DL4CV Final Project: Airbnb listing price prediction using ViT

Noam Azmon, Michal Geyer, Tal Sokolov

Abstract

Airbnb is an online marketplace that connects people who want to rent out their homes with people who are looking for accommodations in specific locales. Founded in 2007, it gained massive popularity across populations and housing genres. Thereby, the tips and tricks for having a successful listing are of great interest to many. In this research project, we focused on the connection between a listing's photo album and its price per night. The nature of this problem is that of a regression task. Specifically, we asked whether a listing's price can be predicted by its photos, and which photo contributes the most to the prediction. An important characteristic of this task is that the prediction for a single listing is done based on an album of multiple photos. Taking advantage of the sequential structure of ViTs, we were able to aggregate each album's photos and extract the album's most important photo for the price prediction through its attention layer. When compared with different naive price estimators, our model yielded better precision. Next, we compared the important photo extracted from the attention layer to the photo chosen as the most important by the listing's host. We passed these photos through a pre-trained room-type classification CNN, and compared the room-type distributions of the most important photos extracted from our network and the ones chosen by the hosts. Furthermore, we analyzed the listings with the highest and lowest predictions, spotting meaningful features. We also inspected the distribution of weights in the attention matrix and found that it correlates with the variance within an album.

Motivation/Problem Statement

On an Airbnb listing page, one may find all sorts of information about the offered accommodation; a free-text description of it, a list of its amenities, its locational benefits, host policies, and importantly: photos of the property. The fact that photos significantly affect listings' appeal was recognized by Airbnb early on. Airbnb offers its hosts a professional photography service, In which they connect a host to a local photographer that takes professional photos of the host's offered accommodation. On their website, Airbnb states that good photos of the property bring up to 20% increase in earnings. A precise price estimation tool, accommodated with an image arrangement recommendation system, can be of value.

In our project, we aim to test the relation between the photos of a listing and its price per night. Our main question is - Can a listing's price be predicted by its photos? Furthermore - what is the most important photo for this kind of prediction? The nature of this question is that of a regression problem, predicting some continuous value for each album. Since, in contrast to classification problems, this is not a common use of CNNs, let alone ViTs, we find this to be an intriguing field of investigation.

Related work

Airbnb as a rich data set with business-related interests has been studied. Different tools were applied in order to categorize its visual data: *Object detection in Airbnb room photos* [1], *Categorizing listing photos into different room types* [2].

Property evaluation is a problem with clear applications which was considered in *Vision-based real estate price estimation* [3].

Referring to a group of photos as a sequence with an assessment of the most important photo was presented in *Photo Albums Event Recognition using Transformers Attention* [4]. We use the latter's architecture as our basis, with the significant change of treating the problem as regression rather than classification.

Method

The input to our model is the first five photos that appear on the front page of an Airbnb listing (Image 1.). The order of the five photos representing an apartment is shuffled to rule out order biases. These five photos are chosen as they dominate the user's first impression.

Image 1. An example for Airbnb's listing main page photo album. The first photo is larger at scale.

Inspired by the PETA architecture [4], we apply PyTorch's pre-trained ResNet152 [5] backbone, trained on ImageNet [6], as a feature encoder. We get the embeddings $x_k = backbone(I_k) \mid k \in \{1,5\}$.

For each image embeddings x_k a learnable positional encoding is added: $z_k = x_k + e_k^{pos} \mid k \in \{1,5\}$.

To apply regression using transformers, a price token (PRC) is concatenated to the whole listing representation. The input to the first layer of the transformer encoder is then $(PRC, \{Z_k\}_{k=1}^{5})$. The output of the transformer encoder through the MLP head is a price prediction.

The photo's importance is determined using the attention layer (At). Recall that the PRC was concatenated to the transformer's input as the first token. We, therefore, extract photo importance for photo I_k from the first row of the attention matrix, as $\alpha_k = At_{0,k+1} \mid k \in \{1,5\}$. The most important photo is determined using $argmax_k\{\alpha_k\}$.

After exploring a range of learning parameters, we proceeded with 20 listings (100 photos) batch-size, a single transformer block, an exponential learning rate scheduler, ADAM optimizer, and MSE loss. Batches' order was shuffled between epochs to avoid learning of order-specific correlations.

