Class-Selective Mini-Batching and Multitask Learning for Visual Relationship Recognition

S. Josias and W. Brink

Abstract—**An image can be described by the objects within it, and interactions between those objects. A pair of object labels together with an interaction label is known as a visual relationship, and is represented as a triplet of the form (subject, predicate, object). Recognising visual relationships in images is a challenging task, owing to the combinatorially large number of possible relationship triplets, which leads to an extreme multi-class classification problem. In addition, the distribution of visual relationships in a dataset tends to be long-tailed, i.e. most triplets occur rarely compared to a small number of dominating triplets. Three strategies to address these issues are investigated. Firstly, instead of predicting the full triplet, models can be trained to predict each of the three elements separately. Secondly a multitask learning strategy is investigated, where shared network parameters are used to perform the three separate predictions. Thirdly, a class-selective mini-batch construction strategy is used to expose the network to more of the rare classes during training. Experiments demonstrate that class-selective mini-batch construction can improve performance on classes in the long tail of the data distribution, possibly at the expense of accuracy on the small number of dominating classes. It is also found that a multitask model neither improves nor impedes performance in any significant way, but that its smaller size may be beneficial. In an effort to better understand the behaviour of the various models, a novel evaluation approach for visual relationship recognition is introduced. We conclude that the use of semantics can be helpful in the modelling and evaluation process.**

Index Terms—**mini-batch construction, multitask learning, visual relationship recognition**

I. Introduction

THERE exists a variety of effective computer vision methods for locating and labelling objects in an image [1], [2]. In order to further develop the image understanding pipeline one can consider methods for recognising interactions or relationships between different objects in the same image.

A visual relationship is defined as a triplet of the form (subject, predicate, object) and describes some visible interaction between a pair of objects in an image. The image in Fig. 1, for example, contains the visual relationship (boy, on top of, surfboard). Such visual relationships can be used to construct scene graph representations [3] for advanced visual reasoning in tasks such as image retrieval, visual question answering, and automated surveillance.

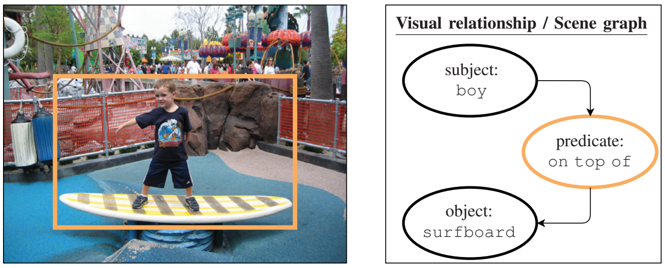


Fig. 1. An example of visual relationship recognition. The task is to label the subject, the predicate (relationship) and the object, given an image and a bounding box around a pair of objects.

Visual relationship recognition is the problem of producing (subject, predicate, object) triplets from a given image. It is often coupled with the object localisation problem, but the focus of this paper is on the labelling task and we will therefore assume knowledge of tight bounding boxes around pairs of objects (a sensitivity analysis on this assumption follows later in the paper). Bounding boxes around objects can be generated by an off-the-shelf object detector (e.g. [1]) and merged pairwise in a straightforward manner.

Visual relationship recognition is challenging for a number of reasons. Firstly, predicates tend to be slightly more abstract than the subjects and objects, often making their visual representations difficult to model and recognise. Secondly, the number of possible relationships explodes combinatorially and leads to what is known as an extreme multiclass classification problem. For example, 100 possible subject and object labels, and 70 possible predicates, amount to 700,000 possible triplets. Such a large number means that it is difficult to collect data representative of all those triplets. In fact, one of the first datasets on visual relationship recognition, called VRD [4], represents a 700,000 class problem (taking all possible combinations of subjects, predicates and objects into account) but contains data of only around 15,000 unique visual relationships. A substantially larger dataset called Visual Genome [5] contains 75,729 unique objects but only 40,480 unique visual relationships.

The third challenge in visual relationship recognition is that the distribution of triplets in a dataset typically exhibits a long tail: the vast majority of possible triplets occur only a few times (or never) in the training set, while a small number may be much more frequent. A long tail in the distribution of training data is problematic for optimisation-based learning, because an undesired local optimum to the objective can be found quickly by merely predicting the dominant classes most often.

Based on “Batch construction and multitask learning in visual relationship recognition”, by Shane Josias and Willie Brink which was published in the Proceedings of the SAUPEC/RobMech/PRASA 2020 Conference held in Cape Town, South Africa 29 to 31 January 2020. © 2020 SAIEE

This work was supported by the CSIR/SU Centre for Artificial Intelligence Research (CAIR).

Shane Josias and Willie Brink are with the Department of Mathematical Sciences, Stellenbosch University, Stellenbosch, South Africa (e-mail: josias@sun.ac.za, wbrink@sun.ac.za).

Another challenge is that there can be an inherent ambiguity in the labelling of visual relationships. For example, the visual relationship (`boy`, `on top of`, `surfboard`) can legitimately be labelled (`boy`, `riding`, `surfboard`) or (`surfboard`, `under`, `boy`). Multiple semantically correct classifications of a visual relationship, with typically only a single ground truth label in the dataset, make both the modelling and evaluation of visual relationship recognition difficult.

The aim of this paper, therefore, is to investigate a number of strategies to deal with these challenges. To address the combinatorially large set of possible classes, models can be designed to predict the elements of a triplet separately, instead of the triplet as a whole. This strategy also allows for a multitask design where the different elements can be predicted with shared model parameters, potentially resulting in inductive transfer and statistical data amplification [6] for improved generalisation. For model training we also implement selective mini-batch construction, in an effort to better capture the long tail of the distribution over visual relationships.

The performance of the proposed multitask model is then compared against multiple single-task models, while the performance of the proposed mini-batch construction strategy is compared against standard uniformly random mini-batch sampling. To partially address the problem of semantic ambiguity in the labels of visual relationships a new performance metric is also presented. Finally the sensitivity of the best-performing model is measured against perturbations in the assumed bounding box coordinates.

The contributions of this work can be summarised as follows:

1) we demonstrate that multitask learning in the setting of visual relationship recognition is effective at reducing model complexity, without a significant positive or negative impact on performance;
2) we show that the proposed approach to mini-batch construction is useful as a simple strategy to improve performance on underrepresented relationships;
3) we introduce a performance metric that seems to improve the understanding of model behaviour in visual relationship recognition.

II. Related work

The literature on visual relationship recognition can be grouped broadly into three common approaches. The first involves the learning of a visual-semantic embedding space. Such embeddings can be achieved by imposing criteria such as small distances between similar relationships [4], by modelling a relationship as vector translation between embedded objects [7], or by minimising a triplet-softmax loss [8]. Visual-semantic embedding allows for few- and zero-shot learning, and can therefore be suitable for modelling a long-tailed distribution; however a separate classifier would still need to be trained on top of the embedding.

