On Modeling Context from Objects with a Long Short-Term Memory for Indoor Scene Recognition

Camila Laranjeira, Anisio Lacerda, and Erickson R. Nascimento
Universidade Federal de Minas Gerais (UFMG), Brazil
Email: {camilalaranjeira, anisio, erickson}@dcc.ufmg.br

***Abstract*—Recognizing indoor scenes is still regarded an open challenge on the Computer Vision field. Indoor scenes can be well represented by their composing objects, which can vary in angle, appearance, besides often being partially occluded. Even though Convolutional Neural Networks are remarkable for image-related problems, the top performances on indoor scenes are from approaches modeling the intricate relationship of objects. Knowing that Recurrent Neural Networks were designed to model structure from a given sequence, we propose representing an image as a sequence of object-level information in order to feed a bidirectional Long Short-Term Memory network trained for scene classification. We perform a Many-to-Many training approach, such that each element outputs a scene prediction, allowing us to use each prediction to boost recognition. Our method outperforms RNN-based approaches on MIT67, an entirely indoor dataset, while also improved over the most successful methods through an ensemble of classifiers.**

I. Introduction

The ability to recognize the environment around us might seem effortless for humans, but research on scene recognition shows otherwise for computers. According to [1], a scene is defined as any place a human being can act within or to which one could navigate, ranging from house rooms to islands, stadiums, cathedrals, among many others. Different from other classification tasks, such as recognizing objects or faces, scenes can be quite difficult. Besides the usual challenges such as lighting, angle of image acquisition, occlusion, to name a few, the image of a scene can be abundant in highly variable local information. This is specially true for indoor scenes, since their global structure can be very ambiguous among classes (e.g., house rooms), while local information from objects allows to distinguish between them. Each object can present itself in a variety of manners, and their disposition in the environment can also be very diverse, putting indoor scene recognition as an even harder challenge, requiring approaches specially tailored for the task.

With the rise of Convolutional Neural Networks (CNN) [2] as the most promising approach for classification on images, many attempts have been made to apply such networks to tackle the issue of scene recognition. With the introduction of a large scale scene-centric dataset [3] the expectations were even higher for a CNN to be the best solution. However, even though the results were promising, future works gained greater prominence by taking advantage of high level semantic knowledge, usually conveying object-level information and their intricate relationship [4]–[6].

More recently we witnessed the surge of Recurrent Neural Networks (RNN) and its variations. The ability to correlate information from parts of a sequence was designed to solve a whole new class of problems, mainly the ones that presented temporal dependencies. Text [7], audio [8], and time sequences such as stock market prices [9] were the primary types of data in which an RNN was applied. And, as expected, they benefited a lot from the behaviour of that type of model. However, any data that can be divided into interdependent parts is eligible to exploit the advantages of recurrent models. Hence, in this paper, we propose to classify scenes using a methodology based on a RNN. Specifically, we exploit the advantages of a Bidirectional Long Short-Term Memory (BiLSTM), an advanced recurrent unit [10] that provides predictions of higher quality compared to its unidirectional counterpart [11]. Our assumption is that scenes can be well represented by their composition of objects, therefore we leverage object-level information to compose the input sequence.

The main idea is to use a Region Proposal approach to select Regions of Interest (ROI) from the image, belonging to object parts. Then, extract highly semantic features from each ROI, composing a collection of object features. By training a BiLSTM to perform classification over the sequence of features, it will learn the underlying structure of object parts, building a semantically meaningful representation that is able to distinguish between classes. We perform a Many-to-Many (M2M) training procedure, seeing that it will produce a prediction for each object part relative to the remaining context, allowing us to boost recognition performance by considering how each part relates to the remaining image.

We are also interested in knowing if our proposal can improve over the most successful methods on the literature by pairing our work to a few of them in an ensemble of classifiers. That will allow us to analyze the performance of our method relative to each approach comparing it to a joint strategy of methodologies. As showed by our results, the proposed ensemble boosts classifcation performance for all approaches. We evaluate our method on three datasets, Scene15 [12], MIT67 [13], and SUN397 [1], each presenting different types of scenes and levels of difficulty.

II. Related Work

Scene Recognition has been an active field for over a couple of decades. Early proposals were inspired by the field of image retrieval [14], [15], while others borrowed knowledge

from human psychology as their main influence [16]. Such approaches relied solely on low-level features from the image, composing a global representation of the scene, which was found by the authors to perform poorly on indoor scenes, since they neglect the importance of local information for recognition. Other approaches proposed mid-level representations, bringing up the concept of Bag-of-Features composed by local information [12]. One of the highlights from the literature is the work of [17], a Spatial Pyramid Matching approach that produced multiple Bag-of-Features from three scales. Even so, indoor scenes remained as a greater challenge for such works.

The popularization of deep Convolutional Neural Networks raised the bar on average performance for scene recognition. Early CNN approaches were directed to the problem of object recognition, specially after the release of a large-scale object-centric dataset, ImageNet [18]. Still, researchers attempted to apply such models to the problem of scene recognition, a more distant domain, reaching promising results [19]. After a large-scale scene centric dataset was released, entitled Places [20], there was a lot of investment in CNN approaches as the solution to the problem of scene recognition. Models pre-trained on Places (Places-CNNs) showed great improvement over the state of the art. And although the debut of Places was groundbreaking, researchers were still providing solutions tailored to the specific task of recognizing indoor environments.

