EasyChair Preprint

№ 5096

Visual Semantic Context Encoding for Domain Prediction of Aircrafts

Andreas Kriegler, Daniel Steininger and Wilfried Wöber

March 3, 2021

VISUAL SEMANTIC CONTEXT ENCODING FOR DOMAIN PREDICTION OF AIRCRAFTS

Andreas Kriegler[†,††]
[†]Austrian Institute of Technology
[††]University of Applied Sciences Vienna

Daniel Steininger[†]
[†]Austrian Institute of Technology

Wilfried Wöber[††]
[††]University of Applied Sciences Vienna

ABSTRACT

In existing CV works visual semantic context is often learned implicitly - this work uses an explicit representation instead and makes two distinct contributions: Firstly, it is shown that during data aggregation context can be used to remove irrelevant images. Secondly, extending the idea of context across multiple images, objects can be observed in characteristic domains. An original baseline, supervised CNNs and unsupervised mixture models are used to predict domains of airplanes. A CNN achieves the best classification performance with accuracies from 69% to 85% depending on the dataset variation. The entire framework can be applied to predict arbitrary domains of objects and provide a higher-level sense of scene understanding.

Keywords: semantic context, domain inference, aerial data

1. INTRODUCTION

Visual semantic context describes the relationship between objects and their surroundings in images and influences the objects' meaning. Contextual semantic information has been successfully applied in natural language processing tasks and is equally essential for visual scene understanding [7]. While humans intuitively incorporate contextual information when perceiving their environment, visual context reasoning has proven to be a difficult perception problem [6]. Semantic context information can be obtained around objects of interest but few works exist that focus on the representation of context priors. Additionally, extending this contextual information across multiple images domains become apparent, i.e. common environments of objects. The contributions are as follows:

1. Improve existing methods for obtaining explicit context representations and use statistical measures for automated filtering of irrelevant samples from datasets
2. A proposed novel baseline, mixture models and supervised CNNs are employed for domain prediction
3. Apply the entire procedure on public aerial data, predicting domains *apron*, *runway*, *sky* and *other*

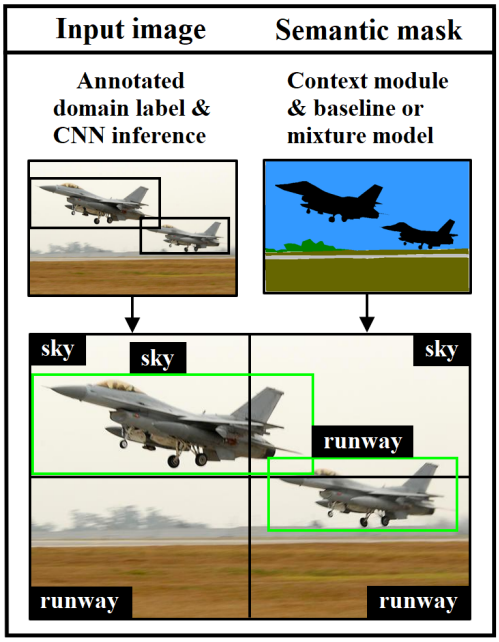

Figure 1: Illustration of the domain prediction process, compatible with multimodal input data.

2. RELATED WORKS

Image semantics are commonly obtained on a micro-level with pixel-wise classification or semantic segmentation. Originally co-occurrence models such as conditional random fields were used, which suffer from only capturing pair-wise relations, to stay tractable, although many context-relationships require richer representations [2]. Recent CNNs [4] can output a complete semantically segmented mask. We argue, that segmented output masks are lower-level in their context expression and are less representative of image content than multi-labels or domains. The concepts of semantic context and visual domains have been treated separately in CV, although the latter can be understood as an extension of the former, collecting context cues across multiple images. Domain adaptation (DA) is commonly used to transfer trained models between two domains from synthetic training data to real test data [3]. In this work domains stem from characteristic (super-) classes

