Combating Human-Trafficking: Using Scene Recognition to Identify Hotels

Shubham Shetty
shubhamshett@cs.umass.edu

Adarsh Kolya
akolya@cs.umass.edu

Abstract

Human trafficking is a form of modern-day slavery which affects up to 1.2 million children worldwide every year. Victims are often photographed in hotel rooms, hence identifying these hotels is important to support trafficking investigations. We propose to use state of the art deep learning models for scene recognition to identify hotel rooms based on query images, while identifying results using nearest neighbor match. Code for this project is available at https://github.com/shubham-shetty/Hotel-ID-Against-Human-Trafficking

1. Introduction

Human trafficking is estimated to have trapped 24.9 million victims.[4]. The pictures of these victims are often clicked in hotel rooms. Identifying these hotels is vital in the combat of human trafficking. We are trying to build a model that identifies the hotel in which the pictures of the victims had been clicked. We will experiment with ResNet50 [5] based and EfficientNet [11] based models to achieve the objective here.

2. Problem Statement

The problem was presented at the FGVC8 workshop [9]. This is more complex than a simple image recognition task, as there are a huge number of classification labels which is the number of hotels in the dataset.

2.1. Dataset

The primary training dataset we will be using is the 2021 Hotel-ID to Combat Human Trafficking Competition Dataset [6]. This dataset consists of photos of hotel room interiors without any people present. The training set consists of a total of 97554 images from 7770 hotels. The training data is divided into sub folders with a dedicated sub folder for each of the 92 hotel chains that the hotels are part of. We will be splitting this data into training and validation sets and use validation set for evaluating model performance.

Additional model training can be done using the Hotels-50k dataset [10], which contains around a million annotated hotel images and was also built for combating human trafficking.

2.1.1 Exploratory data analysis

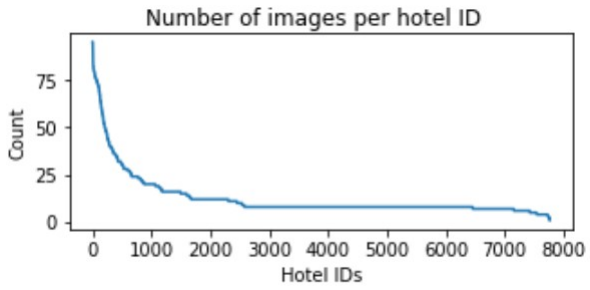

Figure 1. Number of images per Hotel ID

The number of images per class is not fixed. As seen in Fig 1 the number of training examples is higher for hotel IDs 0 - 1000 and the number of examples significantly decrease after that.

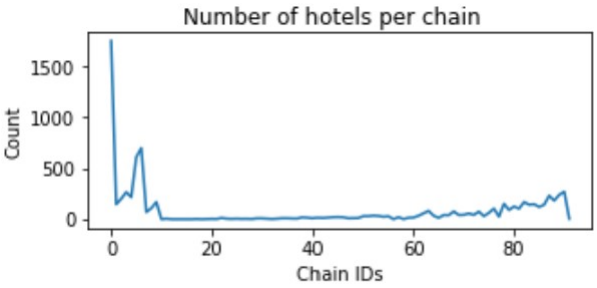

Figure 2. Number of hotels in a chain

The hotels are divided into chains and there are a total of 92 chains. The number of hotels per chain is not fixed either.

Rooms within the same hotel may not look the same as seen in fig 3 and rooms in different hotels (part of the same chain) may look very similar as seen in fig 4. Hence we may have high intraclass variance and low interclass variance. It becomes imperative to minimize the search space efficiently

