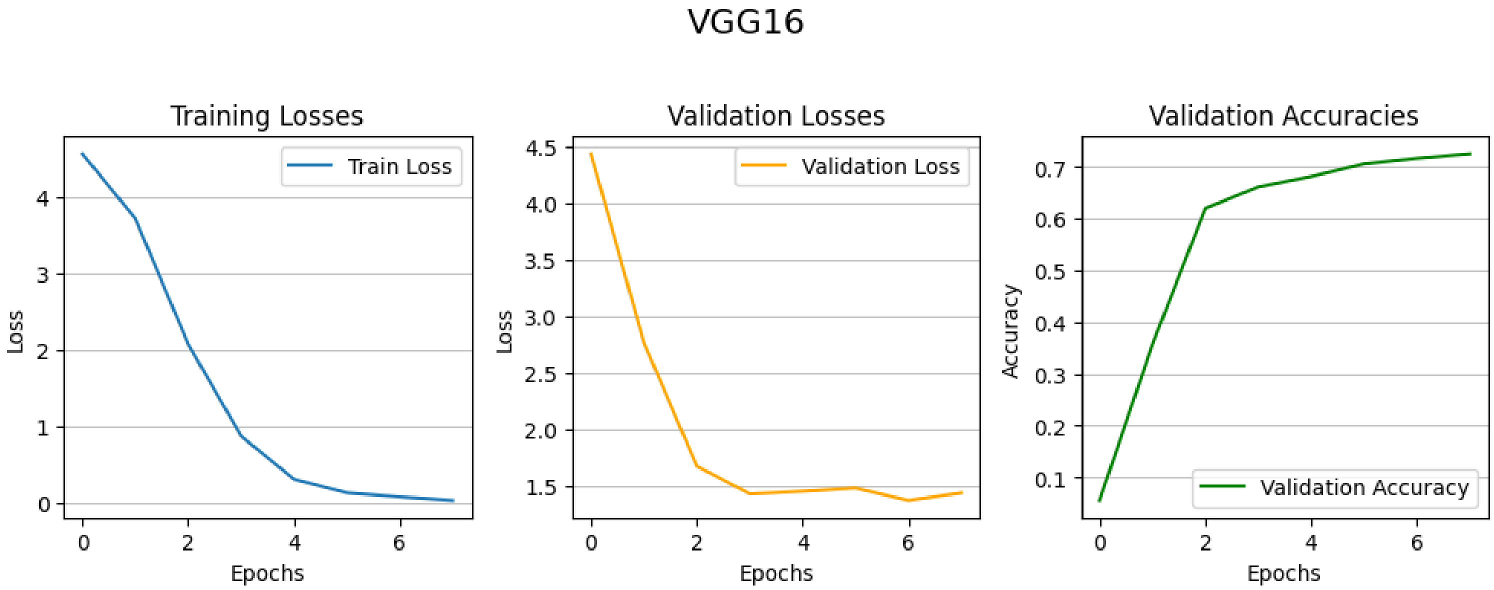
For fine-tuning, we added a fully connected layer matching our number of classes and didn't freeze any layers, allowing all parameters to be modified during the process. This strategy adapts the entire model to our specific dataset and task. We optimized the models using various hyperparameters and settings:

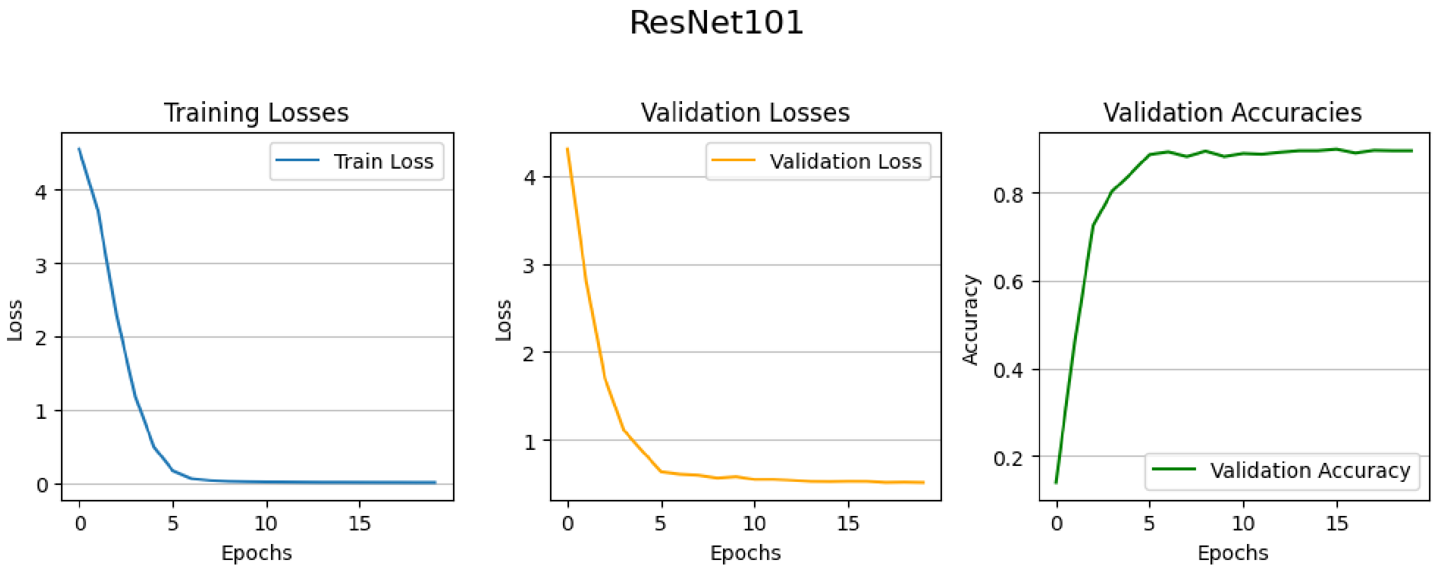
- Learning Rate: 0.001, 0.0001
- Batch Size: 16, 32
- Epochs: 10, 20, 50
- Optimizer: AdamW with Lookahead wrapper to enhance stability and robustness.

- Learning Rate Scheduler: OneCycleLR, dynamically adjusting the learning rate during training.
- Loss Function: Cross Entropy Loss
- Regularization: Early stopping was implemented, halting training if the validation loss didn't improve for 3 consecutive epochs.

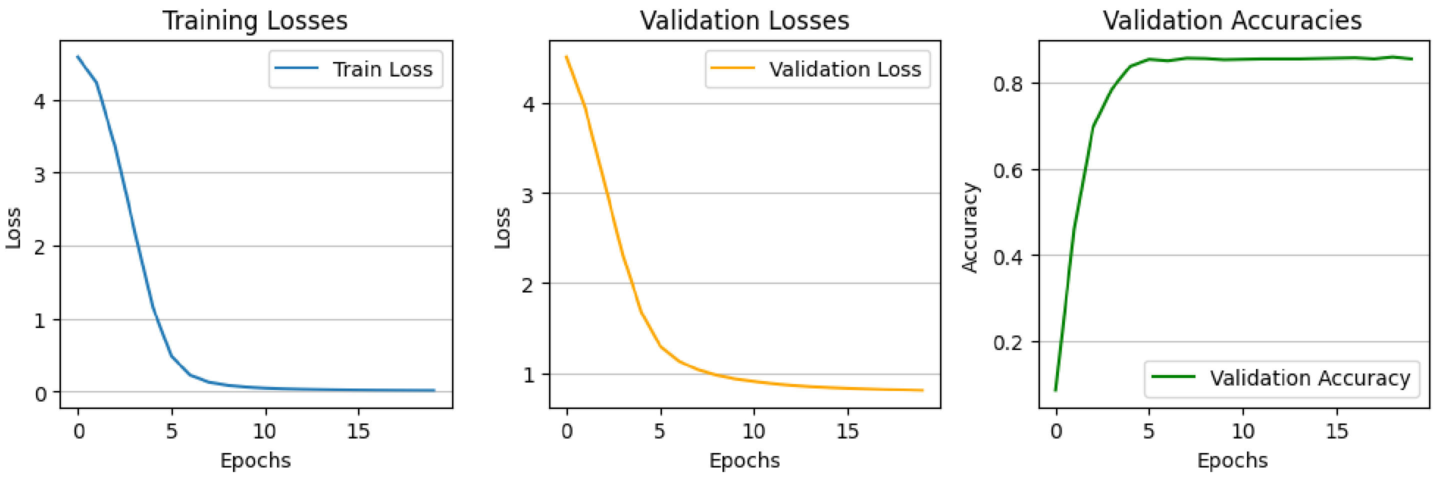
To select the best hyperparameters, we conducted a grid search across different combinations of epochs, batch sizes, and learning rates. Our data was split into training, validation, and test sets using stratified sampling to maintain class distribution.