Combining local information of a given scene was shown very promising in the literature, specially for indoor scenes. [5] proposed a joint strategy of object features for local scales and scene-level features for the entire image, outperforming Places-CNNs. Nascimento et al. [6] followed the same premise on combining scene-level and object-level information on different scales, only now proposing a more robust dictionary-based representation with sparse coding of features. There is also the work of Wang et al. [4], proposing an architecture entitled PatchNet to model the appearance of local patches, and an encoding approach called VSAD as a global descriptor. All three of those methods share the premise of composing a representation with rich local information, such that indoor scenes would not represent a weakness for their approaches.

A. Recurrent Neural Networks

When RNNs gained popularity for image-related problems, it was common to see them applied to inherently sequential data. But soon enough researchers started applying recurrent methodologies to challenges such as multi-label image classification [21], or scene labeling [22], to name a few. For images, modeling the structure of parts is equivalent to learning contextual dependencies, which is highly valuable for problems requiring correlation of local information.

With the success of composing local information to classify scenes, the power of RNNs to model the structure of parts was considered suitable for the problem of scene recognition, specially for indoor scenes. One reference that stands out is the work of Zuo et al. [23], one of the first reports applying a combination of Convolutional and Recurrent layers to correlate semantic features for scene classification. Their method, entitled C-RNN, adopted a quad-directional RNN to correlate intermediate convolutional features, and was entirely pre-trained on object-centric data, being highly competitive to the state of the art at the time. Extending the work of Zuo et al., in 2016 the authors attempted a hierarchical approach [24], entitled C-HRNN. They followed similar steps as their previous attempt, but with a more complex multi-directional RNN along with a hierarchical RNN to correlate information from different scales. It is also worth mentioning the work of Javed and Nelakanti [25], a recent proposal for scene recognition that assumes object features as an ideal source of information to recognize scenes, reinforcing our premise. Their work relied on a method of region proposal to select ROI from the image, and composed a sequence of object features with all ROI from an image. On their experiments the number of ROI was fixed to 10, arguing that it was sufficient as a proof of concept to validate the methodology. Since recurrent models allow inputs of variable size, fixing the number of ROI omits an important aspect of the scene. The amount of object information present in each scene by itself conveys relevant knowledge regarding its category, i.e., some classes can be typically more crowded than others. And although our approach resembles the work of Javed and Nelakanti on the step of sequence composition, our model supports variable length sequences. Additionally, there is little effort towards exploiting the multiple outputs provided by a recurrent approach, which we intend to explore in this work.

III. Methodology

This section describes in details the proposed methodology, as illustrated by Figure 1. In order to build an approach for scene recognition based on a recurrent model, we represent an image as a sequence of elements. Thus, step (a) (refer to Figure 1) of the methodology is dedicated to dividing the image of a scene into parts of scene objects, ordered by significant criteria in the interest of composing a sequence. Then, since our premise is based on representing an image by its composition of objects, for each part we extract high-level object features from a deep CNN, constituting step (b) of our method, composing a sequence $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ of features $\mathbf{x}_t \in \mathbb{R}^d$ where d is the dimension of our chosen deep feature, as it will be detailed later. The composed sequence is used as input to our recurrent model, step (c). We propose a Many-to-Many training approach for a Bidirectional Long Short-Term Memory such that each sequence element $\mathbf{x}_t$ produces an output y_t based on the current input along with accumulated context of the remaining parts. Since we only have scene-level labels, all outputs are an attempt at predicting the category y of the input scene.

At test time we add steps (d) and (e). The first steps generates a single prediction y'_{our} from the recurrent model through a weighted majority voting. We expect to boost classification performance relative to a vanilla voting approach, since not every part of the scene is equally relevant. The voting weights are based on pre-calculated object weights representing the relevance of each object class for a given scene category.

