Zero-shot object prediction and context modeling using semantic scene knowledge

Rene Grzeszick
rene.grzeszick@tu-dortmund.de

Gernot A. Fink
gernot.fink@tu-dortmund.de

Department of Computer Science
TU Dortmund University

Abstract

This work will focus on the semantic relations between scenes and objects for visual object recognition. Semantic knowledge can be a powerful source of information especially in scenarios with less or no annotated training samples. These scenarios are referred to as zero-shot recognition and often build on visual attributes. Here, instead of attributes a more direct way is pursued, relating scenes and objects. The contribution of this paper is two-fold: First, it will be shown that scene knowledge can be an important cue for predicting objects in an unsupervised manner. This is especially useful in cluttered scenes where visual recognition may be difficult. Second, it will be shown that this information can easily be integrated as a context model for object detection in a supervised setting.

1 Introduction

Much progress has been made in the field of image classification and object detection, yielding impressive results in terms of visual analysis. Latest results show that up to a thousand categories and more can be learned based on labeled instances [22]. In comparison, it is estimated that humans recognize about 30,000 visual categories and even more sub-categories such as car brands or animal breeds [18]. Human learning is different from machine learning, although it can be based on visual examples, it is also based on external knowledge such descriptions of entities or the context in which they appear. Recognition systems often omit basic knowledge on a descriptive level. This work will focus on the semantic relations between scenes and objects which can be an important cue for predicting objects. This is especially useful in cluttered scenes where visual recognition may be difficult. Knowing the scene context, which yields a strong prior what can be expected in a given image, the presence of objects in an image will be predicted.

It is known that contextual information can help in the task of recognizing objects [4, 5, 26]. Various forms of context that can improve visual recognition tasks have already been investigated in [5]. Visual context can be obtained in a very local manner such as pixel context or in a global manner by image descriptors like the Gist [17]. Another form of visual context is the presence, appearance or location of different objects in a scene. External context cues are, for example, of photogrammetric, cultural, geographic or semantic nature.

Especially the object level approaches that define context based on the dependencies and co-occurrences of different objects are pursued in several works [4, 9]. In [9] a stacked SVM
classifier is applied that uses the maximal detection scores for each object category in an image in order to re-rank the prediction scores. The work presented in [4] uses a hierarchical tree structure in order to model the occurrence of objects as well as a spatial prior. More global approaches use image level context definitions for object detection or image parsing [13, 16]. Most of these context definitions follow the same approach: They retrieve a subset of training images which are similar to the given image and transfer the object information [13]. A slightly different approach is pursued in [16], where a Context Forest is trained that learns the relation between a global image descriptor and the objects within this image. This allows to efficiently find related images based on the forests leaf nodes and then transfer assumptions about object locations or classes. Most of these works have in common that a considerable effort went into training a state-of-the-art detector and the results of various detections are combined in order to obtain a context descriptor. In [4, 9, 16] these were deformable part based models that build on HOG features. Nowadays, Convolutional Neural Networks (CNNs), like very deep CNNs [22] and R-CNNs [10, 19] show state-of-the-art performance in object prediction and object detection respectively. While methods like data augmentation and pre-training have reduced the required number of samples and weakly supervised annotation schemes lower the required level of detail, still a considerable annotation effort is required to train these models.

A different idea is adding further modalities for context. The most prominent of these modalities is text, for example, image captions or additional tags [12]. These multi-modal approaches allow for answering of visual queries [23, 26], the captioning of images or videos [6, 21] and recognition with limited training samples supported by additional linguistic knowledge [11, 20]. Several of these approaches incorporate additional attributes that allow for transferring knowledge without explicitly annotating a specific class or object label [11, 20, 26]. In [26] a knowledge base system is built that relates scenes with attributes and affordances. It is shown that the association of scenes with attributes and affordances allows for improving the predictions of scenes and their attributes as well as answering visual queries. In [11] attributes that are associated with a set of images or classes are used for uncovering unknown classes and describing them in terms of their attributes. Such methods with no training samples for given objects are also referred to as zero-shot learning approaches. Similarly, given a very small set of training samples, attributes can be used in order to transfer class labels to unknown images [20]. Attributes for images can either be learned directly via annotated training images or indirectly via additional sources of information such as Wikipedia or WordNet [11, 20]. As annotating images with attributes is tedious, especially, the latter allows for scaling recognizers to a larger number of classes and attributes. Furthermore, it has been shown that these attributes can also be derived in a hierarchical manner, i.e. based on the WordNet tree [20].