VISUALIZING MODELS PERFORMANCE

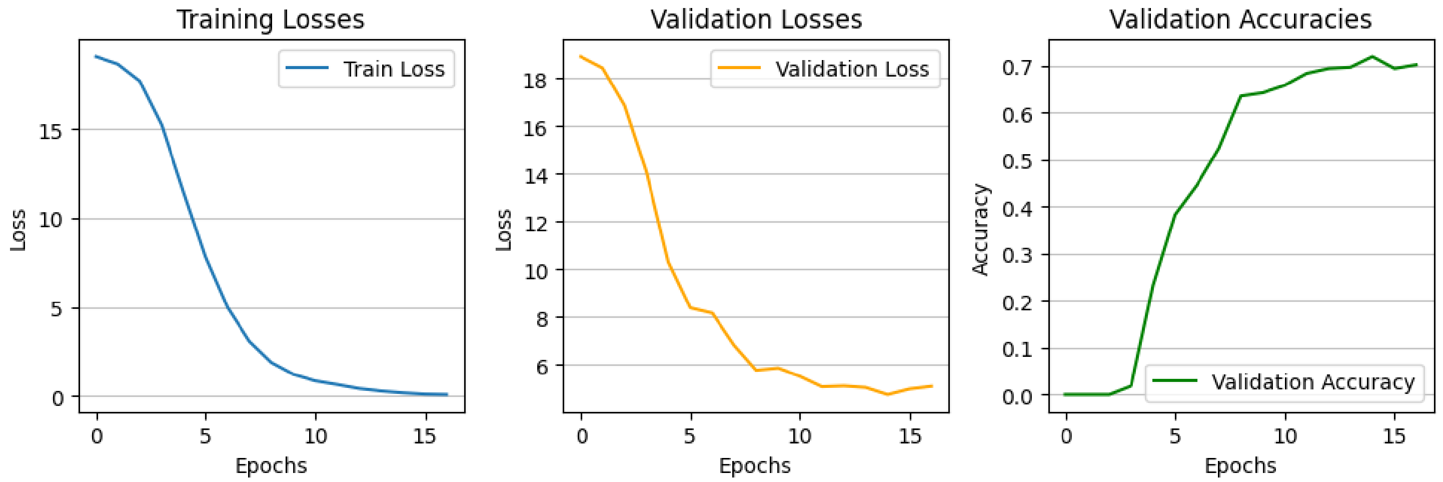




ViT



ArcFace



The graphs reveal that ResNet101 achieved the highest validation accuracy, maintaining strong performance throughout training. VGG16, despite taking only 8 epochs to complete, did not reach a low validation loss or high accuracy, similar to ArcFace. ViT performed well, closely trailing ResNet101 in accuracy. Both ResNet101 and ViT required around 20 epochs for training, while VGG16 was significantly faster but less effective, and ArcFace showed fluctuating performance without achieving high validation accuracy.

HYPERPARAMETER TUNING AND VALIDATION RESULTS

To optimize the performance of our image classification models, we conducted an extensive grid search to find the best hyperparameters. The evaluation metrics used were accuracy, F1-score, and mean Average Precision at 5 (MAP@5). The MAP@5 metric was used in our Kaggle competition on identifying hotels to combat human trafficking. In real-world scenarios, if there is a victim in a hotel, retrieving the top 5 potential hotels for investigation is a reasonable and practical number.

Model	Epochs	Batch Size	Learning Rate	Accuracy	F1	MAP@5
VGG16	8	16	0.0001	72.48	72.33	77.63
ResNet101	20	16	0.0001	**89.57**	**89.34**	**91.71**
ViT	20	16	0.0001	85.45	86.73	88.20
ArcFace	17	32	0.0001	70.20	71.31	71.31

The table shows that a learning rate of 0.0001 was optimal for all models. A batch size of 16 generally yielded better results, except for ArcFace, which used 32 but performed lower. ResNet101 and ViT excelled in accuracy and MAP@5, while VGG16 trained fastest with 8 epochs.