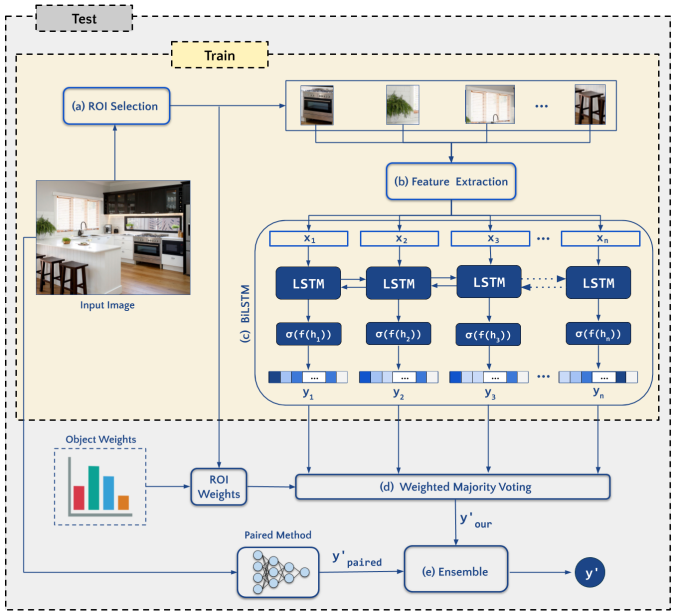

Fig. 1. Overview of our methodology. Steps (a) through (c) constitute the training steps, respectively (a) dividing the image into object parts; (b) extracting high level features from each part; and (c) training a M2M BiLSTM to produce a prediction y_t for each $\mathbf{x}_t$. At test time, step (d) is a weighted majority voting to aggregate predictions, outputting y'_{our}. Finally the ensemble of classifiers, step (e), decides the final prediction between ours and a paired classifier from the literature.

Finally, step (e) is an ensemble of paired classifiers, in which our prediction is paired with a successful approach of scene recognition from the literature that produces an output y'_{paired}. A switch criteria is proposed based on statistical measures over our prediction, such that whenever our inference is determined to be weak, the paired classifier's output is considered as the final prediction y'. This final step aims at improving classification performance over each paired classifier.

A. Composing a Sequence of Object Parts

The goal of our first step is to compose an ordered sequence of ROI from the image, with interdependent object parts. Considering that we do not have available annotations on object labels and bounding boxes for scene images, we chose a well known algorithm for object proposals called Selective Search [26], which yields 99% recall, meaning it selects nearly all object information from the scene. Since the Selective Search algorithm outputs object bounding boxes, it is intuitive to infer that depending on the characteristics of the scene, the number of output regions can vary drastically. For classes such as *deli*, scenes are usually crowded with delicacies up for sale, while categories like *pool inside* present fewer objects other than the pool itself. This is relevant because it means the output sequence based on a region proposal approach has variable length. To the best of our knowledge, there is a couple of references on the topic of scene recognition that exploits such an approach to represent an image as a sequence, and they choose to fix the sequence length despite the aforementioned behavior of region proposal methods [25], [27].

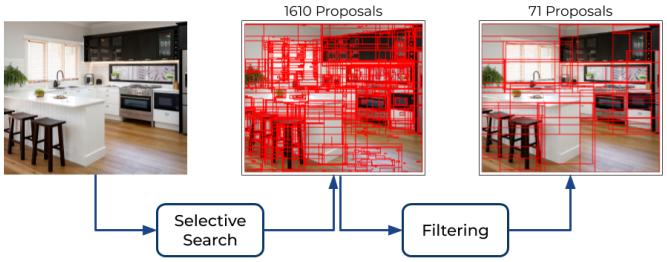

Fig. 2. Step (a) of our methodology, filtering bounding boxes proposed by Selective Search for a feasible model training.

It should also be noted that the number of ROI proposed by Selective Search can reach hundreds or even thousands of bounding boxes, which would be intensely time consuming for a recurrent training due to the difficulty of modeling very long sequences. Therefore, as illustrated in Figure 2, to compose a smaller and more feasible sequence we filter the proposed bounding boxes by their size relative to the entire image. The idea is to define two thresholds t_{lower} and t_{upper} representing the lower and upper percentage limits of patch size. Selective Search provides the size of each segment in pixels, which we call s_{patch}, as an attribute of the output. Thus given the image size s_{img} as the product of its width and height, we allow patches within the following range:

$$s_{img} * t_{lower} < s_{patch} < s_{img} * t_{upper}. \quad (1)$$

The output of Selective Search is decreasingly ordered by the likelihood of a region to contain an object, which we call *objectness*. We maintain the algorithm's order of elements when composing our sequence, meeting the requirement of a consistent order of elements throughout all samples. The final output of this step is a sequence of bounding boxes from the filtered output of Selective Search, decreasingly ordered by objectness.

B. Feature Extraction

After selecting all ROI from the image, the next step is the extraction of highly semantic features from each region. Since our main goal is to input a recurrent model with a sequence of object-level information from the image, the process of feature extraction should convey information of that nature. Deep learning approaches are powerful feature extractors for many applications, and this is specially true for object features. Residual nets were able to take object classification performance even further by adding residual functions to allow training of deeper networks [28]. We exploit the advantages its 50-layer variation, entitled Resnet-50, to serve as feature extractor of our methodology. We perform a forward pass on Resnet-50 pretrained on ImageNet, and extract the last convolutional layer, after average pooling, providing a highly semantic and discriminative object feature of $d = 2,048$ dimensions. After extracting features from each region, the final output of that step is a sequence of object features $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ with $\mathbf{x}_t \in \mathbb{R}^d$, ordered according to the objectness criteria, defined on the previous step.

C. Context Modeling with a BiLSTM

Once the input scene is represented as a sequence $\mathcal{X}$ of features, our goal is to model the image context by correlating all $\mathbf{x}_t \in \mathcal{X}$. We propose to exploit the power of recurrent models to represent the structure of the sequence. Therefore, step (c), presented in Figure 1, consists in training a variation of a Recurrent Neural Networks optimizing it for classification, such that the model will learn the structure of scenes, producing similar intermediate representations for samples from the same category, i.e., the modeled structure will convey semantically meaningful information of the scene. More generally, a recurrent unit is a function of the current input and previous knowledge. Hence it is capable of remembering past information and accumulate knowledge throughout iterations. Although it was designed for data with inherent sequential structure, when applied to images it correlates the given parts just as it would for any other data. As long as the input has structured dependencies between parts, a recurrent approach is capable of modeling it.