The second common approach attempts to generate the scene graph, or collection of interconnected relationships, directly. Xu *et al.* [9] perform probabilistic graph inference with a structural recurrent neural network and an iterative message passing scheme to refine the predictions. Zellers *et al.* [10] observe that natural images usually have certain kinds of structural regularities, which they dub motifs, and propose stacked neural networks (“MotifNets”) to predict graph elements as well as an LSTM to encode global context. Further examples of this approach include the use of associative embeddings [11], graph parsing neural networks [12], and graph R-CNN [13]. Woo *et al.* [14] improve on graph generation strategies by designing an explicit relational reasoning module. Generating a scene graph is more direct than the visual-semantic embedding approach, and end-to-end training to accomplish the intended task directly can lead to superior performance.

The third approach, and the one most relevant to our work, treats the prediction of each element of the visual relationship triplet as its own classification task. Some works use multi-stream architectures for each task [15]–[18], while others employ a single multitask scheme [19], [20] which is similar to what we will investigate.

There seems to be a central theme of transferring knowledge for improved performance through message passing, global context cues, or inductive transfer in multitask learning. The multi-stream and multitask settings can deal with the huge number of classes in visual relationship recognition by making use of multiple outputs of smaller dimensions. It remains unclear, however, whether multitask learning would necessarily provide better performance. Existing approaches also tend to build very large systems, with many parameters, and it is usually not clear exactly how the long tail of typical datasets are dealt with. Within the domain of visual relationship recognition we have not yet come across approaches dealing with the long-tailed nature of training data distributions.

Overall, significant efforts are also being made to construct richer datasets that allow for better learning of visual relationships. The first major dataset released is called VRD [4]. It contains 5,000 images with instances of around 15,000 unique visual relationships. A much larger dataset called Visual Genome, containing 108,077 images with 40,480 unique relationships, was later introduced by Krishna *et al.* [5]. The number of images in Visual Genome is greater than in VRD, and the total number of visual relationship classes have also increased. The long-tailed distribution seems to be inherent in the problem of visual relationship recognition, and can be exacerbated with more data. The Google Open Images Challenge [21] attempts to find a middle ground by considering only 329 possible visual relationship triplets with 375,000 visual relationship instances.

III. Methodology

The aim of visual relationship recognition within the context of this paper is to train a neural network model that takes an image cropped around a pair of objects as input, and generates scores over possible (subject, predicate, object) triplets as output. Training labels are used to define fixed vocabularies for each of the three elements of a triplet. Visual relationship recognition may therefore be treated as a classification problem, and models can be set up to output normalised class scores over triplets. It should be noted that subjects and objects often share the same vocabulary, but this is not a strict requirement.

Instead of attempting to train a convolutional neural network to output one massive vector of scores over all possible triplets, three separate tasks can be considered: predicting the subject label, predicting the predicate label, and predicting the object label. Each of these tasks has far fewer possible classes, and under the simplifying assumption that the tasks are conditionally independent given an image, the normalised output scores can be combined through multiplication. In this way the top scoring triplet can be obtained by combining the top scoring elements from each of the three separate predictions.

As mentioned in section I, typical datasets for training and evaluating visual relationship recognition models exhibit a long tail not only in the distribution over all triplets, but also in each of the marginal distributions over subjects, predicates and objects. A visualisation of this behaviour in the VRD dataset follows in section IV-B.

A. Single-task learning with standard mini-batching

A first approach can be to create three separate neural network models to predict the subject, the predicate and the object from the same image crop. Each network may consist of the convolutional base of a pre-trained network (in our experiments ResNet-18 [22] is chosen for its good balance between size and performance), followed by three trainable, 2,048-dimensional fully-connected layers and a softmax output layer. Refer to Fig. 2 for an illustration.

In order to train each model a cross-entropy loss function can be minimised with mini-batch gradient descent [23]. For each training iteration a mini-batch of some prespecified size is sampled, without replacement, uniformly across all samples in the training set.

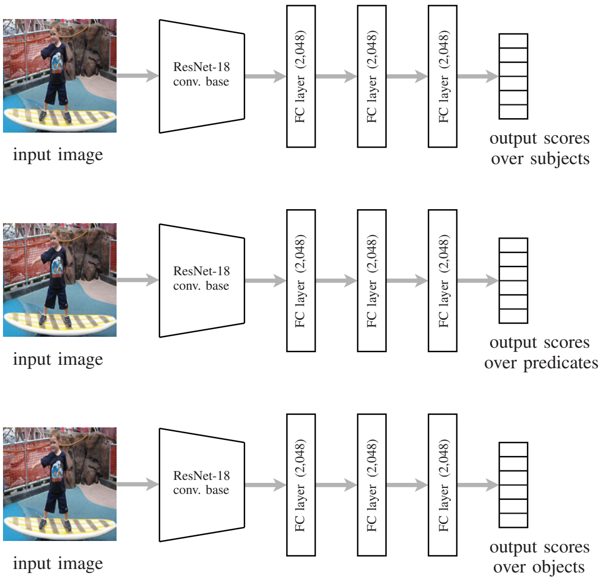


Fig. 2. In the single-task learning setting, three separate models learn to predict respectively the subject, predicate and object from a given image crop.

In the case of visual relationship recognition, where the data is often heavily skewed and exhibits a long-tailed distribution over labels, this "standard" approach to mini-batch selection is likely to pick samples mostly from a small number of frequently occurring classes. The networks may thus learn these dominant classes very well, but would be unable to recognise the vast majority of classes in the long tail of the data distribution.

B. Class-selective mini-batch construction

In an effort to mitigate the potential problem with standard mini-batch selection mentioned above, and expose the network to more classes in the tail of the dataset, the following mini-batch construction strategy is proposed. For a particular task (which can be to predict either the subject, the predicate, or the object) we sample at every training iteration n classes from the vocabulary of that task, uniformly at random. We then randomly select m samples from each of those n classes, for a mini-batch of size mn. Fig. 3 illustrates this strategy on a small example.

Constructing mini-batches in this manner would allow a network to learn from all the classes in a particular task, in roughly equal measure. The hypothesis is that this construction may lead to better performance on the many rare classes in the long tail of the data, potentially at the expense of reduced performance on the small number of dominant classes. Of course, there is now a risk of biasing the network against the true distribution of the data and impede its ability to generalise properly. These issues are investigated experimentally in section V.

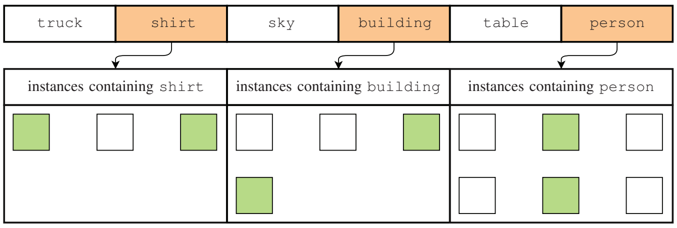

truck
shirt
sky
building
table
person
instances containing shirt
instances containing building
instances containing person

Fig. 3. For a vocabulary of size $N = 6$, we randomly select $n = 3$ classes. From each of these, $m = 2$ instances are randomly selected to form a single mini-batch (green boxes).

C. Multitask learning

We also explore the efficacy of multitask learning, which can be thought of as an inductive form of transfer learning where knowledge is transferred across the three visual relationship recognition tasks.