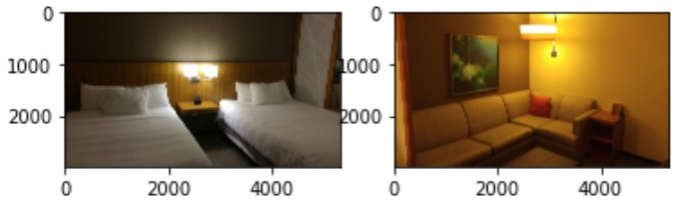

Figure 3. Two images that belong to same class

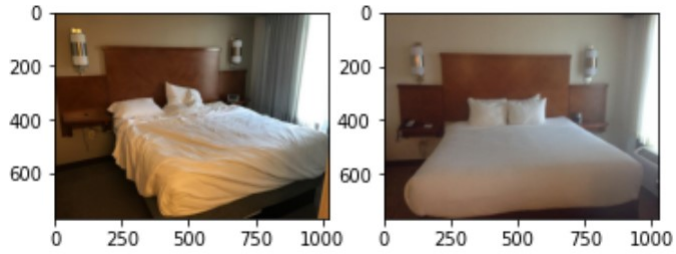

Figure 4. Two images that belong to diff classes

in order to identify possible matches within a reasonable margin of error.

2.2. Expected output

This is a classification problem. For each image in the validation set, we will predict a list of 5 hotel IDs that the image could belong to. This list is sorted so that the probability of the image belonging to the first hotel ID in the list is the highest.

3. Technical Approach

3.1. Related Work

The research team at Nexocode [3] used transfer learning with fastai to classify hotel images. Top layers of ResNet50 (trained on ImageNet data) were used for transfer learning, while fastai was used to normalise data. Using dropout, weight decay regularisation, and data augmentation, they were able to achieve a validation accuracy of 70.4%, rising to 80% with further data cleaning.

Ma et al [7] showed that CNNs can identify more complicated features in hotel images such as rectangles or circles in intermediate layers, which can be stacked for more complex scene recognition. A 152 layer ResNet model was used for final image feature representation.

Nvidia [8] experimented with a triple hinge loss to compare images by learning and understanding aesthetics. Hotels from the same chain should follow similar aesthetics and hence this study becomes relevant for identifying hotels.

3.2. Proposed Approach

3.2.1 Loss Function

According to our data analysis, the data has high intra-class variance and low interclass variance. Additive Angular Margin Loss (ArcFace) [2] is a function proposed for face recognition tasks, which can fit in this scenario as well. ArcFace loss is constructed by modifying the softmax loss.

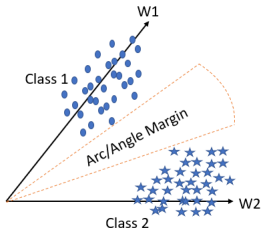

Figure 5. Geometric Representation of ArcFace

The distance between two images can be calculated using cosine distance θ, which is the inner product of two normalized vectors. The features and layer weights are normalized using L2 normalization and inner product is calculated, giving us $\cos\theta$. Mathematically, ArcFace transforms logits $W_j^T x_i = ||W_j||\,||x_i|| \cos\theta_j$, where θ_j is the angle between the weight W_j and the feature x_i. Additionally x_i is rescaled to s after L2 normalization. An angular margin penalty of m is added to prevent weight of fully connected layer to be overly dependent on input data.

Final loss is given by applying softmax to $\cos\theta$ along with the added angular margin penalty.

$$L_3 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\left(\cos\left(\theta_{y_i}+m\right)\right)}}{e^{s\left(\cos\left(\theta_{y_i}+m\right)\right)}+\sum_{j=1,j\neq y_i}^{n}e^{s\cos\theta_j}}$$

3.2.2 Data Preprocessing

Following data preprocessing can be done on the input data -

- Input images resized to 256x256 pixels from 1024. This reduces the size from initial 26 GB to around 8GB.
- *hotel_id* column can be label encoded.
- Training images are shuffled, resized to 64x64, and normalized across RGB channels.
- Data augmentations like flipping, rotation, brightness, etc. can be applied on the data.
- Data can be processed in batches of 32 or 64.

3.2.3 Model

Our proposed model will use a pretrained backbone to generate embeddings for input images. Embeddings can then be used to generate cosine similarity between the test/validation images after which the most similar images can be pulled as matches.

For the intermediate milestone, we will be experimenting with an EfficientNet backbone [11] and ArcFace as the loss function.