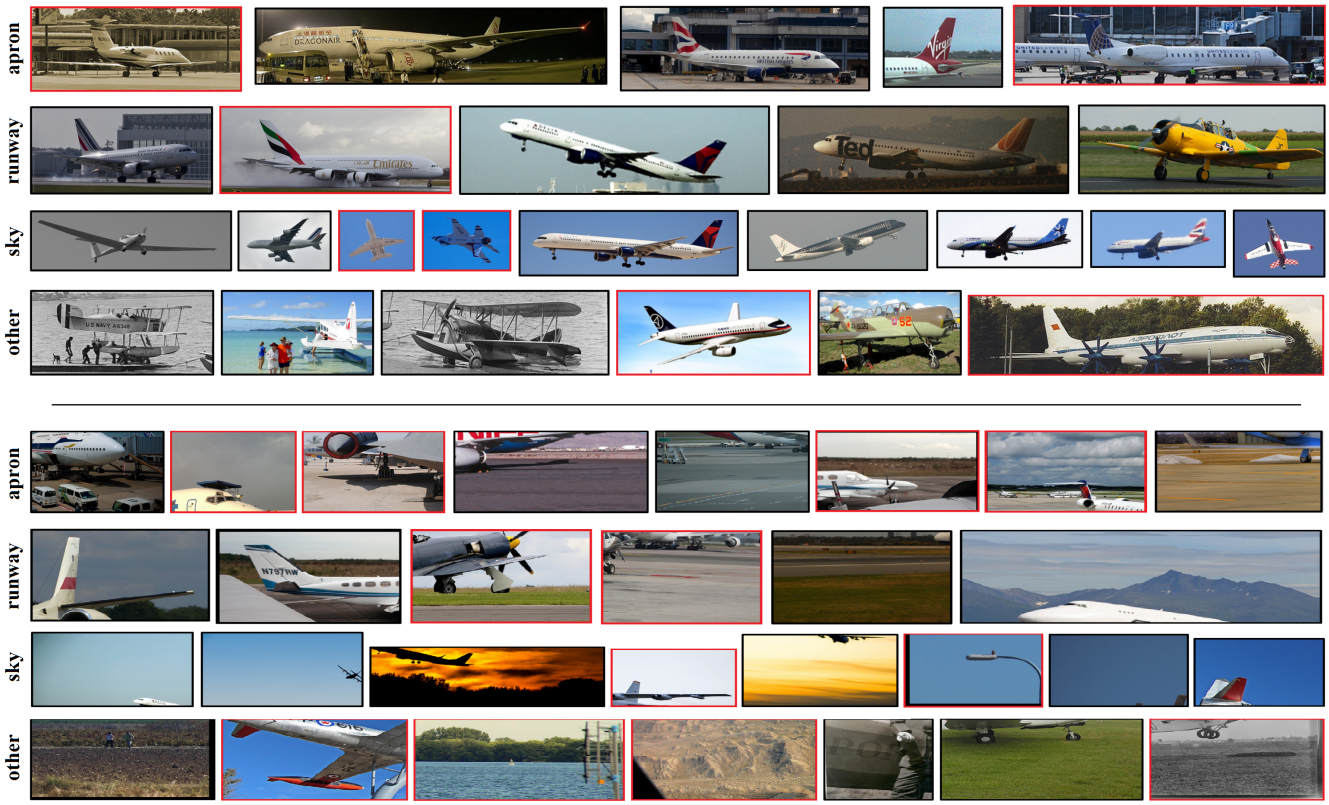

Figure 2: A randomly drawn selection of instances (top) and quadrants (bottom) from *SemanticAircraft*. Samples that were wrongly classified by any of the three algorithms are highlighted in red.

and are specified by visual features surrounding relevant objects, airplanes in this case. Sikirić et al. [9] predicted traffic scenes, a task similar to domain inference.

3. SEMANTIC CONTEXT FOR DATA HARVESTING

Three public datasets were chosen that include semantic masks of airplanes: ADE20K-SceneParsing [11], MS COCO-Stuff [1] and PASCAL-Context [8]. Besides *apron*, *runway* and *sky*, the domain *other* was created to hold out-of-context image-patches, which are prevalent in public datasets. For obtaining semantic context, an improvement on the label frequency outlined in [10] was used. Context is extracted across whole images, image-quadrants, instances and instance-quadrants and further averaged across all three datasets. Results show, that airplane-images from all three datasets are dominated by sky-like context classes, especially in the upper image quadrants. Datasets were merged and classes mapped to form *SemanticAircraft* (for example images see Figure 2). It was decided to proceed with two variations: 30%-increased instances and individual image quadrants (for context statistics see Table 1). A number of filters were implemented to improve the dataset coverage, namely to remove indoor patches and those with a high number of void pixels. Context statistics were calcu-

Table 1: Context in images $\mathcal{I}$ image-quadrants $\mathcal{Q}$ and instance bounding-boxes $\mathcal{B}$ of *SemanticAircraft*, pre-filtering. The difference in *void* pixels between upper and lower image halves is significant. This is due to the fact that scenery images are naturally more cluttered in the bottom half, with many buildings, pavement and plant variations.