Multitask learning makes the assumption that the predictive model should have an ability to explain multiple tasks. This assumption is also referred to as an inductive or learning bias [24]. The premise is that it may lead to a more robust model, capable of better generalisation [6]. Three arguments support this premise.

1) **Data amplification:** Even though the separate tasks share the same input features, there are more training signals when compared to the single-task setting. Training signals in this context refer to the loss induced by the prediction of each task.
2) **Representation bias:** Multitask learning introduces an inductive bias that favours a hypothesis (or model) explaining multiple tasks. By using a shared representation trained on all three visual relationship recognition tasks, we anticipate improved recognition of relationship triplets.
3) **Regularisation:** Training signals for different tasks have different noise patterns [6]. As a result, the learning procedure is likely to be regularised by the aggregation of multiple noise patterns.

In the case of visual relationship recognition a single network with multiple output vectors can be used, instead of multiple networks each performing a single task. For our experiments we use the convolutional base of ResNet-18, add two trainable, 2,048-dimensional fully-connected layers, and then split the network into three parts. Each part has its own additional 2,048-dimensional layer and a softmax output over the subjects, predicates and objects, respectively. Fig. 4 illustrates this multitask architecture. The first two fully-connected layers are thus shared and may learn effectively from the training signals of the three different tasks. The network can be trained to minimise the sum of cross-entropy losses over the three output vectors, by means of mini-batch gradient descent.

In multitask learning it is common to define a main task together with less important auxiliary tasks. For visual relationship recognition one may want to regard each of the three tasks equally important. However, when using mini-batch construction as described in section III-B, we have to sample the n classes from a single task's vocabulary, at every training iteration, and then use the triplets from the complete labels of the training samples in the mini-batch. In section V we explore how performance of the multitask model changes depending on which task's vocabulary is used for class-selective mini-batch construction.

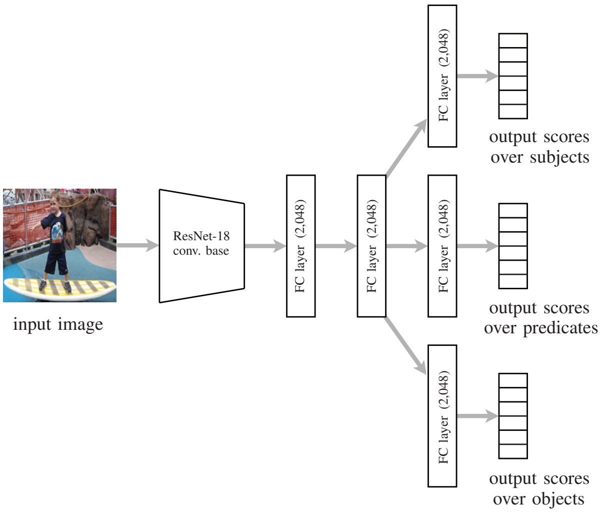


Fig. 4. In the multitask setting, a single model learns to output three score vectors over the subject labels, predicate labels and object labels.

IV. Implementation

This section provides a description of how the various models introduced in the previous section were implemented and trained. The dataset and evaluation metrics used in the experiments are also discussed.

A. Model training

All models are implemented in the PyTorch framework [25]. For standard mini-batching a batch size of 300 is used. For class-selective mini-batch construction we choose $n = 50$ (the number of classes to select per mini-batch) and $m = 6$ (the number of instances to sample per selected class).

There could be a trade-off in performance between the number of classes and sampled instances per class, but informal experimentation showed no significant difference in performance (which is somewhat surprising, although it is possible that effects average out over multiple mini-batches). Gradient descent optimisation is performed by using Adam [26] with a learning rate scheduler that decreases the learning rate every 8 training iterations. The parameter that controls the rate of this decrease is left as the default value. All model training is performed on a single NVIDIA GeForce RTX 2070.

B. Dataset

Models are trained and evaluated on the VRD dataset of Lu *et al.* [4]. It contains 5,000 images and a total of 37,987 visual relationship instances (triplets). Each predicate is an action verb (e.g. `kick`), a non-action verb (e.g. `wear`), a spatial relationship (e.g. `on top of`), a preposition (e.g. `with`), or comparative (e.g. `taller than`). Example images are given in Fig. 5. The semantic ambiguity in ground truth labels mentioned earlier is apparent in some of these. For instance, `taller than` in the fourth example can be replaced by `next to` and still be semantically correct.

The data is split into a training set and a test set. In an effort to ensure representativeness in both sets we consider each predicate label i and split the subset of triplet instances that contain i as a predicate into 80% training data and 20% test data. We base the split on the predicates, since there are fewer samples per class in the tail of the distribution over predicates (as demonstrated below). If the split is based on another element, there is a risk that some predicates may be underrepresented in either the training set or the test set.

There are 100 labels shared between subjects and objects, and 70 labels for predicates, for a total of 700,000 possible (subject, predicate, object) triplet labels. We note that our training set contains only 15,448 unique triplets. However, the manner in which the models are set up to output subject, predicate and object labels separately, potentially enables the recognition of triplets never seen during training.

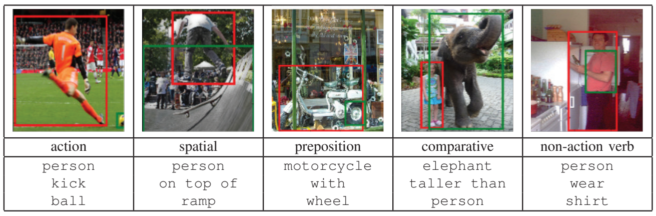


Fig. 5. Examples of the five categories of predicates in the VRD dataset. Green and red bounding boxes are around subjects and objects, respectively.

The long-tailed nature alluded to previously exists in this dataset not only at the relationship triplet level, but also at the level of subjects, predicates and objects, as shown in Fig. 6.

C. Evaluation metrics

Performance of the various models are evaluated first in terms of predicting each of the three elements of a visual relationship, and then in terms of predicting the triplet as a whole (when prediction scores from the three separate tasks are combined).

A standard metric for visual relationship recognition is the recall at k, abbreviated as R@k and sometimes called the top-k accuracy. It measures the percentage of times the correct label occurs in the top k predictions (if ordered by output scores). For the tasks of predicting individual elements, i.e. looking only at the output over subjects, or over predicates or over objects, R@1 and R@3 will be measured on the test set. For the task of predicting the full (subject, predicate, object) triplet, we will measure R@50 and R@100 which seem to be standard practice for a label set of this size [4], [7], [9], [11]. It should be kept in mind that there are 700,000 possible triplets that can be predicted. We found that a random classifier yields an R@100 of approximately 0.026% on this particular non-uniform test set.

In order to evaluate how effectively each model deals with the many rare classes in the tail of the data distribution, the mean per-class accuracy (MPCA) will also be measured on the test set. This metric effectively ignores class imbalance. It will be used only to evaluate the prediction of single elements (subjects, predicates or objects), and not the prediction of full triplets. The large number of possible triplets and the fact that relatively few of them appear in the test set make MPCA less informative in that setting.