This work will show that additional textual information is beneficial for predicting object presence in an image with minimal annotation effort similar to [11, 20, 26]. Here, instead of attributes, a more direct way is proposed exploiting the semantic relations between scenes and objects in a zero-shot approach. The presence of an object is predicted based on two sources: visual knowledge about the scene and the relations between scenes and different objects. For example, a car can hardly be observed in the livingroom or a dining table in the garage. Such knowledge can easily be obtained from additional textual sources. Methods like TextRunner [1] or Reverb [7] allow for processing large text corpora and uncovering nouns, like scene or object names. As a result the tremendous annotation effort that is required for annotating objects in images is no longer necessary. Similar to existing zero-shot approaches, it requires only a descriptive label, the scene name, and works completely un-
supervised with respect to annotations of objects. In the experiments it will be shown that such high level knowledge allows for predicting the presence of objects, especially in very cluttered scenes. Furthermore, in a supervised setting, it will be shown that the presence information can be integrated into a context model for state-of-the-art object detectors.

2 Method

In the proposed method for object prediction, the relations between scenes and objects which are obtained from text sources are used for modeling top down knowledge. They replace the visual information that is typically used for object prediction. An overview is given in Fig. 1. The image is solely analyzed on scene level, which requires minimal annotation effort and no visual knowledge about the objects within the scene. More importantly, a text corpus is analyzed with respect to possible scenes and objects, extracting the relations between them and creating a matrix of objects in a scene context. This information is then used in combination with the scene prediction in order to predict the presence of an object in an image. Furthermore, this information can also be applied for modeling a context descriptor in a supervised object detection setting.

2.1 Relations between objects and scenes

Methods like TextRunner [1] or Reverb [7] allow for uncovering information from text. Here, sentences including possible scene or object categories and their relations will be of further interest. In the following extractions based on Reverb are used. Reverb extracts relations and their arguments from a given sentence. Therefore, two steps are performed: First, for each verb $v$ in the sentence, the longest sequence of words $r_v$ is uncovered so that $r_v$ starts at $v$ and satisfies both a syntactical and a lexical constraint. For the lexical constraint, a dictionary of 1.7 million relations is used. The syntactical constraint is based on the following regular expression:

$$V | VP | WV^* P$$

$V = \text{verb particle? adv?}$

$W = (\text{noun}|\text{adj}|\text{adv}|\text{pron}|\text{det})$

$P = (\text{prep}|\text{particle}|\text{inf. marker})$
Overlapping sets of relations are merged into a single relation. Second, given the relation \( r \), the nearest noun left and right of the relation \( r \) are extracted. If two nouns can be observed for a relation \( r \), this results in a triplet

\[
t = (\text{arg}1, r, \text{arg}2) .
\] (2)

Given a large enough set of text, relations extracted from these texts can be assumed to roughly represent relations that are observed in the real world and henceforth may also be observed on images. Rich text corpora can, for example, be obtained by crawling Wikipedia or any other source of textual information [20]. Such information can then be used for visual object and scene recognition. It is assumed that the likelihood of an object to occur in a given scene is correlated with the number of textual relations between those entities. Given a set of \( S \) scene and \( O \) object names, a matrix \( C \) describing the objects in scene context is created. At the index \( s, o \) it contains the number of relations between the respective scene \( s \) and object \( o \):

\[
C_{s,o} = \sum_r n(o,s) + \sum_r n(s,o) \quad \text{with} \quad n(i,j) = \#\{t = (i, \cdot, j)\} .
\] (3)

In the experiments, some rare cases of self-similarity were observed, in which a scene and an object name are the same (e.g. a scene in a street may also show the object road/street, among others). In these cases the self-similarity is set to the maximum count observed.