On the choice of an RNN variation, gated units tend to be superior than a simple unit, since they are capable of modeling longer sequences due to their ability to avoid the vanishing/exploding gradient problem. As for the difference between the two advanced gated units, i.e.,Gated Recurrent Unit (GRU) [29] and Long Short-Term Memory (LSTM) [10], no significant performance gap was found. Thus, we chose the LSTM variation due to the more extensive literature successfully applying it to different kinds of data. We also exploit the advantages of a bidirectional approach [11], which can accumulate knowledge from more than one direction and produce a better informed inference, as it has already been evidenced by scene recognition approaches from the literature [23], [24].

In practice, a BiLSTM approach means having two LSTM units, each one accumulating knowledge from a different direction, as illustrated in Figure 3. As a result, at every timestep t there is information available from the entire image, the sequence "past" (positive direction) and the "future" (negative direction), which means an output produced at iteration t is a function of the current input and the context of the remaining sequence elements, parts of an image in our case. Based on that, we use a synchronized Many-to-Many (M2M) training procedure, producing a scene classification output y_t for every input $\mathbf{x}_t$. Since each element of our sequence has meaningful semantic information from objects, each prediction can potentially convey information of how such element relates to its context.

As a synchronized M2M procedure, every hidden state $\mathbf{h}_t$ is forwarded through the fully connected layer and activated with a softmax in order to produce one prediction for each input. Likewise, the loss calculation should take into account errors from all timesteps. Since we are optimizing our model to perform classification and we only have scene-level labels, our loss for each timestep t is a Cross-Entropy function between the probability vector y_t and a one-hot encoding representing

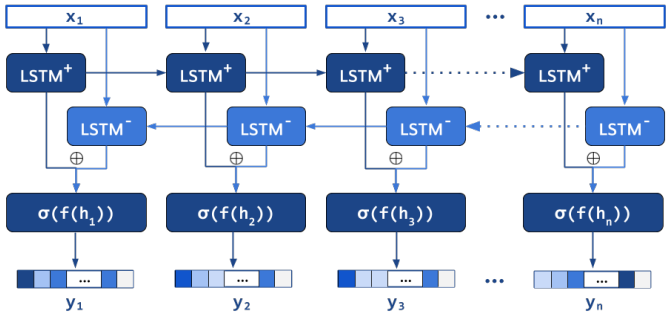

Fig. 3. Representing a Bidirectional Long Short-Term Memory to solve a synchronized Many-to-Many problem. The hidden state output by each LSTM unit is concatenated ($\oplus$) composing a single hidden output for future layers.

the scene category y, as defined by

$$\ell(y_t, y) = -\sum_i (y_t^i \log(y^i) + (1 - y_t^i) \log(1 - y^i)). \quad (2)$$

The final loss is then calculated as an average of every $\mathcal{L}(y_t, y)$ calculated previously, as Equation 3 shows:

$$\mathcal{L}(y', y) = \frac{1}{n} \sum_{t=1}^{n} \ell(y_t, y). \quad (3)$$

D. Weighted Majority Voting

Because our methodology generates n predictions, n being the number of ROI selected from a scene, we still need to output a single prediction to perform inference on test samples. Since a BiLSTM produces an output for each corresponding input relative to the remaining context, each y_t has the potential to predict the correct scene category. However, the prediction can vary throughout timesteps, since any y_t is a function of how the corresponding x_t relates to its context. Thus, we aggregate y_t as a weighted majority voting, taking into account the relative importance of each patch for the image.

To calculate the weights, we use a validation set to build a weight matrix W^{obj} of size n_c x n_o, respectively the number of scene classes on the dataset and the number of all possible objects, for which we considered all $n_o = 1,000$ categories from Imagenet. The goal of this matrix is to contain the relationship between each scene i and each object j. To build W^{obj}, we initialize it with all zeroes, and for each x_t from an image we want to increment cell (i, j) with a weight representing how relevant is the object from that patch to the scene category. We know the scene class i since we work with a validation set, and to find j, we predict the object class from x_t using Resnet-50 predictive ability. Then, we forward x_t through our trained BiLSTM producing a probability vector y_t of size nc, representing the activation strength for each scene category. Finally, we increment cell (i, j) by the probability of class i according to y_t, defined by y_t^i. The rationale is that y_t^i corresponds to the probability of object j belonging to class i. Once W^{obj} is entirely filled by all samples from our set, as a normalization approach we divide the weights from each cell

by the number of patches from the correspondent class used to fill each row of the matrix. This benefits objects that occur more often, which is also an important aspect on the relevance of such object.