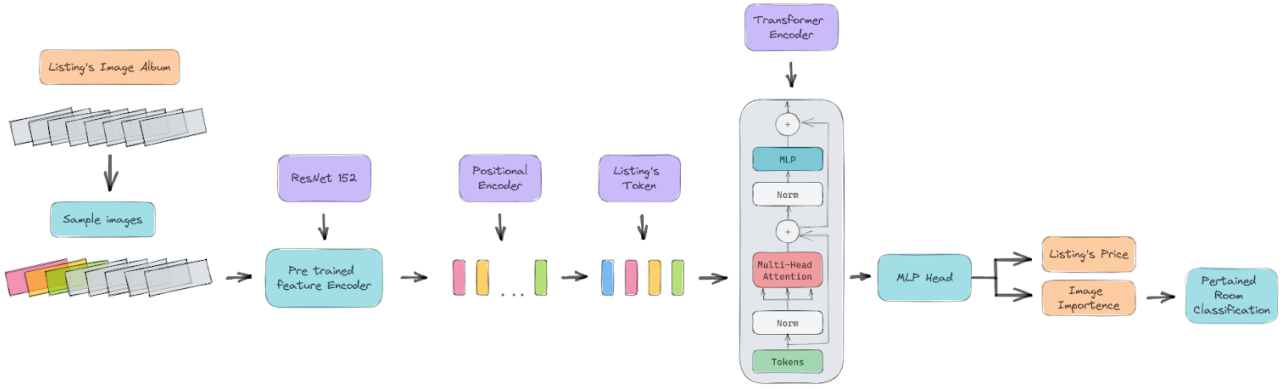

Fig. 1. The architecture

In order to obtain meaningful analysis, the photos are also passed through a pre-trained network to label the photos by room type. Using Monk library [7] that offers a house room classifier [8] based on ResNet18 [9], we labeled each photo with one of seven classes (bathroom, bedroom, dining room, exterior, interior, kitchen and living room). This classifier claims to provide about 85% accuracy on its original validation data, which was acquired via real estate databases [10].

Data Acquisition and annotation

The data of 30,600 Airbnb apartments listings were collected using Kaggle [11] datasets. The apartments are located in New York City, Berlin, Istanbul, Athens, and Toronto, aiming to represent worldwide and cross-continent distributions. Filters were applied to avoid unwelcome effects, as follows. The model relies on 5 input photos so listings with less than that are removed. We focus on rent terms of short vacations, especially in the context of pandemic times where some of the landlords redefine the purpose of their properties to long-term rentals not intended for tourism. Listings with no availability throughout the entire year were also removed, suspected as not active, and not updated. We remove the effect of misleading prices by setting a minimum value for a listing's price. Less than a dozen listings in each city appeared extremely luxurious and overpriced, and were removed as outliers. The applied filters left us with about 12,216 (2,736 in New York, 3,575 in Athens, 3,318 in Istanbul, 1,238 in Toronto, and 1,349 in Berlin) listings, which are 61,080 property photos. A price of a property was evaluated as the mean single night price over the next year. To avoid geographical biases, the prices were normalized within each country using the IQR robust measurement [12] of scale ($\frac{x-\mu}{Q_{0.75}-Q_{0.25}}$, Q_i is the i'th quantile and μ is the median). The resulting price distributions within each county vary, but exhibit similar ranges (Fig. 2). The prices to be predicted are the scaled prices. The Data set was split randomly into 80% train 20% test, with these proportions kept for each country to get a proper representation of the variance.

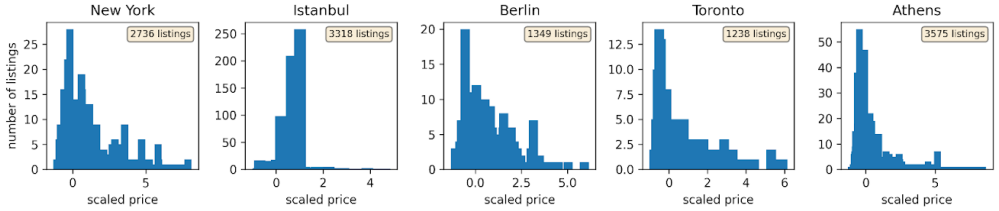

Fig. 2. Histograms of the listings' scaled prices over the five cities.

