Partially-Supervised Novel Object Captioning Using Context from Paired Data

Shashank Bujimalla[1]
shashankbvs@gmail.com

Mahesh Subedar[2]
mahesh.subedar@intel.com

Omesh Tickoo[2]
omesh.tickoo@intel.com

[1] Intel Corporation,
   USA
[2] Intel Labs,
   USA

Abstract

In this paper, we propose an approach to improve image captioning solution for images with novel objects that do not have caption labels in the training dataset. We refer to our approach as Partially-Supervised Novel Object Captioning (PS-NOC). PS-NOC is agnostic to model architecture, and primarily focuses on the training approach that uses existing fully paired image-caption data and the images with only the novel object detection labels (partially paired data). We create synthetic paired captioning data for novel objects by leveraging context from existing image-caption pairs. We then create pseudo-label captions for partially paired images with novel objects, and use this additional data to fine-tune the captioning model. We also propose a variant of Self-Critical Sequence Training (SCST) within PS-NOC, called SCST-F1, that directly optimizes the F1-score of novel objects. Using a popular captioning model (Up-Down) as baseline, PS-NOC sets new state-of-the-art results on *held-out* MS COCO out-of-domain test split, i.e., 85.9 F1-score and 103.8 CIDEr. This is an improvement of 85.9 and 34.1 points respectively compared to baseline model that does not use partially paired data during training. We also perform detailed ablation studies to demonstrate the effectiveness of our approach.

1 Introduction

Image captioning solutions generate natural language description (i.e., caption) of the salient visual concepts in a given image. There has been a rapid progress in using deep neural network models to develop automatic image captioning solutions. The datasets that are used to train these models contain paired image-caption data, i.e., each image consists of a set of human annotated ground-truth captions; we also refer to such data as fully paired data from hereon. These datasets are often limited in the number and diversity of visual concepts; for e.g., MS COCO Captions [10], which is the most popular captioning dataset contains only 91 underlying object classes. This could severely hinder the practical application of such models. There has been an increased effort to improve image captioning solutions on novel object classes that do not have paired image-caption annotations in training dataset. These are referred to as Novel Object Captioning (NOC) solutions, and benchmark datasets such as *held-out* MS COCO [13] and *no-caps* [1] have been created to evaluate them.

A bowl a ***cup*** and two controllers on a table.

A bowl a ***bottle*** and two controllers on a table.

Baseline: A green flower in a glass on a table.
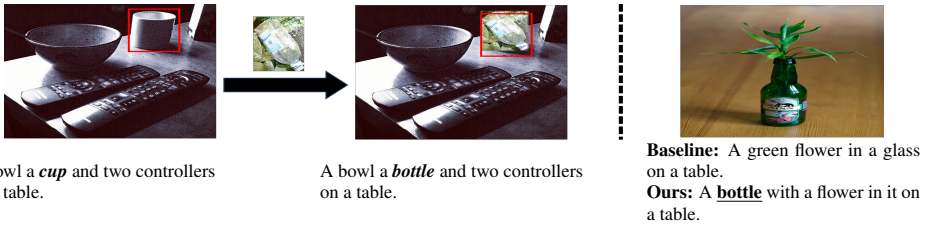**Ours:** A <u>**bottle**</u> with a flower in it on a table.

Figure 1: In our PS-NOC approach, we generate synthetic paired data for novel objects, as shown in 'bottle' example on the left. We then use this additional synthetic data and partially paired data to generate captions for novel objects in unseen images, as shown on the right.

The current NOC solutions broadly belong to two categories. The first category of solutions use specialized model architectures to insert the objects detected in an image into the caption [7, 9, 13, 19, 25, 26, 28]. The second category of solutions use partial supervision to train the models on images with novel objects [4, 8]. The partially-supervised solutions use images from object detection datasets, which include a much wider variety of object classes than the captioning datasets, to train the captioning model. For e.g., Open Images [15] contains 600 object classes that includes human annotations for object bounding boxes along with their class labels. These images are referred to as partially paired data from hereon, since their object class labels can be considered as partial ground-truth captions. In this work, we focus on the second category of solutions and propose our Partially-Supervised Novel Object Captioning (PS-NOC) solution that is agnostic to model architecture.

We first generate synthetic paired image-caption data for novel objects. The current partially supervised solutions for NOC, i.e., PS3 [4] and FDM-net [8], use the images and the class labels of partially paired data, but do not use their bounding box annotations. We use these bounding box annotations and generate synthetic images by replacing objects in fully paired images with novel objects. We also generate the corresponding synthetic captions for these synthetic images by direct word replacement of objects in the fully paired captions with novel objects. We refer to this new paired image-caption data as synthetic data, although they are extracted from natural image-caption pairs. An example is shown in Figure 1.

We then propose a model-agnostic three step training technique that uses this synthetic data and partially paired data to train the captioning models for novel objects. In practical real world situations, there exist captioning models that are pre-trained on fully paired data. Our first step mimics this by training a popular captioning model, Up-Down [5], on fully paired data. In the second step, we augment our fully paired data with synthetic data and warm up the captioning model to novel objects; we notice this considerably improves the captioning model performance on images with these novel objects. Finally, in the third step, we use this model to iteratively generate pseudo-label captions for partially paired images and further fine-tune the model using this additional paired data. Figure 1 shows an example novel object caption generated using our approach.

In summary, our main contributions are as follows:

1) We generate synthetic paired image-caption data for novel objects by using object bounding box annotations and using the context from fully paired data. To our knowledge, our work is the first to study the usage of such synthetic paired data with images for NOC.