At test time, we perform a weighted majority voting between predictions from all timesteps, using matrix W^{obj} to provide the weights. Given an input $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ of features, from each $\mathbf{x}_t$ we predict the object class j, and the corresponding recurrent prediction y_t of size n_c. Let w_t^j represent the j^{th} column of W^{obj} at timestep t. Our weighted prediction will then be defined by Equation 4,

$$\hat{y}_t = y_t \odot w_t^j, \tag{4}$$

where $\odot$ represents the element-wise product of both vectors. Afterwards, the strongest activation from each $\hat{y}_t$ contributes as the vote for class i at iteration t. After all iterations, the majority voting of weighted outputs will provide the final prediction y'_{our}.

E. Ensemble of Classifiers

The prediction y'_{our} from the previous step is sufficient to perform scene recognition, however we are also interested in knowing if our method adds any information over the state of the art. For that purpose, we propose to pair our own approach with methods from the literature, based on a switch criteria that will determine for a given image which of the paired approaches should be considered the final output prediction. We chose successful approaches as paired classifiers since our goal is to see if our reliable predictions can improve over a few of the best approaches in literature. Figure 4 shows an overview of our proposal. It is important to notice that we only apply the switch criteria over our own method, hence our ensemble can be paired with any classifier regardless of their particularities.

Our switch criteria is a random forest trained on a binary problem, with labels $\{0, 1\}$ respectively indicating a correct prediction and a misclassification of our approach. As input, first we build a unidimensional vector p^{max} of maximum activations throughout timesteps. Given a set of outputs $\mathcal{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n\}$ each p_t^{max} corresponds to $\max y_t$ from timestep t. From p_t^{max} we extract statistical measures which then feed the random forest. We perform

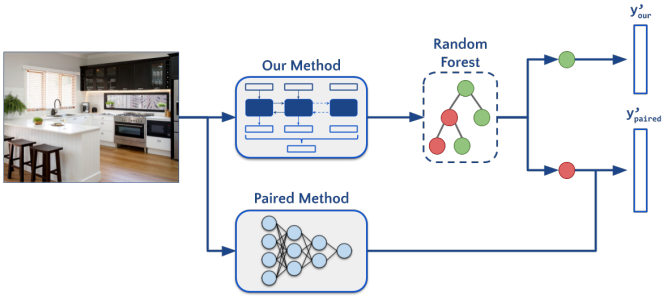

Fig. 4. Overview of ensemble approach. The random forest is a switch criteria to determine prediction reliability. Green circles represent a reliable inference while red circles indicate the paired approach should provide the prediction.

two rounds of training. First, with several empirically chosen metrics to extract the feature importance of each one, and then we choose the best metrics to perform the final round of training that will generate the switch criteria.

IV. Experiments

We tested our approach on three datasets widely known as benchmark for scene recognition, namely Scene15 [12], MIT67 [13], and SUN397 [1]. Scene15 is a small dataset, compared to the MIT67 and SUN397, and it was one of the earlier datasets dedicated to scene recognition. It is composed of 15 classes of indoor and outdoor environments. MIT67 was created due to the need to tackle the specific issue of indoor scenes, and it contains 67 classes of a wide variety of indoor environments. Finally, SUN397 is a dataset of indoor and outdoor enviroments comprising 397 classes. It was motivated by the lack of large-scale scene-centric datasets on the literature, attempting to capture the full variety of scenes.

A. Parameter Settings

For the step of ROI extraction, Selective search mainly requires two parameters: σ and k corresponding respectively to a Gaussian filter parameter and a scale parameter. We use the default values defined as $\sigma = 0.8$ and $k = 300$. Then, to filter larger and smaller patches, we set $t_{lower} = 0.1$ and $t_{upper} = 0.8$ empirically, i.e., patches that account for less than 10% or more than 80% of the image area were discarded. We are aware that such parameters might require optimization, however the intuitive choice drastically decreased the sequence length while maintaining over 90% image coverage.

The input that feeds our BiLSTM has three dimensions: $batch\_size \times seq\_len \times feat\_size$, representing respectively the batch size, sequence length and feature size. The first parameter was fixed to 1 to avoid the need to pad our data, since seq_len varies per sample with the amount of selected ROI. $feat\_size = 2,048$ as defined by the architecture of Resnet-50. There is also a free parameter on the recurrent layer concerning its hidden size. Considering that we tested different architectures and training approaches, h_t was fixed to 512 as proposed by the work of [25]. It is worth reminding that our recurrent layer is bidirectional, which means that although $h_t = 512$, the actual outputs is $2 \times h_t$ since the output of both recurrent units (one for each direction) will be concatenated before feeding the next layer.

As for training settings, we used Adam [30] with its default parameters, except for its initial learning rate that was empirically set as $1e-7$. The train/test split is already defined on the reference of each dataset. However, we needed a validation set in order to generate our weight matrix W^{obj} and to train the switch criteria on the ensemble of classifiers. We created a validation set for each dataset by randomly selecting 15 samples from each class, removing them from training.

