Representing Indoor Scenes as a Sparse Composition of Feature Segments

Camila Laranjeira[1], Erickson R. Nascimento
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
Email: {camilalaranjeira, erickson}@dcc.ufmg.br

***Abstract*—Researchers in the fields of Computer Vision and Pattern Recognition have been trying to tackle the problem of scene recognition for many years. Several approaches rely on the assumption that object-level information can be highly discriminatory, which has been extensively validated in the literature. We propose an approach that merges sparse semantic segmentation features with object features, composing a sparse representation of feature segments, as an attempt to represent the composition of objects of a given scene. Our premise is that by adding sparsity constraints to a semantic segmentation feature, we represent a small amount of well chosen objects or parts of objects. We expect this will add robustness to the final feature, since it will recognize a given scene by its most distinctive segments, thus increasing the generalization power of the representation. According to our results, the methodology seems promising, but it is strongly affected by the poor performance of segmentation features on classes containing small objects.**

I. Introduction

Scene recognition is one of the main challenges in both fields of Computer Vision and Pattern Recognition, and it is considered one of the most difficult classification tasks. As defined by [1], a scene consists in places where humans can act within or navigate. Therefore the capacity to distinguish between the vast amount of existing classes is highly relevant for most applications that intend to operate in the real world. And although there are approaches that surpass human level accuracy [2], it is still regarded as an open challenge.

Several approaches to solve the problem of scene recognition propose the use of Convolutional Neural Network (CNN) features, since they have shown an outstanding performance for a myriad of problems, such as face detection and recognition [3], image super resolution [4], semantic image segmentation [5], to name a few. However, due to the large diversity of scene images, the process of automatic feature learning is much more challenging, requiring large-scale datasets in order to achieve reasonable accuracy [6]. Throughout the years, researchers have found that object-level information can be highly beneficial in the context of scene understanding, being a valuable addition to CNN features, specially for indoor scenes [2], [7]. The rationale is that besides the constituent objects, the image of a room (i.e., an indoor scene) is similar to every scene and it is hard to distinguish among them.

Our approach is based on the assumption that characterizing a small amount of well chosen objects produces a robust representation, since it will rely only on the most distinctive objects of a given scene. To achieve the desired effect, we are leveraging sparse segmentation features along with object features to compose a sparse representation of feature segments. In other words, instead of modeling a dense representation of the entire scene, we add sparsity constraints to the semantic segmentation latent representation, thus selecting only a small amount of segments or parts of segments. Afterwards, we merge this output with object features, in order to characterize only the selected segments. We expect this proposal to produce a sparse composition of objects, in a way that is both discriminatory between scene classes, and robust to either perturbations in the input image or high intra-class variability.

In this work we assess the quality of both features that compose our model (i.e., semantic segmentation and object characterization), and propose a naive approach that merges both features, in order to evaluate how promising this proposal is. To do so, we chose the dataset entitled MIT67 [8], a known benchmark for indoor scene recognition. After performing a per-class assessment on the accuracy of our method, its strengths and weaknesses were highlighted, allowing us to better plan the next steps of our work, in order to propose a more sophisticated methodology.

II. Related Work

Since CNNs started to become a trend, it was expected that researchers would attempt to apply this type of approach to the problem of scene recognition. However, it was only regarded as a promising approach once a large-scale dataset was proposed [9]. The dataset, entitled Places, was used to feed a CNN based approach, which was already very successful for other categorization problems (e.g., objects). However, authors noticed that scenes can vary a lot more, thus the automatic feature learning becomes much more challenging.

The exploitation of object-level information to recognize scenes has been vastly researched for many years. For instance, the work of [10] proposes to represent a scene by combining information from several pre-trained object detectors. A very straightforward approach to represent a scene as a composition

[1]This work is supported by grants from CNPq, CAPES and FAPEMIG. CNPq under Procs. 132779/2016-1 and 456166/2014-9; and FAPEMIG under Procs. APQ-00783-14 and APQ-03445-16, and FAPEMIG-PRONEX-MASWeb, Models, Algorithms and Systems for the Web APQ-01400-14.

of objects, which showed good performance at the time. A more recent approach is the work of [11], which uses a technique for object proposals to select regions of the image that most likely contain objects, along with Long Short-Term Memory (LSTM) units [12] to model a context-aware representation. By incorporating LSTM units to their architecture, the authors were able to model relationships among objects in an end-to-end manner. They also tested the benefits of using object proposals instead of random boxes, in order to validate their premise of capturing knowledge about objects.

Recently, a previous work of ours [7] also showed that local information can be highly beneficial to represent a scene. We showed that by combining scene-centric and object-centric features from different scales, they could outperform previous methods on most benchmark datasets.