Experiments and analysis

Looking at the loss curve over epochs of the train and the test sets (Fig. 3), we can see a learning curve for the train, as expected of a learning model, with quite a noisy test curve that stabilizes slightly above the training loss. To evaluate our prediction, we compared it with the loss of naive models. Models assuming the median or the mean price of all listings achieved a higher loss value of 1.36 and 1.05 accordingly. SciKit's [13] linear regression model receiving the same embeddings used in our model, concatenated, achieved worse results by an order of magnitude. Reverting the scaling of the prices, we derive the loss in USD for each city (Fig. 4). We see that the loss is persistently better than the loss of naive median and mean prediction of each city.

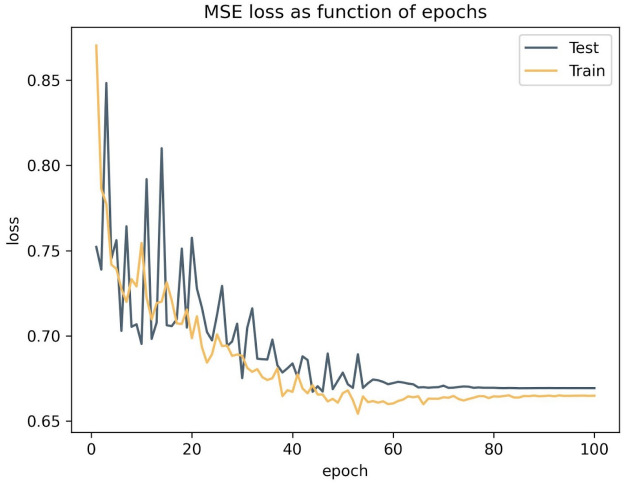

Fig. 3. Loss curve over epochs for the test and train sets.

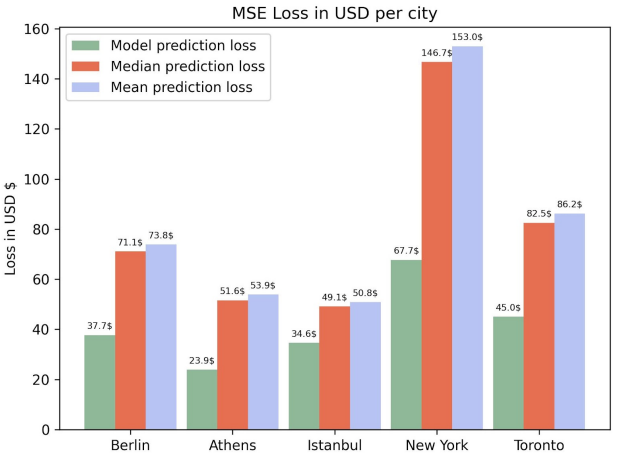

Fig. 4. Loss distribution across cities compared with the naive median and mean prediction loss for each city.

On the front page of a listing, the first photo displayed is significantly larger by size relative to the other four photos. It seems reasonable to assume that the photo chosen as the first photo (Image 0) is considered by the host as the most important photo. We ask what is the distribution of room types of Image 0 of each listing, compared with the distribution of the room types of the photo the network indicated as the most important (Fig. 5a). We can see that the distributions resemble, indicating that even if Image 0 itself was not the model's choice as most important, the room it represents was so. The photo order distribution (Fig. 5c), combining the fact that the order of the photos is shuffled within each listing, can eliminate assumptions of biases regarding such conclusions. A comparison of the most important photo chosen by our model and by the host, for the listings in which our model's loss was the smallest (Fig. 5b), shows the choices as either identical or of the same room. Repeating the comparison for the listings in which our model's loss was the highest didn't show such a correlation.