### 2.2 Presence Prediction

The task of presence detection is concerned with the question whether an object can be observed one or more times in a given image. Multiple objects can occur in a single scene image and the presence predictions for different objects are typically evaluated independently of each other. Since multiple objects may occur in the same scene, the probability \( P(o|s) \) of the object \( o \) to be shown in the scene \( s \) cannot be computed directly using the counts. However, since an image can only depict one scene, \( P(s|o) \) can be estimated from the counts. This allows to compute \( P(o|s) \) based on Bayes theorem as

\[
P(o|s) = \frac{P(s|o)P(o)}{P(s)} \quad \text{with} \quad P(s|o) = \frac{C_{s,o}}{\sum_{s'} C_{s',o}} \quad \text{and} \quad P(o) = \frac{\sum_o C_{s',o}}{\sum_i \sum_j C_{i,j}} .
\] (4)

A prior probability \( P(s) \) for a certain scene cannot be derived from the matrix of objects in scene context as this is rather dependent on the input images. Assume the Matrix \( C \) contains \( N \) relation counts which relate the presence of at least \( N \) objects from \( O \) categories to the set of \( S \) scenes. Given the one to many relation between a scene and objects, the count of scenes cannot be recovered. Therefore, \( P(s) \) is assumed to be uniformly distributed.

In order to be able to predict an object in a scene where no relations have been previously observed, unobserved events need to be handled. Therefore, the probability of an \( o \) object to occur in a scene \( s \) is smoothed by

\[
P^*(o|s) = (1 - \alpha) P(o|s) + \alpha P(o) ,
\] (5)

similar to the smoothing of probability distributions for statistical natural language processing (cf. [14]). The process is based on an interpolation factor \( \alpha \) which is estimated based on the number of events with only a single occurrence:

\[
\alpha = \frac{\#\{C_{o,s}|C_{o,s} = 1\}}{\sum_{o',s'} C_{o',s'}}
\] (6)
Furthermore, as the counts are obtained from an unrelated data source, there remains a degree of uncertainty and, more importantly, the matrix representation does not cover intra-scene variability. In order to model these two issues, \( \hat{P}(o|s) \) is sampled by \( D \) draws from a normal distribution

\[
\hat{P}(o|s) = \frac{1}{D} \sum_{n=1}^{D} n \quad \text{with} \quad n \sim \mathcal{N}(P^{*}(o|s), \sigma(C)).
\]  

(7)

The variance \( \sigma \) is estimated based on the variance within the matrix \( C \). In order to estimate the probability of an object in a given image \( I \), the presence is then predicted by:

\[
P(o|I) = \sum_{s} P(s|I) \cdot \hat{P}(o|s).
\]  

(8)

The probability \( P(s|I) \) can be predicted by a classifier. In this work, a CNN will be used.

Note that the requirement for a single scene label is very lightweight. Similar to attribute-based zero-shot learning, it is a descriptive abstraction that does not imply any visual knowledge about the objects within the scene.

### 2.3 Context descriptor

Given a more informed setting, where a set of object detectors has been trained and evaluated, the obtained presence probability can also be used for modeling a context descriptor. Following the approach of the Context SVM [9], the output of an object detector can be re-scored. Here, a Scene-Context SVM is trained. Let \((B, y)\) be the detection window of an object detector, i.e. an R-CNN [19], described by a bounding box \( B \) and a score \( y \). Then, for each detected bounding box a context descriptor is derived:

\[
g = (y, P(o_1|I), ..., P(o_n|I))
\]  

(9)

The location and object co-occurrence from the re-scoring in [9] are replaced by the scene priors. One SVM is trained for each class on the true positive and false positive (defined by one percentile of the true positive scores) detections using a linear kernel and Platt’s scaling.

### 3 Evaluation

In the following the proposed object prediction and context modeling is evaluated. Ideally, the evaluation requires a dataset that offers both scene and object labels. The zero-shot recognition has been evaluated on different branches of the SUN dataset [25] which show a broad set of different objects. The approach has then been generalized to the Pascal VOC object detection benchmark where the context descriptor is evaluated as well.

**SUN2012 Pascal**: the dataset contains a set of 397 different scenes and 4,919 different object labels [25]. In total there are 11,426 images for which scene and object annotations are available. There is a great variability with respect to the objects properties. While some of them are well defined (e.g. cars, person), some others describe regions (sky, road, buildings) or highly deformable objects (river, curtain). Moreover, the annotations in all versions of the SUN dataset are very noisy. Some of them contain descriptive attributes, like person walking or table occluded others are singular or plural. For all the experiments, these attributes were removed and all objects and scene names were lemmatized based on the WordNet tree [15].
This leaves 3,390 different objects in 377 different scenes.

**SUN2009 Context:** the dataset contains only about 200 different object categories, of which 107 were used for detection in [4]. The same diversity as for the SUN2012 Pascal version can be observed with respect to the objects properties.