4. Preliminary Results

4.1. Baseline using ResNet

The baseline model uses a ResNet50 [5] architecture that is pretrained on ImageNet [1] dataset. We limited the input images to the ones that belong to chain ID 2 only. This reduced the size of the input dataset from 97554 images to 3889 images. This was further divided into training set and validation set in a 9:1 ratio. The input images are resized to 256x256 pixels and are processed with a batch size of 64. The learning rate is set to 1×10^{-4}. The model was trained using triplet margin loss for 20 epochs.

4.1.1 Results and analysis

For each image in the training set and validation set, we used the baseline model to predict 5 hotel IDs that the image could belong to. Mean average precision @ 5 is used to evaluate model performance. This is the same evaluation metric used by the competition [9]. On the training set, the model achieved mean average precision @ 5 of 0.774. The validation set performance was 0.356. The model seems to be over-fitting due to the relatively small size of training set used in the baseline experiments. As we increase the number of training examples, we are confident that this performance will improve.

4.2. Experiments using EfficientNet

4.2.1 Setup

We ran experiments on a small subset of data (around 5000 images) using EfficientNet as a backbone. ArcFace was used as loss function. Following transforms were performed before running the training step -

- Horizontal and vertical flip;
- Shift, scale, and rotate;
- Optical Distortion;
- Perspective etc.

Aim of our experiment was to check whether model is able to load data and learn the objective, optimizing the loss function with each epoch. The hyperparams used for this experiment were -

- Epochs: 2
- Learning Rate: 1e-3
- Batch Size: 8
- Embedding Size: 4096

4.2.2 Results

After running our training procedure for 2 epochs, we observed a decrease in training loss from 11.1824 to 8.0493 (see Figure 6). Total time taken for processing was 15 minutes on Colab using GPU hardware accelerator.

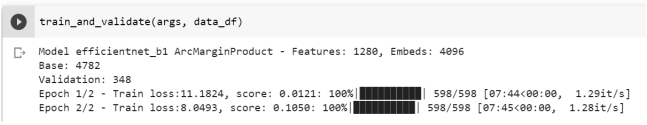

Figure 6. Training Output

4.2.3 Conclusion

For the model using EfficientNet, we are seeing reasonable loss values and a decrease in loss over each epoch. We can confidently move forward and perform the training on complete data with greater number of epochs.

4.3. Next Steps

The next steps for the project, other than performing training on complete data, include -

- Experiment with other backbones such as EfficientNetv2, ResNet, VGG etc.
- Fetch most similar images and predict final hotel.
- Ablation study.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li jia Li, Kai Li, and Li Fei-fei. Imagenet: A large-scale hierarchical image database. In *In CVPR*, 2009. 3

[2] Jiankang Deng, Jia Guo, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *CoRR*, abs/1801.07698, 2018. 2

[3] KORNEL DYLSKI. Transfer learning in practice. image classification for hotel images with fast.ai library, Feb 2019. 2

[4] Human Rights First. Human trafficking by the numbers. https://www.humanrightsfirst.org/resource/human-trafficking-numbers, 2019. 1

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1, 3

[6] Rashmi Kamath, Gregory Rolwes, Samuel Black, and Abby Stylianou. The 2021 hotel-id to combat human trafficking competition dataset. 2021. 1

[7] Yufeng Ma, Zheng Xiang, Qianzhou Du, and Weiguo Fan. Effects of user-provided photos on hotel review helpfulness: An analytical approach with deep leaning. *International Journal of Hospitality Management*, 71:120–131, 04 2018. 2

[8] Appu Shaji. Understanding aesthetics with deep learning, Feb 2016. 2

[9] Abby Stylianou, Rashmi Kamath, Richard Souvenir, and Robert Pless. Hotel-id 2021. https://sites.google.com/view/fgvc8/competitions/hotel-id2021, 2021. 1, 3

[10] Abby Stylianou, Hong Xuan, Maya Shende, Jonathan Brandt, Richard Souvenir, and Robert Pless. Hotels-50k: A global hotel recognition dataset. In *AAAI*, 2019. 1

[11] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *CoRR*, abs/1905.11946, 2019. 1, 3