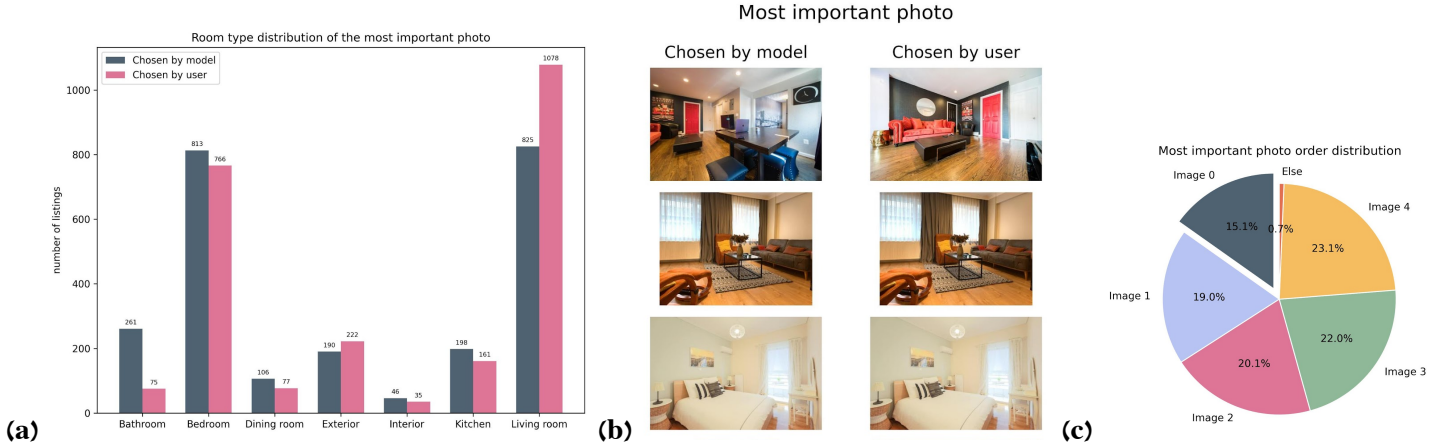

Fig. 5 (a). The most important photo room type distribution compared between the model's and the host's choices. (b) Most important photo chosen by the model and by the host, on the listings on which the model's loss was the smallest (i.e., best predictions of our model). (c) The distribution of the listings' most important photo's order within the album.

We inspect the attention output for the listings with extremal price predictions. Among the sixteen listings that were predicted with the highest scaled price, the most important photos display open spaces, wide angles, bright colors, and natural light. If we allow ourselves for a moment to involve human interpretation, these photos induce a positive atmosphere. On the contrary, among the sixteen listings that were predicted with the lowest scaled price, the most important photos display small, almost claustrophobic, spaces, with dim colors. Repeating the same analysis for the ground truth highest and lowest priced listings, we could not detect any distinct visual characteristics between the two groups.

Fig. 6. Left (right) is the most important photo of the sixteen listings with the lowest (highest) predicted prices.

We analyzed the difference between the attention relative weight of different photos in the album, quantifying the importance of the most important photo (i.e. the photo with the highest attention weight in the album). We focused on listings where the importance (i.e. the attention relative weight) of the most important photo was the lowest (Fig. 7. left) and listings where it was the highest (Fig. 7 right). Note that for listings in which the most important photo was given a low relative weight, the most important photo is consistently similar to the other photos. On the other hand, for listings in which the most important photo was given a high relative weight, the most important photo is significantly different from the others. This may indicate that the attention weight distribution holds the variance between the photos in the album.