B. Ablation Analysis

This subsection unpacks some aspects of our methodology. First, on the choice of recurrent approach, we tested four

configurations shown in Figure 5, varying the training procedure (Many-to-One (M2O) vs Many-to-Many) and output composition. Variations I and II produce a single output from the recurrent hidden states (M2O), with variation I using a unidirectional LSTM and variation II a BiLSTM. Variations III and IV produce multiple outputs (M2M) with a BiLSTM, but they differ on the voting approach, while III uses a simple majority voting, IV is our final proposal with the weighted voting of predictions.

The results for all configurations on Scene15 and MIT67 are presented in Table I. Due to its simple nature, results on Scene15 show slight improvement on performance between configurations. However, it is possible to detect the contribution of each new aspect added by our methodology. Whereas MIT67 allows a more insightful analysis of performance improvement. The greatly improved behavior of a BiLSTM compared to the unidirectional equivalent is consistent with early findings in the literature regarding the benefits of accumulating knowledge from different directions whenever the problem allows it [11]. More importantly, on variation IV, the positive results by weighting the predictions supports our claim that image regions are not equally relevant to inference, and the individual predictions can offer valuable insight to the reasoning process.

For a further understanding of how the weighted majority voting contributes to the prediction process, Figure 6 shows the accuracy performance of our weighted voting relative to a simple voting approach. It is noticeable that for most classes the weights contribute positively, improving accuracy up to 25 percentage points. Figure 6 also highlights the classes that we achieve greatest improvement. They are composed of a large population of inconclusive information, e.g., movietheaters are crowded with chairs, while distinctive objects are underrepresented on the input sequence, e.g., theater curtain. Our matrix of object weights increases the importance of distinctive object parts, and reduces the activation strength for inconclusive patches, serving as a solution for scenarios in which the sequence is dominated by non discriminative parts.

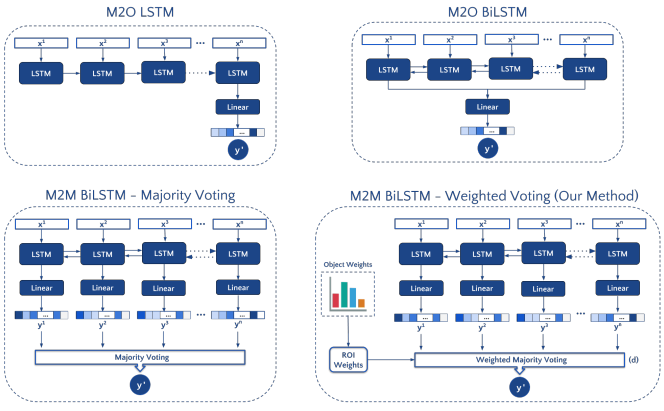

Fig. 5. Variations of recurrent approaches for scene recognition. First, a M2O unidirectional LSTM, followed by variation II, a M2O BiLSTM. Variation III is a M2M BiLSTM with vanilla majority voting, and finally our method adding a weighted majority voting to aggregate predictions.

		M2O		M2M	
		LSTM	BiLSTM	BiLSTM M. V.	BiLSTM W. V.
Scene15	Accuracy	92.00%	93.50%	94.06%	94.29%
	Recall	92.17%	93.50%	94.15%	94.47%
	Precision	92.31%	93.61%	94.29%	94.75%
	F1 Score	92.23%	93.55%	94.19%	94.57%
MIT67	Accuracy	59.66%	72.94%	75.18%	79.52%
	Recall	59.51%	72.85%	75.20%	79.60%
	Precision	47.90%	74.65%	76.09%	80.13%
	F1 Score	53.07%	73.74%	75.64%	79.86%

TABLE I
COMPARING PREDICTION PERFORMANCE ON SCENE15 AND MIT67 OF THE FOLLOWING RECURRENT APPROACHES: UNIDIRECTIONAL LSTM, BILSTM, BILSTM WITH A SIMPLE MAJORITY VOTING (M. V.) AND A BILSTM WITH OUR WEIGHTED VOTING (W. V.).

Next, we experimented with the ensemble of classifiers. Three methods from the literature were chosen to act as paired classifiers for our ensemble. They were chosen either for providing a source code or a trained model. First, a VGG16 [31] pretrained on Places [20] and fine-tuned on each test dataset, a strong baseline proposed by Nascimento et al. [6]. The second method was [5], following the same premise as ours regarding the relevance of objects. And finally [6] one of the best performances on the literature, proposing a sparse coding based methodology. Table II shows our results with the proposed M2M BiLSTM with a weighted voting but without any ensemble, and compares it to each of the paired methods by themselves and as part of our ensemble. We improved the classification accuracy over each method, showing that our methodology contributes positively on the improvement of successful approaches from the literature. We find quite relevant that the best improvement happened on a dataset dedicated to indoor scenes (MIT67), since our work was modeled towards that problem.