Label	$\mathcal{I}$	$\mathcal{Q}_1$	$\mathcal{Q}_3$	$\mathcal{B}$
sky	53.2	**74.5**	30.7	50.7
pavement	15.9	2.6	**30.2**	12.7
soil	6.9	1.6	**12.4**	5.5
void	5.4	3.9	**6.7**	4.3
building	4.2	4.5	3.9	**9.8**
plant	3.8	4.0	3.6	**5.3**
indoor	2.9	**3.7**	2.0	3.5
elevation	2.5	2.5	2.5	**3.3**
waterbody	1.8	0.8	**2.8**	1.4
object	1.3	0.8	**1.9**	1.6
person	1.2	0.7	**1.8**	1.1
vehicle	1	0.3	**1.7**	0.8

Table 2: Domain prediction results (recall) on instances and quadrants of *SemanticAircraft*. The left/right sub-columns show results when including/excluding *other* samples.

	Instances		**Quadrants**	
Baseline	58.8 ± 1.5	79.6 ± 1.1	63.9 ± 1.7	$\mathbf{79.9 \pm 0.6}$
VBGMM	58.6 ± 4.8	71.2 ± 6	53.9 ± 2.9	63.7 ± 8.3
ResNet18	$\mathbf{71.6 \pm 1.5}$	$\mathbf{85.4 \pm 1.1}$	$\mathbf{69.2 \pm 1.3}$	77.8 ± 0.6

lated and show similar distributions when averaged across all instances/quadrants: around 58% *sky* context, 17% *pavement*, 6% *building* etc. Correlation matrices show negative occurence-relationship between sky and all other classes, plant-soil and vehicle-pavement pairs are most common.

4. AERIAL DOMAIN PREDICTION

Figure 1 provides a simple overview of the domain inference component. All 3854 instances and 13265 quadrants were annotated in a one-hot fashion. RGB input images and the domain label were then used to train and tune a large ensemble of CNNs. Initial overfitting of the models was alleviated by reducing both model-size (ResNet18 [5]) and batch-size and adding a drop-out layer. For the baseline and mixture models the context statistics themselves were used. The baseline model, an algorithm that assigns domains following the *SemanticAircraft* class-ontology, was simple to implement although parametrization requires domain-knowledge. The context statistics were interpreted as feature vectors and various unsupervised ML models were employed for context clustering, using the silhouette coefficient to evaluate model performance. Gaussian mixture models (MM) outperformed simple K-Means and a variational Bayesian MM was lastly chosen.

5. RESULTS

Model hyperparameters were optimized employing a grid-search approach. Optimized models were compared against each other in a classification sense, using avg. recall. The results, with and without *other*-samples, can be observed in Table 2. As expected, the supervised CNN performs best yielding the highest accuracy. The baseline performs reasonably well and its deterministic nature and simplicity mean quick reproducibility, although the highly-parametric nature makes it somewhat tedious to tune and extend.

Lastly, while the mixture models fall short in classification strength, they provide insight into the context structure with observed clusters not always correlating to prescribed domains. Mixture models thus yield a deeper understanding into the observed (aerial) scenes, unrestrained from a set number of domains, a research direction that could be further explored in the future.

6. SUMMARY

In this work visual semantic context was used to improve the data aggregation procedure and classify domains of airplanes, although the entire procedure can be applied to any desirable domain requiring either (i) annotation of domain labels for the CNN, or (ii) semantic masks for the baseline and mixture model. Visual semantic context was successfully extracted and applied to not only improve dataset characteristics but also provide a higher-level understanding of aerial image scenes using the concept of domains.

[1] H. Caesar, J. Uijlings, and V. Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[2] M. J. Choi, A. Torralba, and A. S. Willsky. Context models and out-of-context objects. *Pattern Recognition Letters*, 33(7), 2012.

[3] G. Csurka. Domain adaptation for visual applications: A comprehensive survey. *arXiv preprint arXiv:1702.05374*, 2017.

[4] J. Fu, J. Liu, Y. Wang, Y. Li, Y. Bao, J. Tang, and H. Lu. Adaptive context network for scene parsing. In *Proceedings of the IEEE international conference on computer vision*, 2019.

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[6] D. Liu, M. Bober, and J. Kittler. Visual semantic information pursuit: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[7] O. Marques, E. Barenholtz, and V. Charvillat. Context modeling in computer vision: techniques, implications, and applications. *Multimedia Tools and Applications*, 51(1), 2011.

[8] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[9] I. Sikirić, K. Brkić, P. Bevandić, I. Krešo, J. Krapac, and S. Šegvić. Traffic scene classification on a representation budget. *IEEE Transactions on Intelligent Transportation Systems*, 2019.

[10] Y. Zhang, P. David, and B. Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[11] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.