Another way to convey object-level information is to use semantic segmentation features. The work of [13] tries to tackle both problems of scene classification and semantic segmentation, showing that they can contribute to the improvement of each other. In other words, segmentation features helped the classification model to achieve state-of-the-art results, while class labels allowed to refine the output of the semantic segmentation.

Compared to the aforementioned approaches, our method not only aims at finding a robust representation for scenes, it also allows for a better understanding of what are the most distinctive parts of a scene. Since we encode information derived from segmentation features with sparsity constraints, it relates to the original image as a selection of best suitable segments for the task of recognition. This information is highly valuable not only to add transparency to the model, but also to guide future approaches that intend to leverage local information.

III. Methodology

Our methodology leverages two types of features: semantic segmentation, and object features. We propose to combine them in order to build a sparse composition of object features characterizing the most distinctive segments of a scene.

Semantic Segmentation Feature

Firstly, we chose to use the pre-trained model proposed in the work of [5], entitled SegNet, illustrated by figure 1. The authors trained it separately in two datasets, CamVid [14] for road scene segmentation and SUN RGB-D [15] for indoor scene segmentation, releasing both models to the public. According to the authors, the problem of indoor scene segmentation is a lot harder, since they vary a lot more than road scenes in shape, size and pose. Additionally, there are frequent partial occlusions, and scenes may contain several small objects which is more challenging to most semantic segmentation approaches. It is important to highlight that even the current state of the art methodology for indoor scene segmentation claims to be a long way from what is expected from such models, which means that any approach relying on those features will be affected by its flaws in performance.

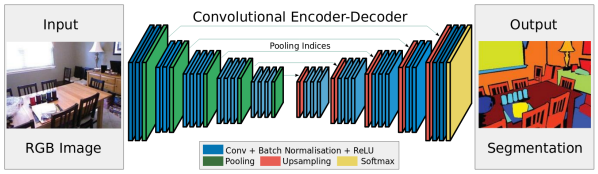

Fig. 1: SegNet architecture [5]. The architecture of the encoder is identical to the VGG16 network.

The rationale behind using segmentation features, comes from the idea of context it conveys. Since it is capable of roughly reproducing the ground truth of semantic segmentation, its latent features must follow a certain structure that separates objects belonging to different classes, and at the same time aggregates regions that belong to the same object. This is very rich information, considering we intend to build a composition of objects as a final feature. In our approach, the semantic segmentation portion of our pipeline is the feature output by the encoder of SegNet.

Object Characterization Feature

Object characterization is a problem vastly studied in the literature in the form of image classification, achieving remarkable results using deep models. Specifically, VGG16 [16] trained on the large-scale object dataset ImageNet [17], became a base model for several methods, including SegNet itself, which uses an identical architecture as its encoder. Therefore, for the object characterization part of our proposal, the output of the first fully connected layer of VGG16 trained on ImageNet was chosen, since it conveys high level semantic information of objects.

Proposed Architecture

We propose a naive approach of putting those two features together, in order to evaluate how promising our premise is. The proposal is illustrated in figure 2. We refer to this architecture as two-stream, one stream of semantic segmentation, and the other for object characterization. Notice that all layers before the merge layer are identical, resembling a siamese network [18]. However, different from a siamese, they do not share weights, acting instead as complementary features. The weights of both streams are frozen until the last convolutional layer, which means only the fully connected layers at the end will adjust its weights for our purpose. The previous layers will function solely as feature extractors.

The pipeline of the proposed architecture functions as follows: firstly, each stream of feature extraction receives the same input: a RGB image of an indoor scene. The only preprocessing step required is resizing the input, in order to fit the needs of the pretrained model, which is $[224, 224, 3]$ for both extractors. Then, during training we adjust the weights of both fully connected layers at the top. Notice that the fully connected layer at the top of the semantic segmentation stream has a sparsity constraint, in the form of a $L1$ activity

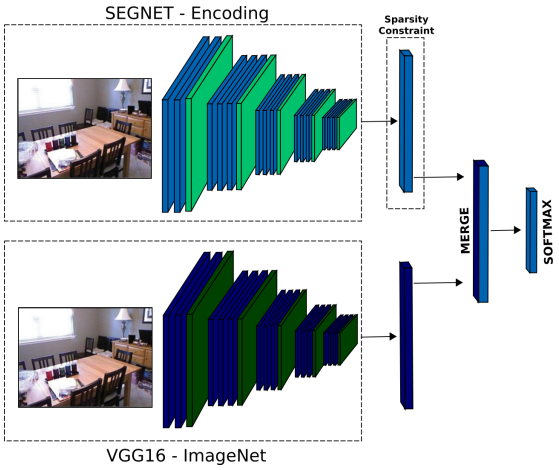

Fig. 2: Proposed architecture. Both SegNet (encoding) and VGG16 share the same architecture, and the final feature merges both of them.

