 DOI: 10.3217/978-3-85125-869-1-14

Case Study: Ensemble Decision-Based Annotation of Unconstrained Real Estate Images

Miroslav Despotovic[1], Zedong Zhang[1], Eric Stumpe[2] and Matthias Zeppelzauer[2]

***Abstract*— We describe a proof-of-concept for annotating real estate images using simple iterative rule-based semi-supervised learning. In this study, we have gained important insights into the content characteristics and uniqueness of individual image classes as well as essential requirements for a practical implementation.**

I. INTRODUCTION

The annotation of unlabeled images is an important task for the assignment of metadata, which can be particularly challenging within a given knowledge domain. Thus, image metadata is being increasingly used in real estate research, e.g., for valuation [9], location analysis [5], or for estimating the condition of a building [4]. In the scientific literature, there are very few contributions on the classification of unlabeled images in the domain of real estate [7]. In this short paper, we present an approach to semi-supervised labeling of images containing interior and exterior views of real estate using simple ensemble classification rule.

II. PROBLEM STATEMENT

To maximize the information potential of the data, it must be tagged with meaningful labels, which in practice can require considerable manual effort. A typical approach for annotating unlabeled data autonomously is semi-supervised learning (SSL), where an initial training set of labeled data $\mathcal{D}_l$ is defined by clustering and/or manual selection and the trained model is used to infer unlabeled data $\mathcal{D}_\cup$ systematically without interactively querying the user (e.g. active learning with embedded Human-in-the-Loop) [6]. Our motivation for this case study is to provide a proof-of-concept for setting up a model for automatic pre-selection of images from large unlabeled datasets that may be used for training ConvNets to learn the visual clues that are indicative of the quality of real estates. This work is therefore intended to serve as the basis for a more extensive follow-up study. Thus, the main incentive is to investigate how the proposed model processes complex intrinsic properties of real estate photographs, as well as which domain-specific labels are generalized well by the classifiers.

*This research was funded by the Austrian Research Promotion Agency (FFG) project 880546 "IMREA" and we are very grateful to DataScience Service GmbH for providing the data for this study.

[1]M. Despotovic and Z. Zhang are with the Kufstein University of Applied Sciences, Kufstein 6330, Tirol, Austria (miroslav.despotovic@fh-kufstein.ac.at; zedong.zhang@fh-kufstein.ac.at)

[2]E. Stumpe and M. Zeppelzauer are with the ICMT Institute of Creative Media Technologies, St. Pölten University of Applied Sciences, St. Pölten 3100, Lower Austria, Austria (estumpe@fhstp.ac.at; matthias.zeppelzauer@fhstp.ac.at)

Real estate images have different resolutions or were taken under different lighting conditions with varying distances and angles to the object. An additional challenge is that there are only a limited number of relevant labels, and it is a priori unclear which classes can even be captured from the images. The data contains noise, samples that cannot be attributed to a specific property characteristic, as well as redundant information because real estate developers in local markets often work with multiple agencies for advertising and sales.

III. APPROACH

We make the naive assumption that empirical error in the decision boundary can be minimized by exploiting the generalization capability of multiple ConvNets, provided that a large amount of training data is available. In this regard, we propose a SSL procedure as follows.

A. Iterative training

We use annotated data to iteratively fine-tune VGG16 [8] and ResNet101v2 [2] (both pre-trained on the large ImageNet dataset), starting from the initial training dataset S_i. That is, after each complete iteration, we infer labels in the unlabeled dataset $\mathcal{D}_\cup$ with fine-tuned networks N_1 and N_2 and enrich training datasets S_1 and S_2 (one set per network) with new instances. Thereby, we select randomly, at a lower threshold of 100% accuracy, 5 predictions per class and network and add them as new instances to the prior training sets. This process is performed sequentially until we obtain training sets S_1 and S_2 with 5000 instances each. The selection of 5 matches per class is deliberate to reduce the target risk due to the learner's prior knowledge [7]. The determination of false predictions in the S_1 and S_2 is carried out within the definition of experiment baselines (see IV-C).

B. Ensemble decision

We build a dataset S_{tr} consisting solely of instances of S_1 and S_2 that are predicted in concordance by both networks. The inference of the SSL model is then evaluated by fine-tuning a VGG16 with S_{tr} and testing it with an independent dataset T_1.

IV. EXPERIMENTAL SETUP

A. Data

The preprocessing of the data initially involves duplicate removal by image-wise assignment of unique hash values and calculating difference using Hamming distance. After this step, our experimental data set $\mathcal{D}_\cup$ eventually comprises 47k images. However, some redundant information remains,

as agencies often add their logos when editing photos or post-processing the image for marketing purposes.

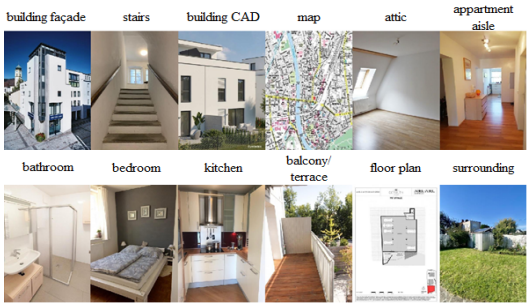

Fig. 1. Experimental selection of real estate classes, Image source: [1]

For our study, we use a manually pre-selected ground truth set $\mathcal{D}_t$ with 12 meaningful classes from the perspective of real estate valuation. Figure 1 shows the experimental class selection. This set is then partitioned into training S_i, validation V_1 and test T_1 datasets with a ratio of 1473-375-240 instances and 12 balanced classes per set. We control our experiment by setting multiple baselines (see IV-C) with training sets S_i, S_3 and S_4 (see Table I). S_3 is a manually selected subset of S_1 where only correctly predicted labels are kept. S_4 is defined like S_3 with the exception that the incorrectly predicted labels are not excluded but manually added to the images with correctly predicted labels from S_1.