ENSEMBLE MODEL

We implemented an ensemble approach by creating four models: one combining all four (VGG16, ResNet101, ViT, ArcFace with ResNet101), one excluding ArcFace, one excluding VGG16, and one with only ViT and ResNet101. We used soft voting to aggregate the predictions, averaging the softmax outputs from each model. This method leverages the strengths of each model to improve overall accuracy and robustness.

Model	Accuracy	F1	MAP@5
All four models	84.31	84.21	86.92
Excluding ArcFace	85.80	85.60	89.18
Excluding VGG16	83.28	84.42	88.62
ResNet101 + ViT	**89.95**	**89.85**	**92.01**

The ensemble of ResNet101 and ViT was the most successful, outperforming both individual models and other combinations. This ensemble demonstrated superior performance, making it the best choice for further analysis and applications among our ensemble models.

DATA AUGMENTATION

To improve model generalization, we used data augmentation techniques on the training set and measured their impact on the validation set.

- Horizontal Flip: Flipping images to vary viewing angles.
- Rotation: Rotating images to account for different camera angles.
- Brightness Adjustment: Adjusting brightness to simulate varying lighting conditions.

Model	Accuracy	F1	MAP@5
VGG16	82.22	82.13	85.56
ResNet101	88.60	88.63	90.89
ViT	88.43	87.32	89.70
ArcFace	76.77	76.99	78.36
Ensemble	**90.00**	**90.22**	**92.52**

Data augmentation significantly improved the performance of VGG16 and ArcFace, boosting their previously suboptimal results. Conversely, ResNet101 and ViT showed no noticeable benefit from augmentation, possibly due to their robust architectures already capturing sufficient variability. Interestingly, the ensemble model, despite individual augmentations not aiding each component, achieved the highest performance seen yet, underscoring the power of combining diverse models for enhanced generalization.

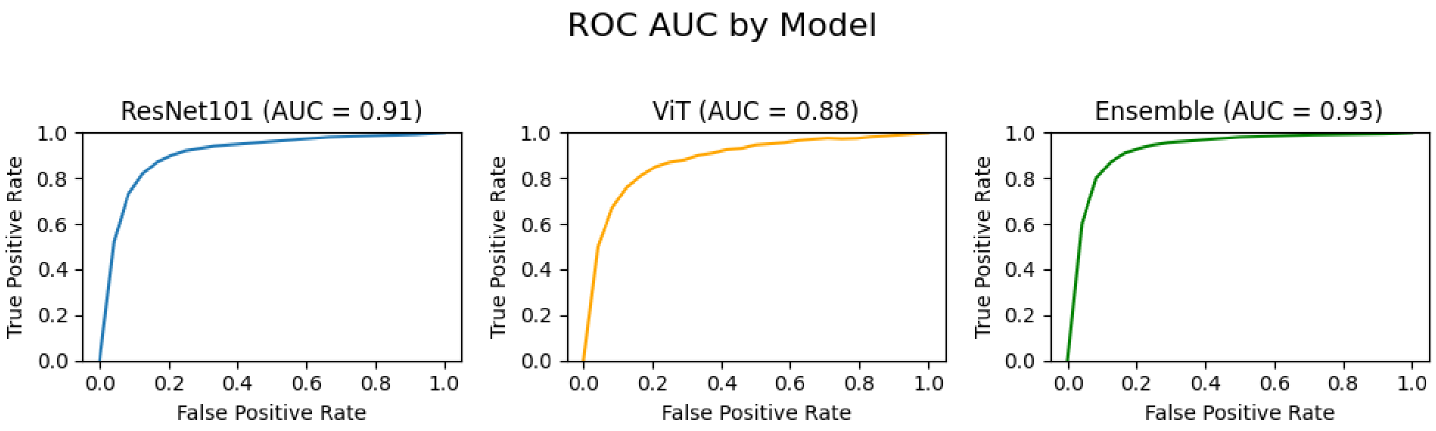
RESULTS

For the testing phase, we decided to evaluate the augmented versions of VGG16 and ArcFace alongside the ensemble model, while using the standard versions of ResNet101 and ViT.