**Pascal VOC 2007:** the Pascal dataset contains 20 well defined object classes in less cluttered scenes than in the different SUN datasets and is a standard object detection benchmark.

### 3.1 Creating a Matrix of Objects in Scene Context

In order to obtain a matrix of objects in scene context, the OpenIE database has been queried. It contains over 5 billion extractions from over a billion web pages\(^1\). Hence, a very diverse dataset that captures the relations between a huge set of nouns has been used. The remaining 377 scene names and 3390 object names from the SUN2012 Pascal dataset were queried for which a total relation count of 1,375,559 has been extracted. Note that the distribution of these relation counts is very long tailed.

### 3.2 Scene prediction

For recognizing objects based on the scene context, the probability of the given image to depict the scene \(s\) needs to be computed. In the following, two different setups are evaluated.

**Known label:** it is assumed that the scene label is known beforehand. Here \(P(s|I)\) is set to 1 for the annotated scene label and to 0 otherwise. Note that these labels might be ambiguous and even human annotators deviate in their decision from the ground truth labels [24].

**Scene-level CNN:** For recognizing scene labels a CNN is evaluated. A VGG16 network architecture has been trained on all scene images from the SUN397 dataset that are not included in the SUN2012 Pascal version. The network has been pre-trained on ImageNet, as this typically improves the performance. The exclusion of the SUN2012 Pascal images leaves a set of 97,304 training images. The training images have been augmented using random translations (0 − 5%), flipping (50% chance) and Gaussian noise (\(\sigma = 0.02\)) in order to achieve a better generalization. In total 500,000 training images have been created. The learning rate has been set to \(\alpha = 0.0001\) using 25,000 training iterations of batch size 39 (= 10% of the number of classes). The CNN yields an accuracy of 62.2% for the scene images in the SUN2012 Pascal set.

### 3.3 Object prediction

In the following experiments, the presence detection is evaluated. Only scene labels are used in the training so that it is completely unsupervised with respect to object occurrences.

**SUN2012 Pascal** The results for the object prediction on the SUN2012 Pascal dataset are shown in Fig. 2. The most frequently occurring 100 up to all 3390 objects categories from the dataset were considered, some of them being comparably rare. On the left the experiment using the annotated scene labels are shown and on the right the scene labels as predicted by the CNN are used. The accuracy of the top \(k\) predictions is evaluated. Only images with \(k\) annotated objects were used for the evaluation. Interestingly, predicting the scene label by the CNN improves the result. This can be explained as the CNN computes a soft assignment

\(^{1}\)For a demo see http://openie.allenie.org
to a set different of scenes. The obtained distribution of objects in scene context, does not completely match the ground truth distribution as observed in the dataset and, therefore, a distribution over a soft assignment to many scenes is more robust. Even without any knowledge of the visual appearance of an object, the experiments reveal that the highest ranking objects have a precision of up to 52.6% when considering a set of 100 objects and 35.9% when considering as much as 3,390 different object categories. However, as mentioned before, some of the objects describing regional objects tend to be very general and can safely be assumed to occur in most scenes.

Exemplary results showing the five highest scoring object predictions for a given image are shown in Fig. 3. It can be seen that although a large set of objects is annotated in the SUN dataset, the annotations are noisy and not at all complete. Some of the predictions that cannot be found in the ground truth annotations might be deemed as correct. Predictions that are not shown in the image, are often at least plausible guesses what else could be found in the scene. For example, parking and even people might be related to a highway or street scene or a sofa might be related to an indoor scene. The example in the middle is a typical case where the CNN predicts a label incorrectly, which is, however, not only visually but also semantically related. The example on the right shows a typical example of ambiguity in natural language as well as in the provided annotations.

Figure 3: Exemplary results showing the five highest scoring object predictions: (green) correct (red) wrong (red & italic) wrong according to annotations, but can be seen in the image. In the bottom row: (left) Annotation (right) highest scoring CNN prediction.
In order to provide a more detailed analysis, different sets of object categories are evaluated based on the VOC mean average precision (mAP) criterion [8]. The results for the 20 to 100 most common objects in the dataset is given in Tab. 1. The mAP of predicting an object by chance is shown, indicating how frequently these objects occur in the dataset. For comparison, the ground truth distribution of all objects in a given scene has been evaluated using the same model. This can be seen as an indication for an upper bound of the performance that could be obtained by solely using the proposed model of scene and object relations.