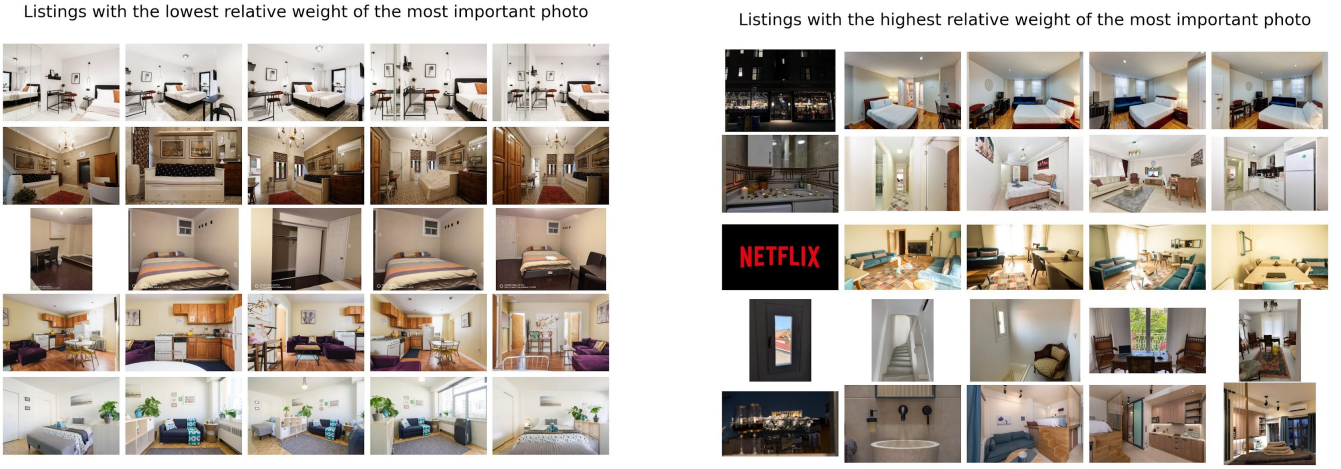

Fig. 7. Left (right) is the photos of the listings with the lowest (highest) relative weight of the most important photo.
Each row is a single listing album, and the leftmost photo of each row is the most important one (i.e. with the highest attention weight).

Conclusion

The use of ViT in our project allowed us not only to create a model that considers the sequential structure of our data (albums) and outperforms naive models, but also to inspect and analyze different aspects of the price prediction task. This was also enabled by the accessibility of powerful tools such as PyTorch's ResNet and Monk pre-trained networks. Analysis of the room type distribution of the most important photo showed that even when our model did not choose the exact same photo as the host to be the most important, in most cases, it chose the same room type (Fig. 5). Analysis of the photos marked as most important by the attention layer in our model exhibited that the model learned to associate visual features with high and low prices (Fig. 6). Finally, quantification of how-much-more-important was the most important photo, showed that the attention distribution holds information on the variance between photos within an album (Fig. 7). It is essential to mention that, to begin with, the problem of price prediction can not necessarily be solved properly using visual information solely. For example, the information on the location attractivity (e.g., city center, neighborhood safety, etc.) might be crucial for an accurate prediction. In addition, the aspiration to model a worldwide phenomenon might have been a far-fetched easing of conditions. For example, the normalized price distribution shows that Istanbul's distribution differs from the others (Fig. 2), which might have affected our results, as can be detected in the loss analysis (Fig. 4). Modifications considering these points are interesting follow-up research directions.

Code can be found at https://github.com/azmonoam/airbnb_dlp

References

[1] Shijing Yao, 2019, *Amenity Detection and Beyond — New Frontiers of Computer Vision at Airbnb*
https://medium.com/airbnb-engineering/amenity-detection-and-beyond-new-frontiers-of-computer-vision-at-airbnb-144a4441b72e

[2] Shijing Yao, 2018, *Categorizing Listing Photos at Airbnb*
https://medium.com/airbnb-engineering/categorizing-listing-photos-at-airbnb-f9483f3ab7e3

[3] Omid Poursaeed, Tomas Matera, Serge Belongie, 2018, *Vision-based Real Estate Price Estimation*
https://arxiv.org/abs/1707.05489

[4] Tamar Glaser, Emanuel Ben-Baruch, Gilad Sharir, Nadav Zamir, Asaf Noy, Lihi Zelnik-Manor, 2021, *PETA: Photo Albums Event Recognition using Transformers Attention*
https://arxiv.org/abs/2109.12499

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, 2015, *Deep Residual Learning for Image Recognition*
https://arxiv.org/abs/1512.03385

[6] *ImageNet*
https://www.image-net.org/

[7] *Monk repository*
https://github.com/Tessellate-Imaging/monk_v1

[8] *Monk House room type Claasifier*
https://github.com/Tessellate-Imaging/monk_v1/blob/master/study_roadmaps/4_image_classification_zoo/Classifier%20-%20House%20room%20type%20Claasification.ipynb

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, 2015, *Deep Residual Learning for Image Recognition*
https://arxiv.org/pdf/1512.03385.pdf

[10] Omid Poursaeed, Tomáš Matera, Serge Belongie, 2018, *Vision-based real estate price estimation*
https://omidpoursaeed.github.io/publication/vision-based-real-estate-price-estimation/

[11] *Kaggle Datasets*
https://www.kaggle.com/datasets

[12] Jason Brownlee, 2020, *How to Scale Data With Outliers for Machine Learning* Blog post
https://machinelearningmastery.com/robust-scaler-transforms-for-machine-learning/

[13] *SciKit-learn model*
https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html