C. State-of-the-art Approaches

On this subsection we compare our work to state-of-the-art approaches, presenting works of two different natures: CNN-based and RNN-based. Table III shows the accuracy of all

	Scene15	MIT67	SUN397
M2M BiLSTM Weighted Voting	94.29%	79.52%	54.00%
VGG16 (Places)	93.87%	80.88%	66.90%
Ensemble (VGG16)	**94.40%**	**84.60%**	**68.08%**
Herranz et al., 2016 [5]	95.18%	86.04%	70.17%
Ensemble (Herranz)	**95.96%**	**86.47%**	**71.35%**
Nascimento et al., 2017 [6]	95.73%	87.22%	71.08%
Ensemble (Nascimento)	**96.30%**	**88.25%**	**71.81%**

TABLE II
ACCURACY RESULTS FOR THE ENSEMBLE OF PAIRED CLASSIFIERS. OUR METHOD WAS PAIRED WITH THREE LITERATURE APPROACHES, IMPROVING OVER EACH OF THEM.

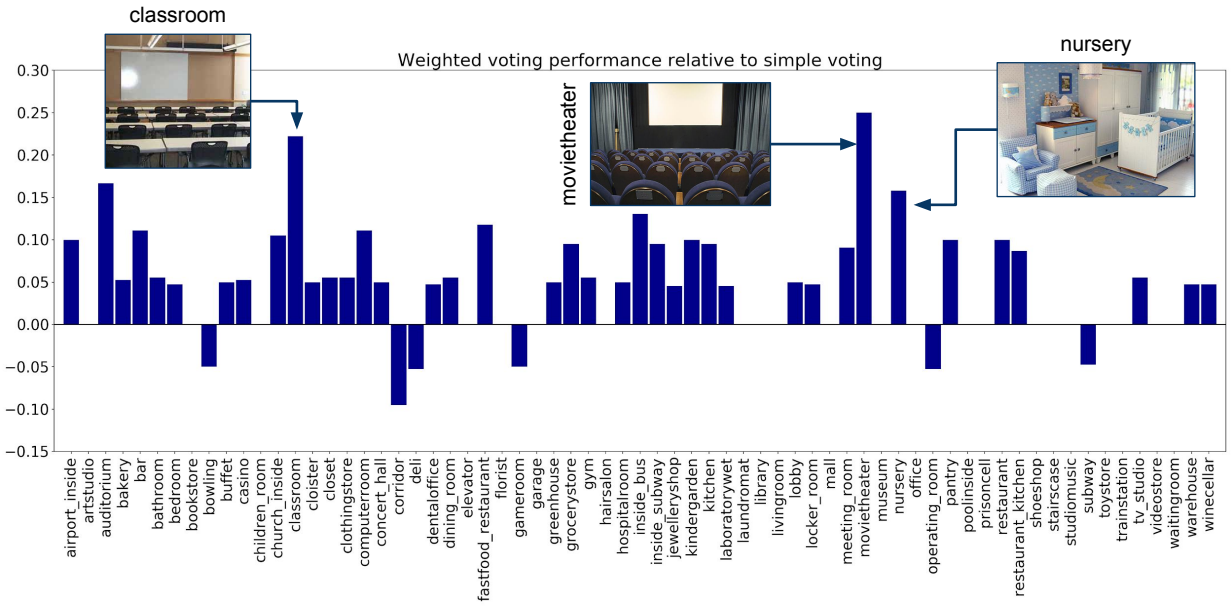

Fig. 6. Weighted voting accuracy performance relative to a simple voting, highlighting classes with the highest gains.

methods including the two possible outputs produced by our proposal: M2M BiLSTM with a weighted majority voting, and the best result with an ensemble of paired classifiers presented on Table II, which for all cases is by pairing with [6].

Our method, even without the ensemble, performs better than any other RNN-based approach on MIT67, which is an entirely indoor dataset. That result is very positive since all approaches rely on the same premise of correlating interdependent image parts. None of the RNN approaches presented results for Scene15, but they did for SUN397, which showed interesting results in comparison to ours. The work of Zuo et al. [24], presented twice at Table III, pretrains its model on two different datasets: ImageNet (ILSVRC), with object-centric samples, and Places, a scene-centric dataset. Our performance is better than its ILSVRC variation for both datasets (MIT67 and SUN397), whereas the scene-centric pretraining beats our accuracy on SUN397 by a large margin. From that, we can infer that since SUN397 has over half of its samples dedicated to outdoor scenes (55.41%), a methodology based on correlation of object parts has little capacity to compete with features that encode global structures.

As for CNN-based approaches, Table III starts by presenting the performance of an Support Vector Machine (SVM) classifier trained and tested with features from a Resnet-50 pretrained on object images (ILSVRC) and scene images (Places). Our method outperforms both of them without any ensemble on Scene15 and MIT67. As for SUN397, scene-centric features seem to have higher quality than correlating object features, serving as further evidence that outdoor scene recognition benefit from global scene-centric features. Following, we show competitive results to VGG16 (Places), the baseline proposed by [6], without any ensemble, showing that correlating object information can be as valuable as fine-tuning a CNN on the target dataset.