**SUN2009 Context** In order to emphasize the difficulty of detecting objects with a huge variability, as depicted in the SUN dataset, the approach has also been evaluated on 107 object categories of the SUN2009 Context dataset. The presence detection of the proposed zero-shot method after predicting a scene label for each of the scenes using the scene-level CNN is compared to supervised object detectors. On this dataset, different object detectors based on deformable part based models were trained in a fully supervised manner and evaluated in [4]. The results are shown in Table 2. As the original evaluation protocol contained three classes that were filtered by the stemming (bottles, stones and rocks), both results are displayed. Although R-CNNs have surpassed deformable part based models as the state-of-the-art in object detection [10], it is surprising that a model that is solely based on scene-level predictions can achieve comparable results to object detection models that are trained completely supervised.

**VOC 2007** The limitations of the proposed object prediction can be observed on the VOC 2007 dataset. Here, the approach is generalized to a completely different dataset using the same scene and object categories as gathered for the SUN dataset. Then the presence of the 20 classes of the VOC benchmark is predicted with the results shown in Table 3. For predicting the classes, some of the labels were mapped to the best matching synonym (‘pottedplant’-
Table 3: Mean average precision for the presence prediction (left) and detection using context rescoring (right) of the 20 objects in the VOC2007 dataset. (*) The CNN predicts one of the 397 scenes from the SUN397 dataset without any knowledge about the VOC dataset.

'plant', 'diningtable'- 'table' 'aeroplane'- 'airplane' and 'tvmonitor'- 'television'). Note that when generalizing to this dataset, it is assumed that the scenes in the SUN dataset and their characteristics are also representative for the scenes found in the VOC dataset. Furthermore, the VOC task is especially designed for visual object detection and the classes are not as ambiguous as the ones in the SUN dataset. They can therefore be learned more efficiently by a visual classifier. However, the proposed approach is still able to make an 'educated guess'.

3.4 Object Detection with Scene Context

Although the presence detection on the VOC dataset appears to be difficult in an unsupervised manner, it can be shown that the object prediction scores are still a powerful context cue. The results for an R-CNN object detector [19] using a Region Proposal Network with 300 proposal have been re-scored by the proposed Scene-Context SVM and are shown in Tab. 3. For comparison, a Context SVM as introduced in [9] has been evaluated. Both approaches were trained on the VOC2007 trainval set. Although the theoretical upper bound allows for some improvement, the low performance of the Context SVM emphasizes the difficulty of re-ranking the CNN’s predictions. This is mainly due to the fact that the network’s confidence is usually quite high. Still the proposed Scene-Context SVM that integrates a strong prior is able to slightly improve the detection scores.

4 Conclusion

In this work a novel approach for predicting object presence in an image is presented. The method works in a zero-shot manner and only relies on scene level annotations from which a probability for an object’s presence is derived. The probability is based on the relations between scenes and objects that were obtained from additional text corpora. As a result the proposed method is completely unsupervised with respect to objects and allows for predicting objects without any visual information about them. In the experiments it has been shown that it is possible to predict the occurrences for as many as 3390 objects. On tasks that are very difficult for visual classifiers, such as cluttered scenes with not very well structured objects, the approach yields similar performance to visual object detectors that were trained in a fully supervised manner. Furthermore, it could be demonstrated that the obtained presence predictions can form a powerful context descriptor for object detection.
5 Acknowledgment

This work has been supported by the German Research Foundation (DFG) within project Fi799/3. The authors would like to thank Prof. Kristian Kersting for his helpful comments and discussions.

References

[1] Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. Open information extraction for the web. In *IJCAI*, volume 7, pages 2670–2676, 2007.

[2] K Chatfield, V Lempitsky, A Vedaldi, and A Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *Proc. British Machine Vision Conference (BMVC)*, 2011.

[3] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *British Machine Vision Conference*, 2014.

[4] Myung Jin Choi, Antonio Torralba, and Alan S Willsky. A tree-based context model for object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(2):240–252, 2012.

[5] S K Divvala, D Hoiem, J H Hays, A A Efros, and M Hebert. An empirical study of context in object detection. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1271–1278, 2009.

[6] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[7] Oren Etzioni, Anthony Fader, Janara Christensen, Stephen Soderland, and Mausam Mausam. Open information extraction: The second generation. In *IJCAI*, volume 11, pages 3–10, 2011.