B. Setup & Training

For the training we utilize extensive data augmentation including centering, rescaling and shifting. Training parameters for both nets are learning rate of 0.001, decay of 0.001, momentum of 0.9 and a batch size of 40 for N_1 resp. 100 for N_2 . All nets were trained with cross-entropy loss and adamax optimizer [3]. A full SSL iteration was initially set to 200 epochs and successively reduced: $I_1 = 200, I_2 = 200//2, I_3 = 200//3, I_4 = 200//4, ..., I_n = 200//4$. Since we observed higher loss/accuracy variability in the earlier and later training phases, a larger number of epochs was deliberately chosen. Thus, we do not apply early stopping for regularization but select the training stage with the best performance.

C. Evaluation

We aim at answering following research questions: (1) are the individual classes sufficiently discriminative to achieve an acceptable generalization of the classifier? and (2) can the proposed experimental SSL approach achieve a comparable result to the established baselines? To measure the performance of the model, we set up multiple baselines whose performance was evaluated with the test set T_1. The lower baseline is defined as the performance of a fine-tuned VGG16 trained on initial training set S_i. The mid baseline is specified through the performance of a fine-tuned VGG16 trained on S_3. Finally, we define an upper baseline as the performance of a fine-tuned VGG16 trained on S_4.

V. RESULTS

In the Figure 2 showing Receiver Operating Characteristic (ROC) for each predicted class, a larger deviation is noticable for class 4 (map), followed by class 12 (surrounding) and class 10 (balcony/terrace). These are basically classes that do not represent interior spaces. An expected confusion can be seen between class 1 (building facade) and class 3 (building CAD). On the other hand, all classes with interiors were particularly well recognized by the classifier, indicating their discriminative visual content. However, false-positive test results point to a minor misinterpretation for classes attic and staircase.

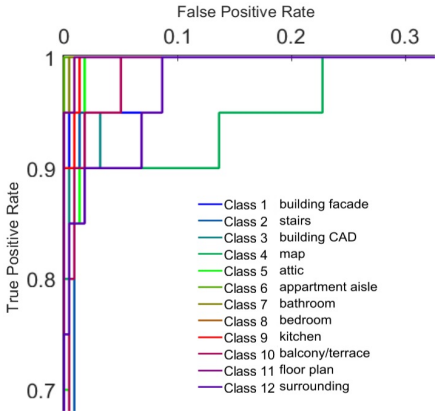

Fig. 2. ROC of individual classes.

Table I shows that the SSL model slightly underperforms lower and middle baseline, but the performance is almost consistent with the upper baseline. This is attributed to the larger proportion of false positives for classes stairs, building facade and building CAD in S_{tr} (compared to S_1 and S_2) and thus the inconsistent class balance during the training. Notably, the overall class balance in S_{tr} (intentionally not supervised) expressed by coefficient of variation CV (18%) is smaller than CV for S_3 (30.2 %) and S_4 (41.7 %).

With this study, we have gained first insights into the challenging task of enriching metadata from real estate images. We intend to build on the results of the presented approach in a more comprehensive follow-up study to gain further valuable evidence.

TABLE I
COMPARISON OF CLASSIFICATION ACCURACY (IN %) FOR SSL MODEL AND BASE MODELS.

training dataset	VGG16 lower baseline			training dataset	VGG16 mid baseline		
	sample size	validation	test		sample size	validation	test
S_i	1473	97.28	92.08	S_3	2661	96.2	91.67
training dataset	VGG16 upper baseline			training dataset	VGG16 ResNet101v2 ssl		
	sample size	validation	test		sample size	validation	test
S_4	3448	95.92	90.00	S_{tr}	3005	94.84	89.17

REFERENCES

[1] Justimmo- einfach makeln! B&G Consulting & Commerce GmbH. [Online]. Available: https://www.justimmo.at/

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[3] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2015.

[4] D. Koch, M. Despotovic, M. Sakeena, M. Döller, and M. Zeppelzauer, "Visual estimation of building condition with patch-level convnets," *Proceedings of the 2018 ACM Workshop on Multimedia for Real Estate Tech*, 2018.

[5] V. Muhr, M. Despotovic, D. Koch, M. Döller, and M. Zeppelzauer, "Towards automated real estate assessment from satellite images with cnns," in *Proceedings of the 10th Forum Media Technology (FMT)*, vol. 2009, 2017, pp. 14—-23.

[6] C. Persello and L. Bruzzone, "Active and semisupervised learning for the classification of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 11, pp. 6937–6956, 2014.

[7] P. Pourashraf and N. Tomuro, "Use of a large image repository to enhance domain dataset for flyer classification," in *ISVC*, 2015.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2015.

[9] Y. Zhang and R. Dong, "Impacts of street-visible greenery on housing prices: Evidence from a hedonic price model and a massive street view image dataset in beijing," *ISPRS International Journal of Geo-Information*, vol. 7, no. 3, p. 104, 2018.