regularizer, which will limit the amount of nonzero values in its output. The goal of the sparsity constraint is to achieve the desired effect of selecting a very small amount of segments or parts of segments. Since we are optimizing the whole model for classification, the constraint will try to find the best choice of segments that increases classification accuracy. The final loss function, optimizing for both categorical cross-entropy and $L1$ regularizer can be represented as follows:

$$[-y_k^{(i)} \log({y'_k}^{(i)}) - (1 - y_k^{(i)}) \log(1 - ({y'_k}^{(i)}))] + \lambda |FC6_s|_1, \quad (1)$$

with $|FC6_s|_1$ representing the $L1$ activity regularizer over the semantic segmentation dense feature, with $\lambda = 1e-5$, empirically set. $y_k^{(i)}$ and ${y'_k}^{(i)}$ are respectively the ground truth class of the scene, and the predicted class.

The second last layer is responsible for merging both features. We chose an element-wise product of both vectors, such that the sparse segmentation feature will function as a mask to the object features. In other words, all segments that were assigned zero values, will cancel the respective object feature. This means that the final feature will only attempt to characterize the selected segments, leading to what we call a sparse composition of feature segments.

IV. Experiments

To test our proposed methodology we chose the dataset MIT67 [8], a known benchmark for scene recognition containing 67 classes of indoor scenes. We split this dataset for training and testing, selecting a total of 5360 images to train our model, uniformly distributed throughout classes, and 1340 for testing. To test for robustness we also created two corrupted test sets, following the protocol of [7], which adds noise and occlusion to the test images. An illustration of such perturbations is showed in figure 3. A region of the image was randomly selected to be corrupted, with a fixed window of size $[\frac{w}{2}, \frac{h}{2}]$, with w and h as the width and height of the original image. Black squares were added for occlusion, while salt and pepper was used to produce the noise.

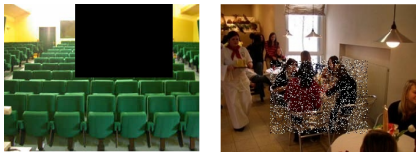

Fig. 3: Examples of corrupted images. Occlusion (left) and Noise (right)

TABLE I: Average accuracy of each feature on MIT67, a benchmark dataset for scene recognition.

	Segnet_{enc}	VGG16-ImageNet	Ours	Nascimento et al.
MIT67	42.00%	59.38%	52.33%	87.22%
MIT67 Noise	33.40%	55.82%	48.19%	82.74%
MIT67 Occlusion	32.28%	51.27%	48.27%	84.76%

We tested both features that compose our methodology, in order to assess their individual quality and compare to the performance of our proposed feature. The first feature was the latent variables of SegNet. A feature of dimensions $[7, 7, 512]$ from its last layer was extracted for all training and test images, which fed a linear SVM model for classification. The penalty value C of the SVM was set to $1e-2$ after performing a grid search for parameter optimization. We also tested the feature output by VGG16 trained on ImageNet. Similarly, features from its last layer, $FC7$ were extracted for training and testing and fed a linear SVM for classification. Coincidentally, the optimization of the penalty value C led to the same result $1e-2$. Finally, we tested our merged feature, proposed on this paper, using a Softmax layer as a classifier, since our model was already trained end-to-end on the target dataset.

Preliminary Results. Table I shows the results for all three features we evaluated, plus the methodology of Nascimento et al. [7], representing the current state of the art. As noticeable, none of the features are competitive with the state of the art. It is worth highlighting that neither the segmentation feature nor VGG16 trained on ImageNet were built for scene understanding, however we intended to evaluate their individual quality relative to our proposed method. Our method did not show a satisfying performance, reaching only 52.33% of average accuracy. On the other hand, our feature shows a higher level of robustness when perturbations are added to the input image, an inherent characteristic of sparse representations.

In order to understand the poor performance of our method, and plan our future work, we evaluated its performance on each class of MIT67. The results are shown in figure 4. Even though our training set was balanced for all classes, this analysis shows the model performs very differently for each class, reaching a maximum of 94% accuracy on the class *greenhouse*, and a minimum of 25% for the class bakery. Judging by this result, our model performs poorly on classes that contain a large amount of small objects (e.g. bakery, deli, toystore), while showing remarkable performance for classes such as greenhouse, cloister and bowling, which is composed