[8] M Everingham, L Van Gool, C K I Williams, J Winn, and A Zisserman. The PASCAL visual object classes challenge 2011 (VOC2011) results. http://www.pascal-network.org/challenges/VOC/voc2011/workshop/index.html, 2011.

[9] P F Felzenszwalb, R B Girshick, D McAllester, and D Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[10] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Region-based convolutional networks for accurate object detection and segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(1):142–158, 2016.
[11] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465, 2014.

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proc. European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014.

[13] Ce Liu, Jenny Yuen, and Antonio Torralba. Nonparametric scene parsing: Label transfer via dense scene alignment. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1972–1979. IEEE, 2009.

[14] Christopher D Manning and Hinrich Schütze. *Foundations of statistical natural language processing*, volume 999. MIT Press, 1999.

[15] G A Miller and Others. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[16] Davide Modolo, Alexander Vezhnevets, and Vittorio Ferrari. Context forest for object class detection. 2015.

[17] Aude Oliva and Antonio Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 155:23, 2006.

[18] Stephen E Palmer. *Vision science: Photons to phenomenology*. MIT press Cambridge, MA, 1999.

[19] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015.

[20] Marcus Rohrbach, Michael Stark, and Bernt Schiele. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1641–1648. IEEE, 2011.

[21] Marcus Rohrbach, Qiu Wei, Ivan Titov, Stefan Thater, Manfred Pinkal, and Bernt Schiele. Translating video content to natural language descriptions. In *IEEE International Conference on Computer Vision*, 2013.

[22] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[23] Qi Wu, Chunhua Shen, Anton van den Hengel, Peng Wang, and Anthony Dick. Image captioning and visual question answering based on attributes and their related external knowledge. *arXiv preprint arXiv:1603.02814*, 2016.

[24] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3485–3492. IEEE, 2010.
[25] Jianxiong Xiao, Krista A Ehinger, James Hays, Antonio Torralba, and Aude Oliva. SUN Database: Exploring a Large Collection of Scene Categories. *International Journal of Computer Vision (IJCV)*, pages 1–20, 2014.

[26] Yuke Zhu, Ce Zhang, Christopher Ré, and Li Fei-Fei. Building a large-scale multimodal knowledge base system for answering visual queries. *arXiv preprint arXiv:1507.05670*, 2015.
6 Supplementary Material

This supplementary material of the submission contains additional evaluation results and a few more examples.

6.1 SUN2012 Pascal - Exemplary results

Figure 4: Exemplary results showing the five highest scoring object predictions: (green) correct (red) wrong (red & italic) wrong according to annotations, but can be seen in the image. In the bottom row: (left) Annotation (right) highest scoring CNN prediction.
6.2 SUN2012 Pascal - VOC Objects

Mean average precision for the presence of the 20 VOC objects classes on the SUN Pascal-format dataset. The presence predictions are based on the number of relations between the scenes and the objects: (left) The scene label, (right) the predicted scene labels using a CNN. For comparison the prediction by chance is shown and the results using the ground truth distribution of objects in the SUN scenes using the same model. The ground truth distribution can be seen as an indicator for an upper baseline.

| Objects     | Scene Label mean AP [%] | Scene Label CNN mean AP [%] | Chance mean AP [%] | GT Distribution mean AP [%] |
|-------------|-------------------------|-----------------------------|-------------------|-----------------------------|
| Person      | 19.6                    | 18.7                        | 17.6              | 45.0                        |
| Bird        | 0.2                     | 0.2                         | 0.1               | 3.6                         |
| Cat         | 0.2                     | 0.1                         | 0.1               | 0.4                         |
| Cow         | 28.0                    | 37.2                        | 0.1               | 28.0                        |
| Dog         | 5.7                     | 2.1                         | 0.4               | 28.2                        |
| Horse       | 2.7                     | 2.2                         | 0.1               | 39.1                        |
| Sheep       | 20.7                    | 5.6                         | 0.1               | 7.7                         |
| Airplane    | 11.9                    | 14.9                        | 0.6               | 81.7                        |
| Bicycle     | 1.6                     | 1.2                         | 0.4               | 11.8                        |
| Boat        | 8.2                     | 9.5                         | 1.2               | 18.6                        |
| Bus         | 2.4                     | 6.0                         | 0.4               | 4.7                         |
| Car         | 48.6                    | 59.1                        | 8.1               | 66.1                        |
| Motorbike   | 1.5                     | 2.6                         | 0.3               | 4.3                         |
| Train       | 0.2                     | 0.2                         | 0.2               | 88.8                        |
| Bottle      | 19.1                    | 19.3                        | 5.2               | 23.1                        |
| Chair       | 37.8                    | 39.3                        | 18.1              | 59.4                        |
| Table       | 36.2                    | 36.2                        | 13.9              | 53.9                        |
| Plant       | 16.2                    | 17.9                        | 14.6              | 26.2                        |
| Sofa        | 32.5                    | 49.6                        | 4.5               | 47.2                        |
| Television  | 7.6                     | 7.7                         | 2.9               | 11.3                        |
| VOC Mean    | 15.0                    | 16.5                        | 4.5               | 32.3                        |