For an indication of how the models fare on the rare classes only, a subset of the test set is constructed by keeping only those triplets for which the subject, predicate and object each has fewer than 1,000 instances across the full dataset (refer to Fig. 6). We use counts over the full dataset merely as a proxy for rarity, and remind the reader that elements in the training set are distributed similarly to those in the full set.

V. Experimental results

This section consists of four parts: (A) quantitative results where the performance metrics mentioned above are reported for various versions of the model; (B) a behaviour analysis where a richer investigation into the performance of visual relationship recognition models is introduced; (C) qualitative results where a few example outputs are discussed; and (D) a sensitivity analysis of the assumption that bounding boxes around object pairs are available.

A. Quantitative evaluation

Results from the various models for the three tasks of predicting the subject, the predicate and the object over all the samples in the test set are presented in Table I.

The MPCA values show that class-selective mini-batch construction offers a significant improvement in performance on the long tail-end of each individual task, but only if mini-batches are constructed according to those same tasks. Fig. 7 demonstrates this, in that models with mini-batch construction (the orange bars) perform consistently better across all models than their standard mini-batch counterparts. Class-selective mini-batch construction ensures a roughly uniform label distribution for a particular task and as a result, MPCA is improved. However, mini-batch construction based on a different task seems to reduce MPCA for the subject and object tasks. It is not clear how mini-batch construction based on labels from one task influences the distribution of other tasks, making it difficult to explain the poorer performance.

Relatively lower accuracies from all models for the prediction of predicates verify the suspicion that predicates are harder to recognise visually, possibly due to the greater diversity in their visual representation. The interaction between objects is also in a sense abstract, and the manner in which pre-trained models have been fitted to object classification datasets may prevent effective transfer learning for predicates. Models can be trained from scratch to test this hypothesis, but it will require a dataset larger than VRD.

The results in Table I also indicate higher R@1 and R@3 scores for models trained with standard mini-batching compared to those that implement class-selective mini-batch construction. This can be seen clearly in Fig. 7 where the bars corresponding to standard mini-batching are consistently higher. There seems to be a trade-off: class-selective mini-batch construction contributes to better generalisation on the many rare classes at a cost of accuracy on the small number of dominant classes. Moreover, mini-batch construction with respect to the object labels deals with this trade-off better than in other tasks, as is evident from the smaller difference between standard mini-batching and mini-batch construction in Fig. 7. This could be due to the fact that objects are easier to recognise than predicates, and suffer from a less severe long-tailed distribution than the subjects. The severity of that long tail in the distribution of the subject labels induces interesting behaviour that will be discussed in section V-B.

Furthermore, it is noted that multitask learning does not seem to significantly improve or worsen mean per-class accuracy in the prediction of individual elements. Generally speaking, it is not yet clear under which circumstances a multitask model will improve performance but there are arguments suggesting that more uniform label distributions in the auxiliary tasks might be preferred for multitask learning to be

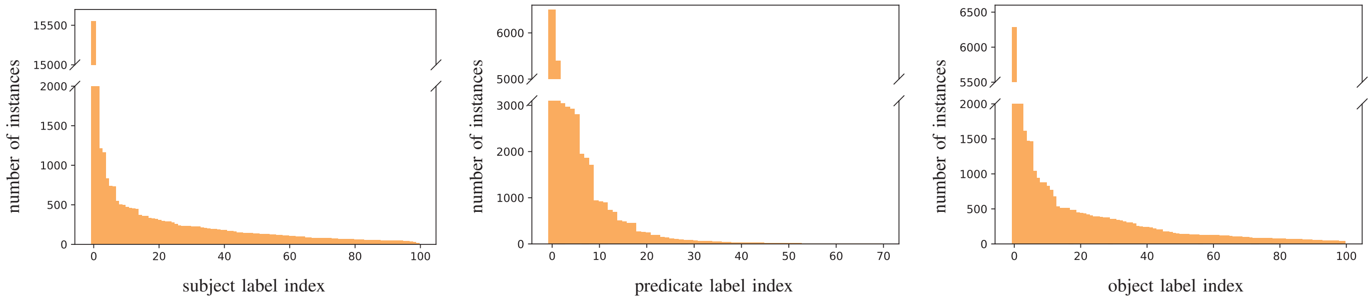


Fig. 6. Plots of the number of relationship instances containing each subject label, predicate label and object label, across the entire VRD dataset.

TABLE I
QUANTITATIVE TEST RESULTS FROM VARIOUS VERSIONS OF OUR MODELS, ON PREDICTING SINGLE ELEMENTS OF VISUAL RELATIONSHIPS.

Model	Description	Predicting the subject			Predicting the predicate			Predicting the object		
		MPCA	R@1	R@3	MPCA	R@1	R@3	MPCA	R@1	R@3
ST-SB	single-task, standard mini-batching	19.09	53.29	73.63	4.51	31.96	56.77	28.92	40.23	68.17
ST-BC-S	single-task, batch construction from subject labels	33.13	16.14	39.39	4.13	19.26	41.36	22.44	38.20	63.74
ST-BC-P	single-task, batch construction from predicate labels	16.70	49.55	68.86	17.01	10.78	31.72	25.20	34.99	61.50
ST-BC-O	single-task, batch construction from object labels	16.67	52.66	71.43	5.24	27.93	51.39	40.72	26.62	50.58
MT-SB	multitask, standard mini-batching	19.96	53.44	74.62	4.74	32.24	57.12	28.34	40.03	68.94
MT-BC-S	multitask, batch construction from subject labels	32.83	17.37	43.41	4.09	19.35	40.70	22.46	38.46	64.92
MT-BC-P	multitask, batch construction from predicate labels	17.18	50.26	70.95	17.54	12.71	32.39	26.24	35.59	62.65
MT-BC-O	multitask, batch construction from object labels	17.52	53.05	72.08	6.27	28.34	52.06	40.60	27.33	51.91

effective [27]. In our case, the multitask models do provide similar performance to the multiple single-task models, which is useful if there are limitations on model size and complexity. Reduced model capacity can also act as a form of regularisation.

Table II lists results from the different models predicting full (subject, predicate, object) triplets. R@50 and R@100 on the test set are reported, as is standard in the literature, and we remind the reader that there are 700,000 possible classes in this case and a random classifier would get an R@100 of about 0.026%. The results are quite similar to related work, but since we focus only on the labelling of visual relationships, and not on the localisation of individual objects, a direct comparison would not mean much. Sensitivity to the tight bounding box assumption is evaluated in section V-D.

As before, standard mini-batching produces better R@50 and R@100 compared to class selective mini-batch construction. However, when focusing only on the long tail-end of the distribution (as explained at the end of section IV-C), we find that mini-batch construction does offer an improvement.

TABLE II
QUANTITATIVE TEST RESULTS FROM THE MODELS, ON PREDICTING FULL VISUAL RELATIONSHIP TRIPLETS.