Model	Accuracy	F1	MAP@5
VGG16 (aug)	80.10	79.99	84.22
ResNet101(aug)	88.43	88.46	91.12
ViT	83.87	83.88	87.20
ArcFace	74.93	75.95	76.20
Ensemble(aug)	**91.41**	**91.30**	**93.19**

- ❖ (aug) refers to a model that has been trained on augmented data in addition to the original data.

The augmented versions of VGG16 and ArcFace showed improvement. ResNet101 and ViT performed better overall, and their ensemble surpassed all individual models.



These ROC AUC curves illustrate the performance of our top 3 models: ResNet101, Vision Transformer (ViT), and the Ensemble model. The curves show that all three models perform well, with high true positive rates and low false positive rates. Notably, the Ensemble model appears to have the best overall performance, as indicated by its ROC curve being closest to the top-left corner.

EXPLAINABILITY

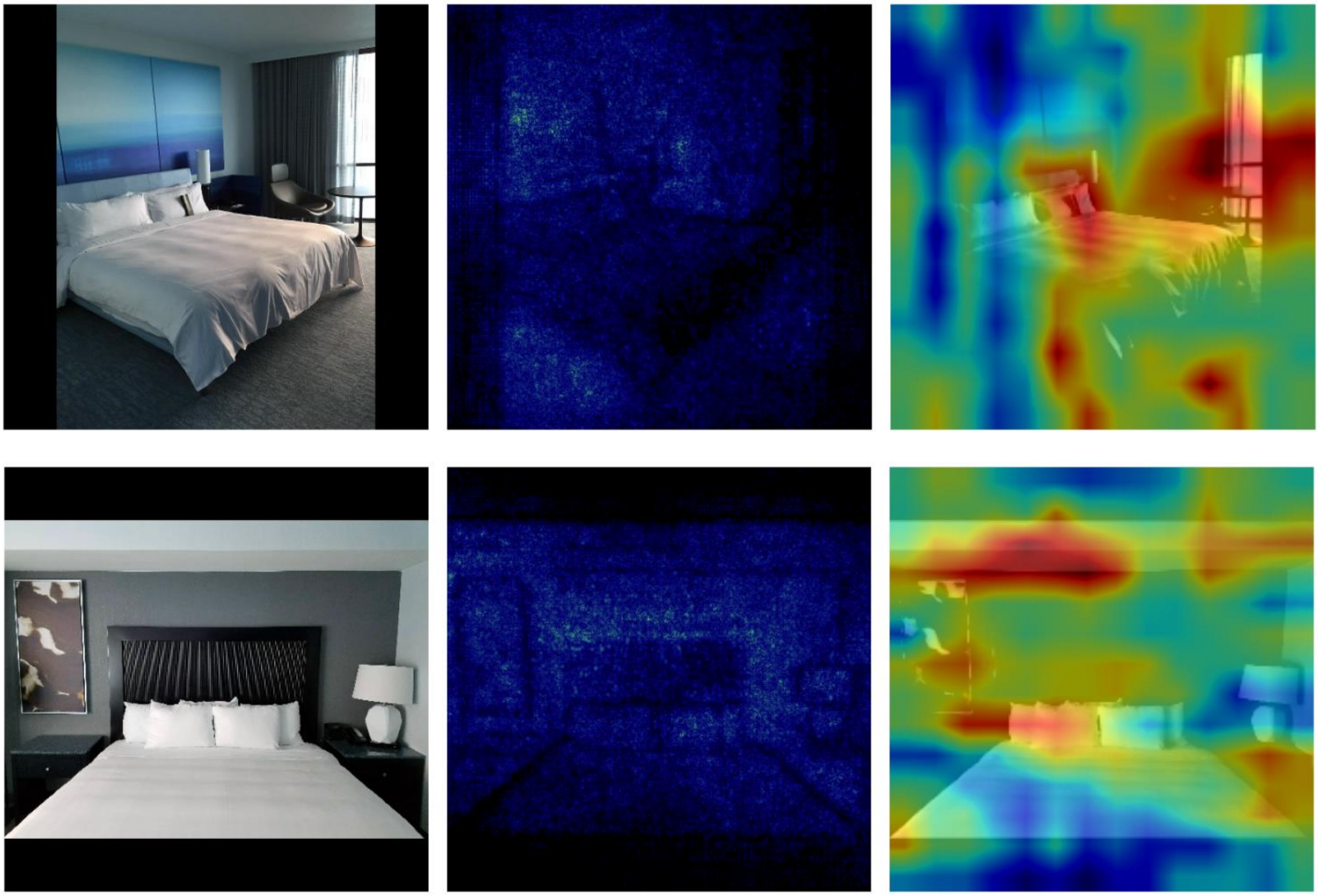
To gain insights into our model’s decision-making process, we generated both saliency maps and attention maps for a subset of our test images. Our leading model is an ensemble of ResNet101 and ViT.