Table 4: Mean average precision for the 20 VOC objects classes on the SUN Pascalformat dataset.
6.3 VOC Presence

Mean average precision for the presence of the 20 VOC objects classes on the VOC2007 testset. The presence predictions are based on the number of relations between the scenes and the objects after predicting one of the 397 SUN scene labels using a CNN.

| Objects | Scene Label CNN + Objects in Context |
|---------|-------------------------------------|
| Person  | 49.21                               |
| Bird    | 12.84                               |
| Cat     | 8.34                                |
| Cow     | 33.73                               |
| Dog     | 13.70                               |
| Horse   | 22.64                               |
| Sheep   | 25.91                               |
| Aeroplane | 80.55                       |
| Bicycle | 12.90                               |
| Boat    | 43.33                               |
| Bus     | 13.14                               |
| Car     | 51.61                               |
| Motorbike | 6.38                        |
| Train   | 5.35                                |
| Bottle  | 16.31                               |
| Chair   | 47.36                               |
| Diningtable | 44.50                  |
| Pottedplant | 7.92                     |
| Sofa    | 60.10                               |
| Tvmonitor | 12.94                   |
| VOC Mean | 28.44                             |

Table 5: Mean average precision for the presence detection of the 20 VOC objects classes on the VOC 2007 dataset.
6.4 VOC Object Detection

Mean average precision for the detection 20 VOC objects classes on the VOC2007 testset. Note that the high confidence scores that are typically produced by CNNs make the task of re-ranking the bounding box prediction comparably difficult.

| Objects   | R-CNN [10] | + Context SVM [9] | + Scene-Context SVM | Upper Bound |
|-----------|------------|-------------------|---------------------|-------------|
| Person    | 75.73      | 75.56             | **75.76**           | 80.74       |
| Bird      | **69.42**  | 68.94             | 69.13               | 80.71       |
| Cat       | 85.03      | 84.27             | 85.01               | 90.61       |
| Cow       | 76.48      | 75.48             | **77.20**           | 90.17       |
| Dog       | 81.11      | 81.33             | 81.26               | 90.63       |
| Horse     | **83.58**  | 83.22             | 83.32               | 90.46       |
| Sheep     | 70.48      | 64.40             | **70.94**           | 81.82       |
| Aeroplane | **68.74**  | 68.48             | 68.72               | 71.89       |
| Bicycle   | 78.33      | 78.41             | **78.36**           | 81.54       |
| Boat      | 54.55      | 54.92             | **55.62**           | 72.36       |
| Bus       | 79.62      | 77.80             | **79.69**           | 90.72       |
| Car       | **79.79**  | 79.51             | 79.76               | 80.84       |
| Motorbike | **76.21**  | 76.16             | 75.98               | 81.27       |
| Train     | **77.88**  | 78.42             | 77.38               | 81.47       |
| Bottle    | 49.70      | 49.42             | **49.92**           | 62.68       |
| Chair     | **50.98**  | 50.30             | 50.96               | 71.54       |
| Diningtable | 65.34    | 62.91             | **65.64**           | 81.56       |
| Pottedplant | **38.40** | 37.36             | 38.34               | 62.81       |
| Sofa      | 65.80      | 64.40             | **66.14**           | 88.52       |
| Tvmonitor | 65.70      | 63.38             | **65.76**           | 81.43       |
| VOC Mean  | 69.64      | 69.07             | **69.75**           | 80.69       |

Table 6: Mean average precision for the detection of the 20 VOC objects classes on the VOC 2007 dataset.

The Training Parameters for the Scene-Context SVM were:

*Linear Kernel, No shrinking heuristic, C=1, Train Tolerance=0.01*