Model	Predicting the full triplet			
	R@50	R@100	Tail R@50	Tail R@100
ST-SB	49.18	58.18	13.10	17.74
ST-BC-S	23.87	30.84	20.96	27.82
ST-BC-P	31.79	42.10	16.93	23.58
ST-BC-O	40.66	48.58	18.95	24.59
MT-SB	50.27	59.69	12.50	18.95
MT-BC-S	24.95	32.37	19.35	27.21
MT-BC-P	33.56	44.08	17.94	26.41
MT-BC-O	41.83	49.47	20.76	26.20

One may postulate that the predicate is most representative of the visual relationship, but it appears that mini-batch construction with the object labels is a better strategy. The ResNet-18 layers might have an influence here, since they were pre-trained for object classification and thus potentially less suited for the more abstract concept of a predicate. With that in mind, one may expect to gain a similar benefit from subject-based mini-batch construction, but as before the severity of the long tail over subject labels hinders learning.

A fundamental difference between the single- and multitask settings is that the latter receives training signals from all three elements of the visual relationship triplet simultaneously. In some sense the multitask model learns to perceive the full visual relationship, yet does not yield significantly better scores compared to the single-task setting. This shows that it is not immediately obvious that multitask learning will give improved metrics, which again corroborates previous findings [27].

A major limitation of these quantitative evaluations is that they compare model predictions to a particular ground truth label, despite the fact that visual relationships are often ambiguous and a prediction different from the ground truth may therefore not be completely wrong.

B. Behaviour analysis

The evaluations above were concerned with whether the correct (ground truth) triplet occurs in the top 50 or top 100 predicted triplets. If this is not the case, there are a few specific outcomes that may still be of interest. To gain deeper insights into the behaviour of the models we consider five mutually exclusive events, each predicated on all preceding events not taking place. The following list describes each event, where the predicate element is coloured according to whether it is correct (green) or incorrect (red).

1) **subject**, **predicate**, **object**: The correct visual relationship triplet occurs in the top 50 predictions. This is what is picked up when R@50 is determined.
2) **subject**, **predicate**, **object**: Event 1 does not occur, but the correct subject and object appear together in the top 50 predictions with an incorrect predicate.
3) **object**, **predicate**, **subject**: Events 1 and 2 do not occur, but the three correct elements appear together in the top 50 with the subject and object swapped.
4) **object**, **predicate**, **subject**: Events 1, 2 and 3 do not occur, but the correct subject and object appear together in the top 50, swapped and with an incorrect predicate.
5) **other**: Events 1, 2, 3 and 4 do not occur, that is, the correct subject and object do not appear together (in order or swapped) in the top 50 predictions.

Fig. 8 shows the percentages of these event occurring for all the different models across the test set. Despite significant differences in design, all models find the correct subject and object with an incorrect predicate (event 2) at a similar rate. This once again suggests that predicting the predicate is more challenging, even when mini-batch construction ensures a uniform label distribution over the predicates. The same happens when the subject and object labels are confused (event 4): the correct predicate is not found in the top 50 predictions at a similar rate when compared to standard mini-batching. Perhaps there is not enough visually discriminative information in a predicate for a vision-based classifier to be effective. Considering the many ambiguous predicate labels in the VRD dataset, it is also likely that incorrect but semantically sensible predicates are being predicted. Quantitatively evaluating semantic similarity would require significant manual labour or the clever use of a language model. The latter could make for fruitful future research.

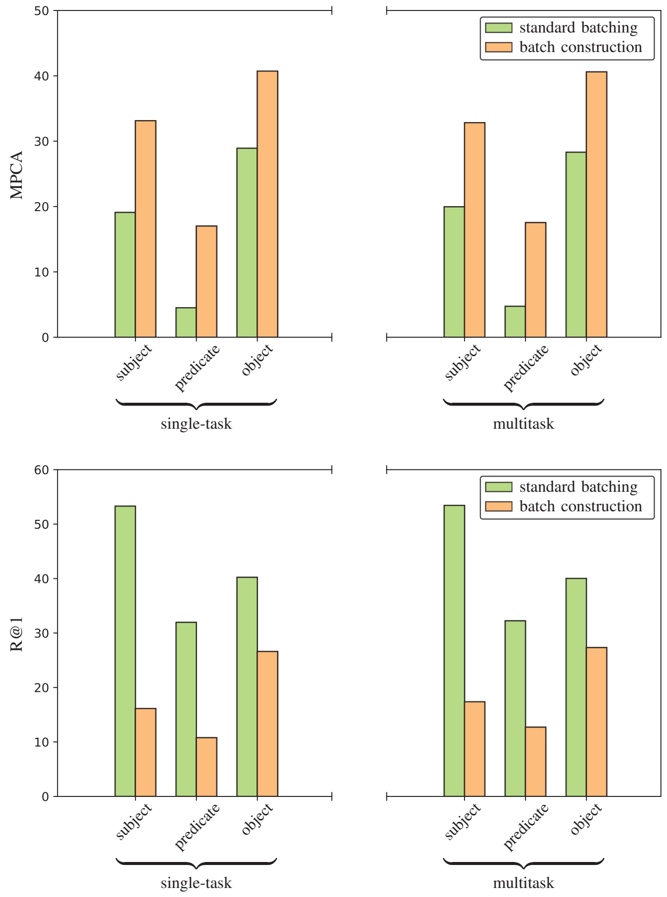


Fig. 7. MPCA (top) and R@1 (bottom) in predicting single relationship elements, with standard mini-batching vs class-selective mini-batch construction with respect to the same element that is being predicted. Bars that are connected can be compared directly.

Subject-based mini-batch construction seems to confuse the object and subject labels at a consistently greater rate than all other models. This may again be attributed to the severity of the long tail in the distribution over subject labels. The dominant subject class has around 15,000 samples, compared to the dominant object and predicate labels that have around 6,000 each. Performing subject-based mini-batch construction lessens the strong bias that exists in the subject labels, and as a result the models may confuse the subjects and objects.

When a model confuses the subject and object, missing the correct predicate (event 4) happens more frequently than picking up the correct predicate (event 3). This might be an indication that models can swap the subject and object labels and reverse the predicate. For example, it is possible to have (`giraffe`, `taller than`, `person`) as the ground truth but have (`person`, `shorter than`, `giraffe`) as an acceptable answer that would be regarded incorrect by a quantitative evaluation. It remains difficult to say precisely what proportion of outcomes that are categorised by the given events should be classified as semantically correct predictions. An option here would be to construct a mapping of semantically similar predicates for event 2, predicates with a reflective property for event 3, and a mapping of inverse predicates for event 4. The construction of such mappings can potentially be automated with the aid of a computational language model based on a lexical database like WordNet [28].

The analysis in this section provides a means of measuring specific kinds of misclassifications which can potentially be semantically correct. Measuring performance with such semantic ambiguities in mind can provide more reliable results and comparisons, and more insight into model behaviour. Moreover, a single human-generated label may be unable to capture the rich semantics necessary to describe visual relationships and therefore be unable to provide an optimal training signal. Again, it seems natural to incorporate a language model in order to address this problem.

In light of the ambiguities that exist in visual relationships, comparing model predictions to a particular ground truth label does not offer the complete picture. The next section provides further insight into the behaviour of our various models by means of a qualitative evaluation.