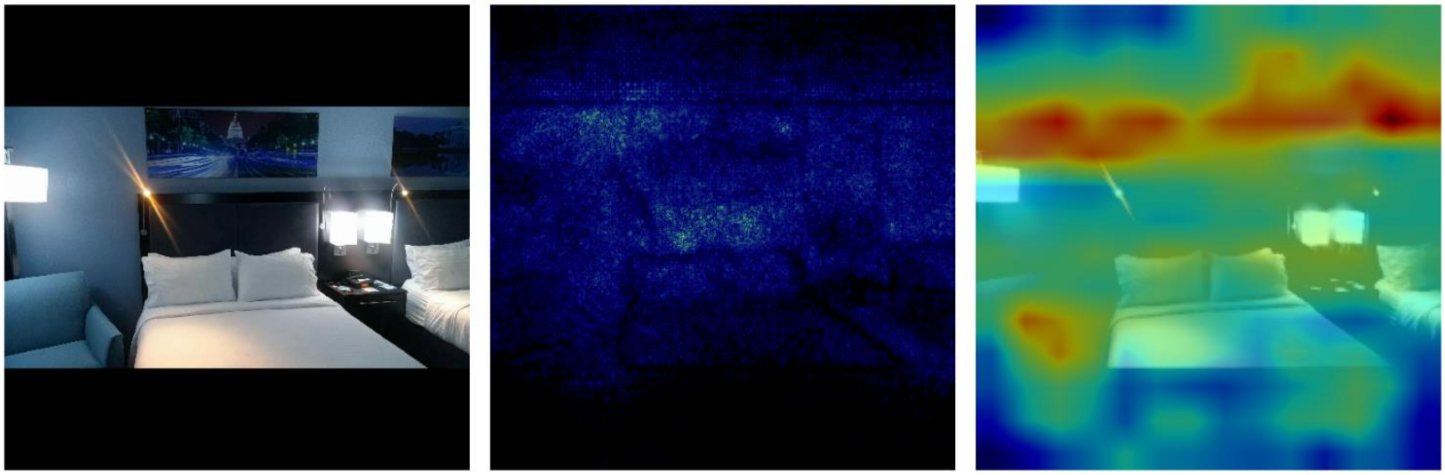
For ResNet101, we used saliency maps, which highlight regions of the input image that have the highest influence on the model’s predictions. These maps allow us to see which parts of the image the model focuses on when making decisions.

For ViT, we used attention maps. Attention maps show which parts of the image are most relevant to the model's output by illustrating how the model distributes its focus across different regions of the image. This helps us understand how the model processes the entire image to make its predictions.