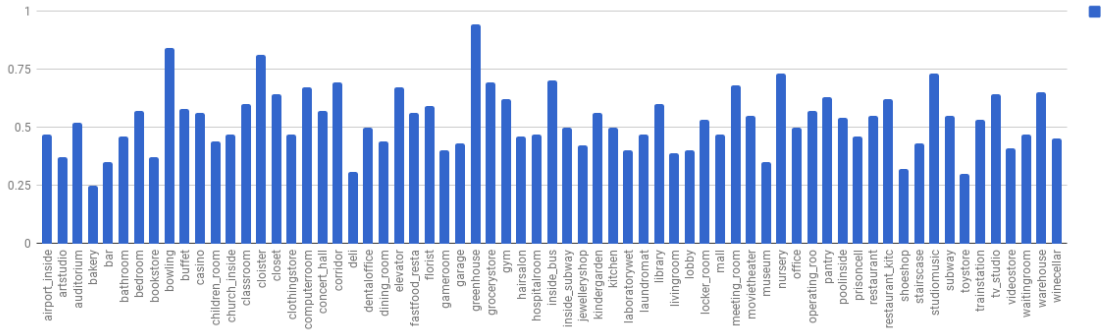

Fig. 4: Detailed performance assessment on per-class average accuracy.

mostly of large segments. Our guess is that the segmentation feature performs poorly for small objects, thus compromising the performance of our final feature. Figure 5 is an evidence that supports our guess, comparing the segmentation output for a sample of class greenhouse, and a sample of bakery. Nevertheless any definite conclusions requires further testing.

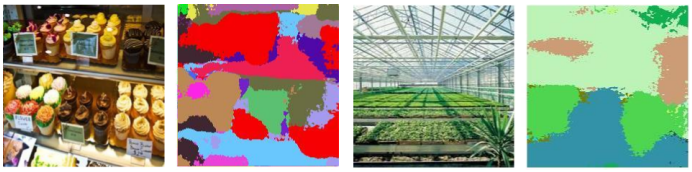

Fig. 5: Performance of SegNet segmentation on different classes of MIT67. Left: bakery, Right: greenhouse.

V. Conclusions

We proposed a two-stream architecture that leverages semantic segmentation features and object characterization features, combining them in order to build a sparse composition of feature segments. We added sparsity constraints to the segmentation feature as an attempt to select the most distinctive segments of a scene, thus building a robust feature. The average accuracy of our model performs poorly compared to the state of the art. However, a detailed assessment of per class accuracy showed that the performance of our method might be correlated to the size of the objects present in the scene. This indicates that poor quality on the segmentation feature can compromise our model. When testing for robustness, the performance of our model showed little decrease in accuracy, an inherent characteristic of sparse representations. We strongly believe that characterizing distinctive segments of a scene can provide a robust feature, thus we intend to propose a more sophisticated methodology to exploit this proposition.

References

[1] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*. IEEE, 2010, pp. 3485–3492.

[2] Z. Wang, L. Wang, Y. Wang, B. Zhang, and Y. Qiao, "Weakly supervised patchnets: Describing and aggregating local patches for scene recognition," *IEEE Transactions on Image Processing*, 2017.

[3] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "From facial parts responses to face detection: A deep learning approach," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3676–3684.

[4] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *European Conference on Computer Vision*. Springer, 2014, pp. 184–199.

[5] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *arXiv preprint arXiv:1511.00561*, 2015.

[6] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Advances in neural information processing systems*, 2014, pp. 487–495.

[7] G. Nascimento, C. Laranjeira, V. Braz, A. Lacerda, and E. R. Nascimento, "A robust indoor scene recognition method based on sparse representation," in *22nd Iberoamerican Congress on Pattern Recognition. CIARP*. Valparaiso, CL: Springer International Publishing, 2017, to appear.

[8] A. Quattoni and A. Torralba, "Recognizing indoor scenes," vol. 0, pp. 413–420, 2009.

[9] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 487–495.

[10] L.-J. Li, H. Su, L. Fei-Fei, and E. P. Xing, "Object bank: A high-level image representation for scene classification & semantic feature sparsification," in *Advances in neural information processing systems*, 2010, pp. 1378–1386.

[11] S. A. Javed and A. K. Nelakanti, "Object-level context modeling for scene classification with context-cnn," *arXiv preprint arXiv:1705.04358*, 2017.

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," vol. 9, pp. 1735–80, 12 1997.

[13] Y. Liao, S. Kodagoda, Y. Wang, L. Shi, and Y. Liu, "Understand scene categories by objects: A semantic regularized scene classifier using convolutional neural networks," in *Robotics and Automation (ICRA), 2016 IEEE International Conference on*. IEEE, 2016, pp. 2318–2325.

[14] G. J. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," *Pattern Recognition Letters*, vol. 30, no. 2, pp. 88–97, 2009.

[15] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 567–576.

[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 248–255.

[18] J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Säckinger, and R. Shah, "Signature verification using a "siamese" time delay neural network," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 7, no. 04, pp. 669–688, 1993.