C. Qualitative evaluation

Fig. 9 shows a number of test image samples and the top five predicted triplets from four of the models. The first three

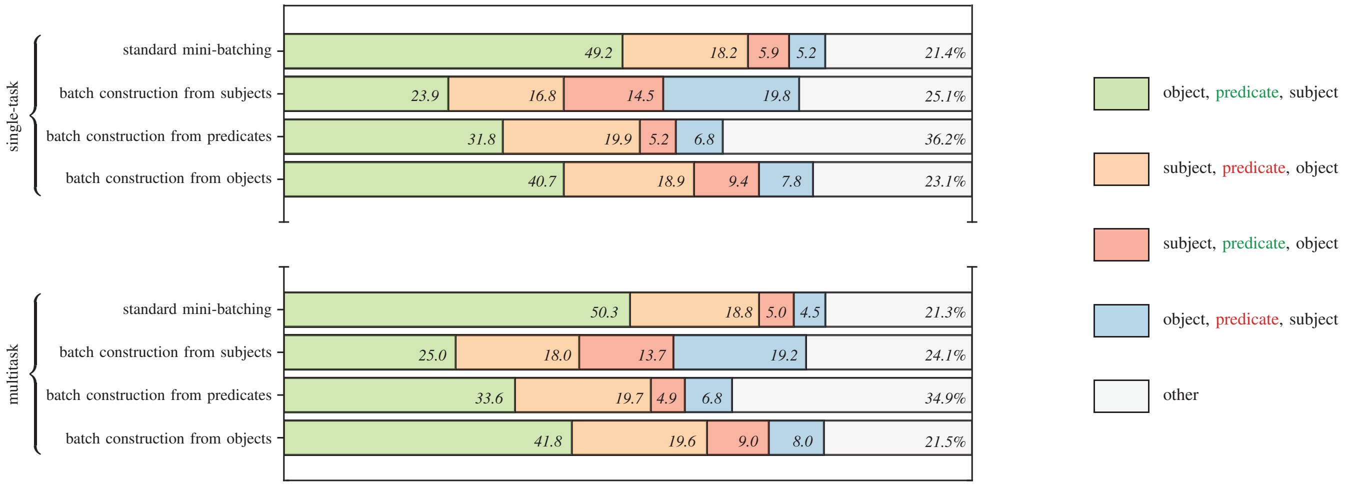


Fig. 8. Percentage of occurrences of the five events described in section V-B, over the test predictions of the various models. Green and red in the legend indicate whether the predicate is correctly or incorrectly classified.

examples were chosen randomly from those for which the ST-SB model returned the correct visual relationship in its top five predictions. The other three were chosen randomly from those for which the ST-SB model did not return the correct relationship in its top five predictions. Here mini-batch construction based only on the object labels is highlighted, since it outperformed those based on subjects and predicates.

For the second example shown in Fig. 9, with ground truth label (`giraffe`, `taller than`, `giraffe`), ST-BC-O gets the subject and object right but confidently predicts `in front of` as the predicate; perhaps a forgivable error. Similarly sensible errors can be seen throughout the examples in Fig. 9, and further demonstrate the ambiguities present in visual relationships. It is interesting to note that the `behind` and `in front of` predicates, although appearing in the top five predictions for the giraffe example, have very different confidence scores. This behaviour is undesirable and further motivates the inclusion of multi-modal semantics in the modelling process.

The entity `person` is the dominating subject class, and is predicted correctly in almost all applicable cases shown in Fig. 9. Some predicted relationships for the (`person`, `on`, `horse`) example may seem nonsensical, like (`person`, `behind`, `horse`) or (`person`, `next to`, `person`). There is actually another person in the background of the image, and the models may be recognising the overrepresented `person` class despite a lack of sufficient visual cues.

Models trained with class-selective mini-batch construction appear to make predictions with relatively high confidence scores. There are 700,000 normalised confidence scores, so high scores in the top five predictions imply exceptionally low scores for the remaining 699,995 relationships. It is interesting that the confidence scores are this heavily skewed under an arguably more uniform training data distribution.

For the bottom three examples in Fig. 9 we see some examples where the models swap the positions of the subject and object labels and misclassify the predicate, as investigated in section V-B. This is seen in the example where three of the models predict (`tree`, `next to`, `bear`) instead of the ground truth (`bear`, `adjacent to`, `tree`). These misclassifications may again be forgivable, especially since `adjacent to` is one of the more obscure predicates in the dataset.

In the first example in the bottom row of Fig. 9, the models return predicates other than the ground truth `on`, despite `on` being the dominating predicate in the dataset. This may be due to the strong visual cues in favour of interactions between the person and their items of clothing, rather than `skateboard` which is less common in the dataset.

The predicates `feed` and `adjacent to` are rare tail-end predicates that are misclassified even under the class-selective mini-batch construction strategies. Nevertheless, for those examples many of the top predicted relationships do seem to be in line with the content of the images. Perhaps this kind of error can be mitigated with the inclusion of multi-modal semantics in the training procedure.

D. Bounding box perturbation

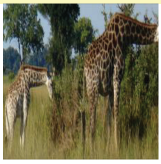
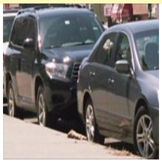
All the models developed in this paper assume knowledge of tight bounding boxes around object pairs in images. The final experiment is to test the sensitivity of one of the models to perturbations on the bounding box coordinates, for an indication of how the model might respond in practice when combined with an automatic object detector.

The intersection over union (IOU) can be used to measure the degree of perturbation. For two bounding boxes B_1 and B_2, the IOU is defined as

$$\mathrm{IOU} = \frac{\mathrm{area}(B_1 \cap B_2)}{\mathrm{area}(B_1 \cup B_2)}, \tag{1}$$

and gives the percentage overlap between the two bounding boxes. IOU, also referred to as the Jaccard index, is a standard metric used in the evaluation of object detectors.

We add varying levels of Gaussian noise to the four corners of each bounding box in the VRD dataset. Specifically, we

Model	person, on, horse 		giraffe, taller than, giraffe 		car, behind, car 	
ST-SB	person, on, horse	12.0	giraffe, taller than, giraffe	25.1	car, behind, car	20.6
	person, ride, horse	7.0	giraffe, in front of, giraffe	20.8	car, in front of, car	13.1
	person, wear, horse	5.3	giraffe, next to, giraffe	9.5	car, next to, car	7.2
	person, has, horse	5.2	giraffe, above, giraffe	7.6	car, on, car	4.7
	person, on, person	3.1	giraffe, behind, giraffe	7.2	car, near, car	4.3
ST-BC-O	person, on, horse	18.7	giraffe, in front of, giraffe	98.6	car, next to, car	8.5
	person, has, horse	11.8	giraffe, taller than, giraffe	0.4	car, behind, car	7.8
	person, wear, horse	7.7	giraffe, behind, giraffe	0.4	car, in front of, car	5.0
	person, in front of, horse	4.3	giraffe, next to, giraffe	0.1	car, next to, van	3.8
	person, next to, person	3.7	giraffe, beside, giraffe	0.1	car, has, car	3.7
MT-SB	person, wear, horse	9.3	giraffe, taller than, giraffe	45.4	car, behind, car	14.1
	person, on, horse	6.8	giraffe, in front of, giraffe	18.9	car, in front of, car	11.6
	person, wear, person	3.4	giraffe, next to, giraffe	8.6	car, next to, car	7.2
	person, behind, horse	3.1	giraffe, behind, giraffe	7.3	car, on, car	4.6
	person, has, horse	2.6	giraffe, under, giraffe	2.6	car, near, car	3.4
MT-BC-O	person, on, horse	13.2	giraffe, in front of, giraffe	92.5	car, behind, car	8.7
	person, above, horse	12.0	giraffe, taller than, giraffe	6.0	car, in front of, car	6.9
	person, behind, horse	6.3	giraffe, behind, giraffe	0.9	car, behind, van	5.3
	person, ride, horse	5.3	giraffe, next to, giraffe	0.3	car, in front of, van	4.3
	person, has, horse	4.8	giraffe, beside, giraffe	0.07	car, on, car	4.2