		Scene15	MIT67	SUN397
CNN-based	Resnet-50 (ILSVRC)	90.87%	69.13%	53.70%
	Resnet-50 (Places)	92.03%	74.73%	60.33%
	VGG16 (Places)	93.87%	80.88%	66.90%
	Herranz et al., 2016 [5]	95.18%	86.04%	70.17%
	Wang et al., 2017 [4]	-	86.20%	**73.00%**
	Nascimento et al., 2017 [6]	95.73%	87.22%	71.08%
RNN-based	Zuo et al., 2015 [23]	-	65.07%	51.14%
	Zuo et al., 2016 [24] (ILSVRC)	-	69.25%	52.78%
	Zuo et al., 2016 [24] (Places)	-	75.67%	60.34%
	Wang et al., 2017 [27]	-	71.86%	57.72%
Our Method	M2M BiLSTM Weighted Voting	94.29%	79.52%	54.00%
	Ensemble	**96.30%**	**88.25%**	71.81%

TABLE III
COMPARING THE ACCURACY OUR PROPOSED APPROACH WITH METHODS FROM THE LITERATURE. RESULTS WERE SEPARATED BY THE MAIN METHODOLOGY NATURE: CNN-BASED AND RNN-BASED.

The remaining CNN-based approaches showed in Table III are more sophisticated methodologies, some of them used as paired classifiers on our ensemble. Their performance are outstanding on all three datasets. That could be attributed to the more extensive history of applying CNN approaches to the problem of scene recognition, allowing the field to grow on a fast pace throughout the years relative to RNN-based methods. Essentially, there is much yet to be researched on recurrent approaches, which rose after CNNs. Judging by the performance increase of RNN-based methods throughout the years, it is important to unravel the full potential of recurrent

methods for image classification problems. By experimenting with CNN approaches as part of our ensemble, we found that there is still room for improvement on such methods that can be provided by high quality correlation of local information.

V. Conclusions

In this paper, we presented an approach for scene classification through context modeling of indoor scenes. Our proposal was based on the assumption that an RNN-based method is suitable for the problem of indoor scene recognition, since it can correlate object information, learning the underlying structure of scenes. Even though there are other approaches on the literature fundamentally based on the same premise, ours achieve the best result amongst RNN-based methods relying solely on object-level features, without adding information from global structures. We performed an extensive ablation analysis of individual aspects of our method, unraveling the advantages of each step of the prosed context modeling. In the future, earlier steps relative to sequence composition should be explored just as much, to evaluate its impact in the overall performance. Finally, our method can improve over state-of-the-art approaches, surpassing their performance by pairing each method with our own in an ensemble of classifiers.

Acknowledgment

The authors would like to thank the agencies CAPES, CNPq, and FAPEMIG for funding different parts of this work.

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] Y. LeCun, P. Haffner, L. Bottou, and Y. Bengio, "Object recognition with gradient-based learning," in *Shape, contour and grouping in computer vision*. Springer, 1999, pp. 319–345.

[3] 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.

[4] 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*, vol. 26, no. 4, pp. 2028–2041, 2017.

[5] L. Herranz, S. Jiang, and X. Li, "Scene recognition with cnns: objects, scales and dataset bias," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 571–579.

[6] 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.

[7] J. A. Pérez-Ortiz, J. Calera-Rubio, and M. L. Forcada, "Online text prediction with recurrent neural networks," *Neural processing letters*, vol. 14, no. 2, pp. 127–140, 2001.

[8] M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSTM modeling," in *INTERSPEECH 2010, Makuhari, Japan*, 2010, pp. 2362–2365.

[9] T.-J. Hsieh, H.-F. Hsiao, and W.-C. Yeh, "Forecasting stock markets using wavelet transforms and recurrent neural networks: An integrated system based on artificial bee colony algorithm," *Applied soft computing*, vol. 11, no. 2, pp. 2510–2525, 2011.

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

[11] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[12] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2. IEEE, 2005, pp. 524–531.

[13] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 413–420.

[14] A. Vailaya, A. Jain, and H. J. Zhang, "On image classification: city vs. landscape," in *IEEE Workshop on Content-Based Access of Image and Video Libraries (Cat. No. 98EX173)*. IEEE, 1998, pp. 3–8.

[15] I. Ulrich and I. Nourbakhsh, "Appearance-based place recognition for topological localization," in *ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation.*, vol. 2. Ieee, 2000, pp. 1023–1029.

[16] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International journal of computer vision*, vol. 42, no. 3, pp. 145–175, 2001.

[17] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 2169–2178.

[18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[19] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," in *IEEE conference on computer vision and pattern recognition workshops*, 2014, pp. 806–813.

[20] 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.

[21] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "CNN-RNN: A unified framework for multi-label image classification," in *IEEE conference on computer vision and pattern recognition*, 2016, pp. 2285–2294.

[22] W. Byeon, T. M. Breuel, F. Raue, and M. Liwicki, "Scene labeling with LSTM recurrent neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3547–3555.

[23] Z. Zuo, B. Shuai, G. Wang, X. Liu, X. Wang, B. Wang, and Y. Chen, "Convolutional recurrent neural networks: Learning spatial dependencies for image representation," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2015, pp. 18–26.

[24] ——, "Learning contextual dependence with convolutional hierarchical recurrent neural networks," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 2983–2996, 2016.

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

[26] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.

[27] Y. Wang and W. Pan, "Scene recognition with sequential object context," in *Chinese Conference on Computer Vision*. Springer, 2017, pp. 108–119.

[28] 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, pp. 770–778.

[29] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

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