BEDROOMS

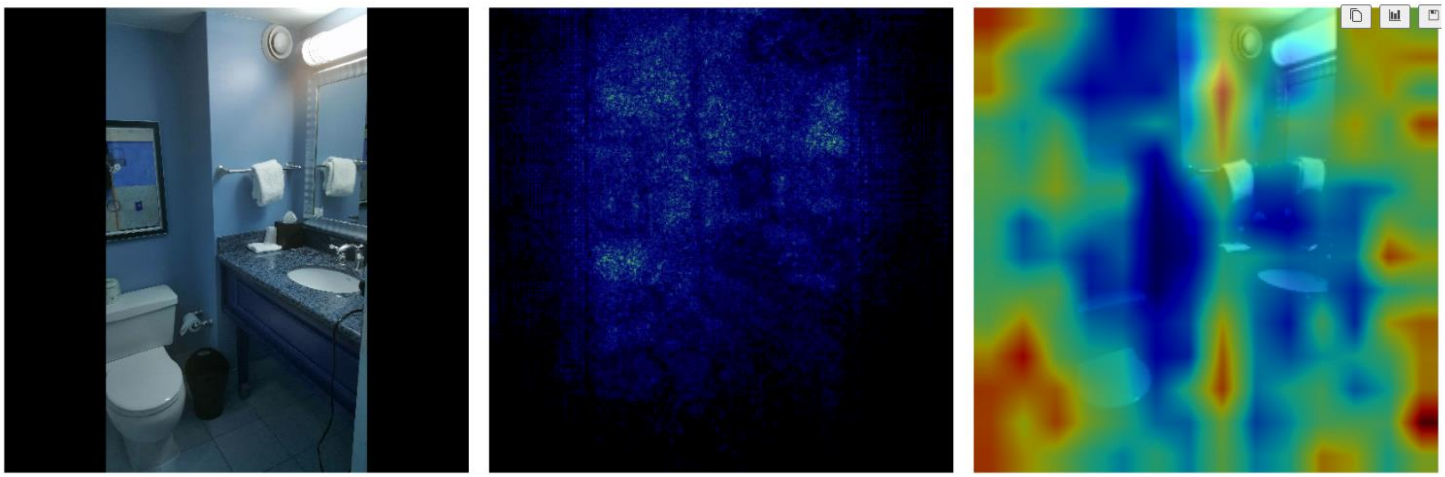
In the following images, the highlighted areas are primarily the walls and other parts of the room rather than the beds. This is likely because beds tend to look very similar across different hotels. Therefore, the model focuses on distinguishing features in the rest of the room to make its identification. Specifically, in the last picture, the model appears to heavily rely on the presence of the lamp as a key differentiating feature.