Model	person, on, skateboard 		bear, adjacent to, tree 		person, feed, elephant 	
ST-SB	person, wear, person	11.8	bear, next to, grass	3.6	person, above, street	4.3
	person, wear, shirt	10.5	bear, on, grass	3.3	person, on, street	4.1
	person, wear, skateboard	10.0	bear, next to, person	2.5	person, under, street	3.0
	person, wear, shoes	5.4	bear, next to, tree	2.3	sky, above, street	1.7
	person, wear, pants	4.4	bear, on, person	2.3	sky, on, street	1.6
ST-BC-O	person, wear, skateboard	25.6	tree, next to, bear	78.7	person, under, elephant	16.4
	person, on, skateboard	10.0	bear, next to, bear	11.2	person, in front of, elephant	16.0
	person, has, skateboard	9.6	tree, near, bear	2.7	person, above, elephant	10.0
	person, ride, skateboard	5.2	person, next to, bear	0.7	person, near, elephant	4.7
	person, wear, shoes	3.5	tree, right of, bear	0.6	person, behind, elephant	4.1
MT-SB	person, wear, shirt	15.5	bear, next to, grass	4.3	person, on, street	4.7
	person, wear, person	9.6	bear, next to, bear	3.5	person, under, street	3.9
	person, wear, skateboard	6.9	bear, on, grass	3.1	person, above, street	3.4
	person, wear, shoes	6.1	bear, on, bear	2.6	person, on, person	2.4
	person, wear, pants	4.1	bear, behind, grass	2.5	person, under, person	1.9
MT-BC-O	person, wear, skateboard	20.0	tree, next to, bear	87.7	person, in front of, elephant	7.4
	person, wear, shoes	14.0	bear, next to, bear	2.7	person, near, elephant	6.9
	person, wear, helmet	12.0	tree, behind, bear	0.8	person, under, elephant	5.1
	person, has, skateboard	3.8	tree, beside, bear	0.8	person, on, elephant	3.4
	person, wear, pants	3.7	tree, next to, grass	0.7	person, above, elephant	2.4

Fig. 9. Top five visual relationship predictions on example test images, with confidence scores, as returned by four of the models. The ground truth label is shown above each image.

fix the mean of a Gaussian distribution from which noise is sampled at zero and use the following set of standard deviations: $\{0, 5, 10, 15, 20, 45, 55, 65, 75\}$. These standard deviations result in intersection over union measures from 100% down to around 50%, which is a standard threshold used in the object detection literature.

When bounding boxes are perturbed, it is possible that the resulting box is outside the image. If after 50 random perturbations any pixels of the resulting bounding box are not within image bounds, the ground truth bounding box for that sample is left as is. We show the resulting mean IOU between the original bounding boxes of the entire test set and their random perturbations as percentages in Fig. 10. It may happen that perturbed bounding boxes remain within the image bounds but no longer contain the visual relationship at all, so we expect R@50 scores to decrease with increased noise levels.

Results are shown for the multitask standard mini-batching

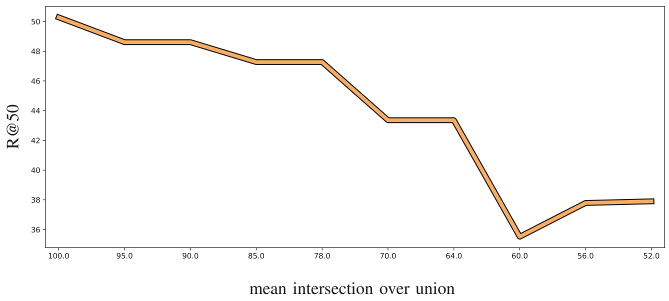


Fig. 10. R@50 on the VRD test set with perturbed bounding boxes, from the multitask with standard mini-batching model. Perturbation IOU ranges from 100% to 52%.

model, since in the previous experiments it performed best on average. Note that the model tested previously is used here without retraining it on the perturbed bounding boxes. As expected there is a decline in R@50 scores, but the decline occurs at a fairly low rate. Performance even remains roughly constant over different noise levels, which indicates some robustness under perturbations of the bounding boxes.

VI. Conclusion

We investigated the potential of class-selective mini-batch construction and multitask learning for the task of visual relationship recognition; a challenging task in computer vision given the large number of possible relationships as well as a typical long-tailed distribution over those relationships.

The proposed mini-batch construction strategy seems to improve performance on the tail of the data distribution, but at the cost of performance on the small number of dominating classes. Multitask learning does not seem to improve or impede performance when compared to single-task learning, but provides other benefits such as a reduced model capacity. Results also suggest that it can be more difficult to model and recognise the predicate of a relationship, and that current pre-trained models may not be suitable for that task.

A novel evaluation approach was introduced to analyse the frequency at which three specific types of errors occur in the top 50 predictions of a full relationship triplet. This approach can be viewed as an extension to the existing recall-at-k metric, and provides deeper insight into the behaviour of models. Further analysis on those types of errors can lead to more effective models. We demonstrated a number of semantically sensible misclassifications through a few test examples.

The sensitivity of the best performing model to perturbations in the bounding boxes around pairs of interacting objects was also investigated, and some level of robustness to such perturbations was found.

It is also possible to further extend the evaluation methods. R@50 is limited in that it relies strongly on hard comparison with ground truth labels, but misclassifications can often be semantically correct due to the ambiguities inherent in visual relationships. It may therefore be useful to design additional metrics that can measure semantic similarity. WordNet [28] is a lexical database of so-called synsets that group nouns, verbs, adjectives and adverbs that express the same concept. Model outputs can then automatically be compared to synsets for better evaluation. The Visual Genome dataset [5] already contains mappings from class labels to WordNet synsets, but this information is not yet commonly used in model evaluation.

Another extension of the work in this paper involves incorporating semantics directly into the modelling process. To this end, a language model can be used to mitigate the inherent ambiguity in visual relationship labels. Lu *et al.* [4] employ a language model to obtain a visual relationship embedding space, but there may be other ways to do so. For example, a language model can be used to re-score the outputs of a classifier and thus encode semantics (as is commonly done in speech recognition). In this way, model confidence in relationship triplets that would be unlikely from a semantic point of view, such as (`giraffe`, `drive on`, `umbrella`), can be suppressed. Techniques such as N-best list re-scoring [29], [30], and lattice re-scoring [31], [32] can be considered as potential re-scoring strategies.

References

[1] Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: IEEE Conference on Computer Vision and Pattern Recognition; 2016. p. 779–788.

[2] Zhang S, Wen L, Bian X, Lei Z, Li S. Single-shot refinement neural network for object detection. In: IEEE Conference on Computer Vision and Pattern Recognition; 2018. p. 4203–4212.

[3] Johnson J, Krishna R, Stark M, Li LJ, Shamma D, Bernstein M, et al. Image retrieval using scene graphs. In: IEEE Conference on Computer Vision and Pattern Recognition; 2015. p. 3668–3678.

[4] Lu C, Krishna R, Bernstein M, Fei-Fei L. Visual relationship detection with language priors. In: European Conference on Computer Vision; 2016. p. 852–869.

[5] Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, et al. Visual Genome: connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision. 2017;123(1):32–73.

[6] Caruana R. Multitask learning. Machine Language. 1997;28(1):41–75.

[7] Zhang H, Kyaw Z, Chang SF, Chua TS. Visual translation embedding network for visual relation detection. In: IEEE Conference on Computer Vision and Pattern Recognition; 2017. p. 5532–5540.

[8] Zhang J, Kalantidis Y, Rohrbach M, Paluri M, Elgammal A, Elhoseiny M. Large-scale visual relationship understanding. In: AAAI Conference on Artificial Intelligence. vol. 33; 2019. p. 9185–9194.

[9] Xu D, Zhu Y, Choy CB, Fei-Fei L. Scene graph generation by iterative message passing. In: IEEE Conference on Computer Vision and Pattern Recognition; 2017. p. 5410–5419.

[10] Zellers R, Yatskar M, Thomson S, Choi Y. Neural motifs: scene graph parsing with global context. In: IEEE Conference on Computer Vision and Pattern Recognition; 2018. p. 5831–5840.

[11] Newell A, Deng J. Pixels to graphs by associative embedding. In: Advances in Neural Information Processing Systems; 2017. p. 2171–2180.

[12] Qi S, Wang W, Jia B, Shen J, Zhu SC. Learning human-object interactions by graph parsing neural networks. In: European Conference on Computer Vision; 2018. p. 401–417.

[13] Yang J, Lu J, Lee S, Batra D, Parikh D. Graph R-CNN for scene graph generation. In: European Conference on Computer Vision; 2018. p. 670–685.

[14] Woo S, Kim D, Cho D, Kweon IS. LinkNet: relational embedding for scene graph. In: Advances in Neural Information Processing Systems; 2018. p. 560–570.

[15] Dai B, Zhang Y, Lin D. Detecting visual relationships with deep relational networks. In: IEEE Conference on Computer Vision and Pattern Recognition; 2017. p. 3076–3086.

[16] Wang G, Luo P, Lin L, Wang X. Learning object interactions and descriptions for semantic image segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition; 2017. p. 5859–5867.

[17] Chao YW, Liu Y, Liu X, Zeng H, Deng J. Learning to detect human-object interactions. In: IEEE Winter Conference on Applications of Computer Vision; 2018. p. 381–389.

[18] Yin G, Sheng L, Liu B, Yu N, Wang X, Shao J, et al. Zoom-Net: mining deep feature interactions for visual relationship recognition. In: European Conference on Computer Vision; 2018. p. 322–338.

[19] Gkioxari G, Girshick R, Dollár P, He K. Detecting and recognizing human-object interactions. In: IEEE Conference on Computer Vision and Pattern Recognition; 2018. p. 8359–8367.

[20] Li Y, Ouyang W, Wang X, Tang X. ViP-CNN: visual phrase guided convolutional neural network. In: IEEE Conference on Computer Vision and Pattern Recognition; 2017. p. 1347–1356.

[21] Kuznetsova A, Rom H, Alldrin N, Uijlings J, Krasin I, Pont-Tuset J, et al. The Open Images dataset V4. International Journal of Computer Vision. 2020:1–26.

[22] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition; 2016. p. 770–778.

[23] Bottou L, Curtis FE, Nocedal J. Optimization methods for large-scale machine learning. SIAM Review. 2018;60(2):223–311.

[24] Mitchell TM. The need for biases in learning generalizations. Technical Report CBM-TR-117, Rutgers University; 1980.

[25] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: an imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems; 2019. p. 8024–8035.

[26] Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv preprint arXiv:14126980. 2014.

[27] Martínez Alonso H, Plank B. When is multitask learning effective? Semantic sequence prediction under varying data conditions. In: Conference of the European Chapter of the Association for Computational Linguistics; 2017. p. 44–53.

[28] Miller GA. WordNet: a lexical database for English. Communications of the ACM. 1995;38(11):39–41.

[29] Stolcke A, Konig Y, Weintraub M. Explicit word error minimization in N-best list rescoring. In: European Conference on Speech Communication and Technology; 1997. p. 163–166.

[30] Verhasselt J, Dercks H. N-best list rescoring in speech recognition. Acoustical Society of America Journal. 2010;128(6):3828.

[31] Sundermeyer M, Tüske Z, Schlüter R, Ney H. Lattice decoding and rescoring with long-span neural network language models. In: Conference of the International Speech Communication Association; 2014. p. 661–665.

[32] Kumar S, Nirschl M, Holtmann-Rice D, Liao H, Suresh AT, Yu F. Lattice rescoring strategies for long short-term memory language models in speech recognition. In: IEEE Automatic Speech Recognition and Understanding Workshop; 2017. p. 165–172.

Shane Josias was born in Atlantis, Cape Town, South Africa in 1995. He received his BScHons and MSc degrees in applied mathematics from Stellenbosch University, Stellenbosch, South Africa in 2017 and 2020 respectively.

From 2015 to 2018, he participated in a number of internships at a telecommunications and information security consulting firm. From 2020 to 2021, he was a Software Engineer in telecommunications, focusing on complex events processing and database integration. He currently holds a position as Junior Lecturer in the Department of Mathematical Sciences at Stellenbosch University, where he fulfills teaching duties while pursuing his PhD degree. His current research is concerned with learning robust representations of data in the context of computer vision.

Willie Brink received BScHons and MScEngSci degrees in applied mathematics from Stellenbosch University, Stellenbosch, South Africa in 2003 and 2005 respectively, and a PhD degree in computer science from Sheffield Hallam University, Sheffield, England in 2008.

In 2008, he held a postdoctoral position in robotics at the Council for Scientific and Industrial Research, Pretoria, South Africa. In 2009, he was appointed as Lecturer of applied mathematics in the Department of Mathematical Sciences at Stellenbosch University. He was promoted to Senior Lecturer in 2012, and to Associate Professor in 2020. His research is in applied machine learning, with a current focus on representation learning and low-resource computer vision.

Prof. Brink is a member of the Pattern Recognition Association of South Africa and the Computer Vision Foundation. He is a fellow of the CSIR/SU Centre for Artificial Intelligence Research, and a co-founder of the Deep Learning Indaba.

CONTENTS PAGE