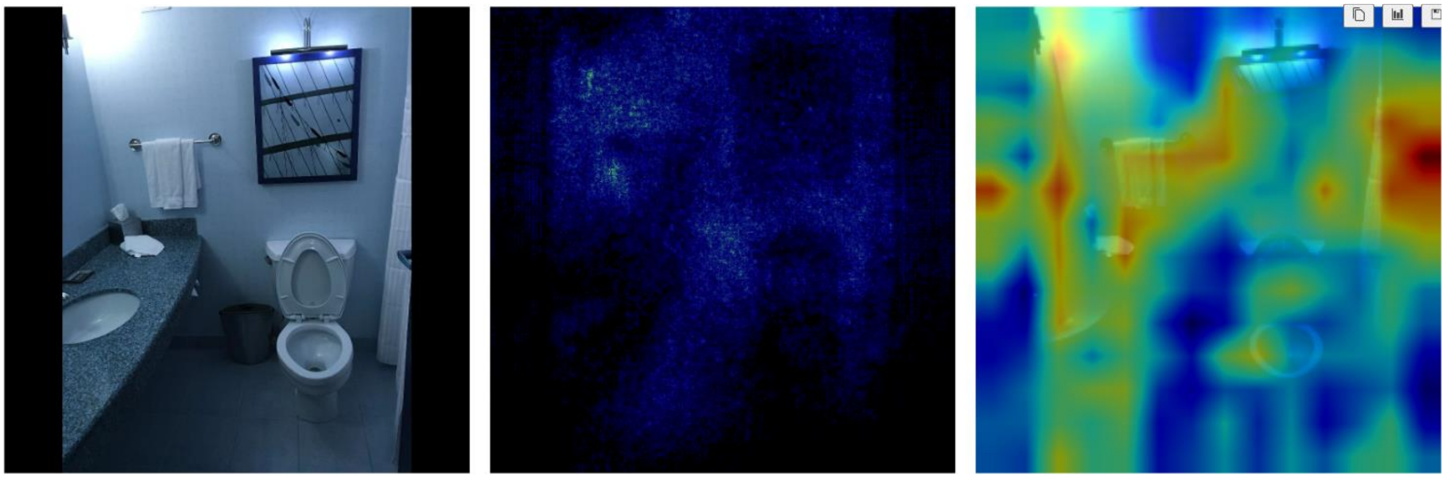


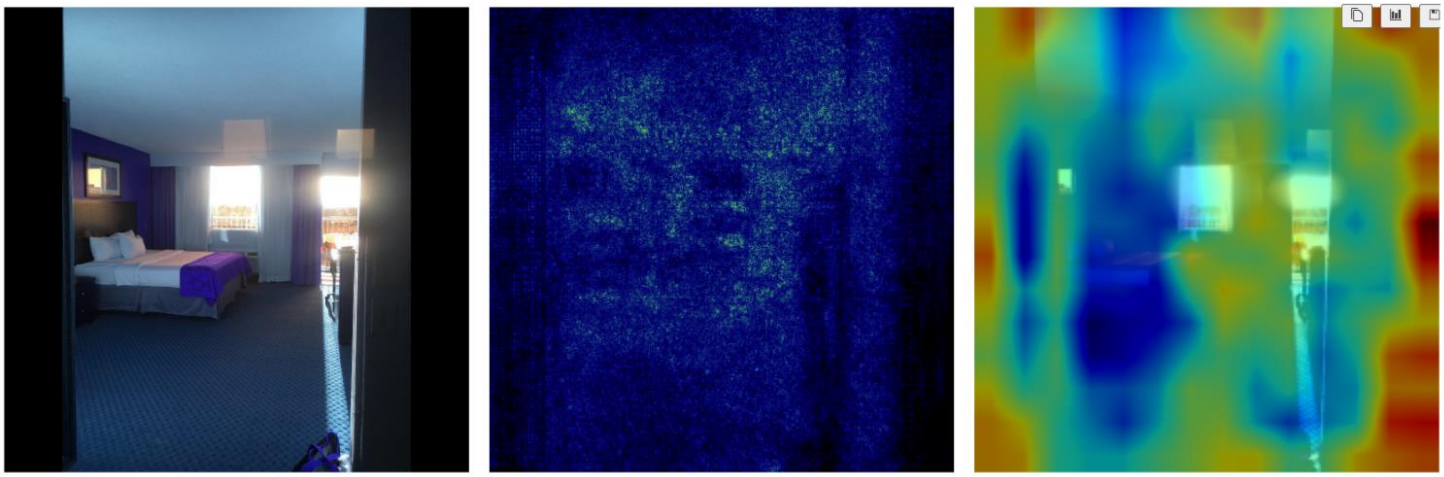


The ViT model occasionally focuses on different parts of the image than the ResNet101, highlighting its complex architecture. However, it performed worse in our task, suggesting it might add more noise than useful insights.

REST OF THE ROOM







For photos of hotel room areas other than the bed, the models spread their attention across a wider part of the image due to greater variance in elements like walls, flooring, windows, and lights. Conversely, they don’t focus much on toilets for the same reason.

FUTURE WORK

While our current model shows promising results in hotel room classification, there are several key areas for future research and development:

- Custom Masking for Obstructions: Develop techniques to handle images where people or objects obstruct key features of the room, improving the model's performance in real-world scenarios.
- Multi-modal Learning: Integrate additional data sources such as geolocation or textual descriptions to enhance classification accuracy.
- Adversarial Training: Implement techniques to make the model more robust against potential attempts to deceive the classification system.

These future directions aim to enhance the accuracy, versatility, and practical applicability of our hotel classification system in combating human trafficking and supporting law enforcement efforts.

CONCLUSIONS

In this project, we applied several image recognition models to classify hotel room images, aiming to aid human trafficking investigations. Through our analyses, we achieved functional results with the best performances from ResNet101 and ViT.

Key findings include:

- Model Efficacy: ResNet101 and ViT provided the highest accuracy and MAP@5 scores, indicating their suitability for handling complex image datasets.
- Impact of Data Augmentation: Horizontal flipping, rotation, and brightness adjustments were effective in improving the generalization capabilities of our models, particularly enhancing models with initially lower validation scores.
- Ensemble Model**:** The ensemble model combining ResNet101 and ViT, particularly when trained with data augmentation techniques, surpassed all other configurations in performance. This

model effectively leveraged the strengths of each component, achieving the highest overall accuracy and MAP@5 scores in our tests. This demonstrates the value of combining multiple models to enhance predictive reliability and accuracy in complex image classification tasks.

- Model Insights: Utilizing saliency and attention maps, we observed that models often focus on distinctive room features other than beds, like walls and decorations, to differentiate between hotels.

Overall, the project demonstrated the practical utility of advanced image recognition technologies in identifying hotel rooms from photographs, which could potentially support legal and investigative processes against human trafficking.

The code for this project is available on GitHub.