2) We propose a model-agnostic three step training approach that uses the additional synthetic data and partially paired data to train the captioning model. This includes novel training techniques: i) SCST-F1, a variant of Self-Critical Sequence Training (SCST) [23],

which directly optimizes F1-score of novel objects [13], ii) Pseudo-labeling for NOC, which makes use of partially paired images, considers both novel object inclusion and caption quality, and is easy to integrate into existing captioning model implementations.

Using Up-Down captioning model as baseline, our PS-NOC approach sets new state-of-the-art results on *held-out* MS COCO out-of-domain (i.e., novel object images) test split with 85.9 F1-score and 103.8 CIDEr.

2 Related Work

Many of the current solutions for NOC use specialized model architectures to generate captions that contain novel objects. For e.g., they use two parallel network pipelines [7, 9, 13, 25], neural slot-filling [19], or copying mechanism [26, 28]. There also exist a few partially-supervised solutions, PS3 [4] and FDM-net [8], that use images from object detection datasets, i.e., partially paired images, to train the captioning model.

PS3 [4] uses constrained beam search (CBS) [2] to generate pseudo-label captions for the partially paired images. CBS forces the inclusion of specified tag words, called constraints, into the model generated captions. To generate these tag words for partially paired images, PS3 uses an additionally trained lexical classifier from [13]. PS3 generates these pseudo-label captions at each mini-batch of training and refer to it as online version. To perform this training in a compute-efficient manner, they maintain two copies of the model with tied weights. In our PS-NOC approach, we perform offline version of pseudo-labeling by generating pseudo labels after training the model for several thousands of mini-batch iterations. Since we use synthetic data to warm-up the model to novel objects before performing pseudo-labeling, we notice that our model fine-tuning converges within a few rounds (for e.g., 1 to 4) of offline pseudo-labeling. This makes our approach easier to integrate into any existing captioning model implementations, as we do not need two copies of our model during training. However, our approach is also applicable to online version, if required. Unlike PS3, we do not use the additional lexical classifier and instead use the object detection labels directly as CBS constraints. In PS-NOC, we also generate an additional pseudo-label caption without using CBS to allow prioritizing caption quality over novel object inclusion.

FDM-net [8] uses partially paired images and generates synthetic data containing novel objects. However, they perform object replacement in image feature space and generate synthetic region features by using object detection model. This restricts FDM-net's application to region based captioning models. Our PS-NOC approach to use synthetic images decouples this dependency, making it more generic and applicable to even the advancing captioning models that use the entire image [17, 27]. Our approach also allows us to include humans in the loop to sign-off or discard the synthetic image-caption data generated using such automation. FDM-net also generates synthetic captions, but trains an additional scene-graph based network to generate them. We instead generate synthetic captions by direct word replacement, and then use simple heuristics to remove noise in these captions.

3 Our Approach: PS-NOC

3.1 Synthetic Data Generation

We first generate paired synthetic data, i.e., image-caption pairs, for novel objects. To generate synthetic images, we use the bounding box annotations of novel objects in partially

paired images and replace them into bounding box annotations of candidate objects in fully paired images. We also generate the corresponding synthetic captions for these synthetic images by directly replacing the objects in fully paired captions with novel objects. Examples of synthetic data generated using our approach are shown in Figure 2. More details are as follows.

Candidate object selection: We follow the approach similar to FDM-net [8] to identify the candidate objects in fully paired data for replacement with a novel object. In this approach, validation split images containing novel objects are input to a baseline captioning model that is trained on fully paired data. For each novel object, the number of occurrences of each fully paired object in the generated captions is used to identify the m most likely candidate objects that the model finds similar to the novel object. We generate up to K synthetic images for each novel object, with K/m images for each novel-candidate object pair. We also heuristically try to minimize the number of times each novel object image is used, so that we generate synthetic images with diverse novel object features and background contexts.

Image selection: The sizes of novel objects and candidate objects could vary widely across images. In order to avoid a big change in resolution and aspect ratio when performing object replacement, we only consider images where novel objects and candidate objects have bounding box annotations whose areas do not differ by more than $\Delta A_{max}\%$; we also set a lower limit, A_{min}, on their areas. We set similar constraints AR_{min} and $\Delta AR_{max}\%$ for their aspect ratios, along with an upper limit on aspect ratio AR_{max}. If multiple occurrences of an object are found in each image, we consider all the pairs for replacement and exclude the pairs that do not meet the constraints. For e.g., multiple occurrences of 'cow' are replaced with multiple occurrences of 'zebra' in Figure 2. We observe that including such images with multiple object occurrences into synthetic data is critical for captioning performance. We also exclude any objects whose bounding box annotations completely contain other annotations; this is to avoid cases where additional objects accidentally sneak into the synthetic image along with novel objects, or get removed along with candidate object, thus resulting in a noisy caption.

Synthetic caption processing: When candidate objects get replaced with novel objects, their descriptions in the caption could be incorrect. We use simple heuristics to process these captions. Before the candidate object is replaced with novel object, we remove any occurence of color words (e.g., 'blue') in the caption; we also remove any adjectives or nouns that occur within certain word radius, ΔR_{adj} and ΔR_{noun} respectively, of the candidate object word. For e.g., 'a frosted cake' is converted to 'a pizza' in Figure 2.

We notice that synthetic images and captions generated using our simple heuristics are effective enough to improve the overall NOC results, as shown in Section 4.

3.2   Training Technique

Figure 3 shows the overview of our training technique that consists of three steps, and leverages synthetic and partially paired data.

Step I. Train using fully paired data: We first train our baseline captioning model $M^{(B)}$ with fully paired data.

Step II. Fine-tune using synthetic data: We warm-up our captioning model $M^{(S)}$ to novel objects using our synthetic data in addition to the fully paired data.

Step III. Generate pseudo-labels on partially paired data and fine-tune: In this step, we use our model from Step 2 that considerably improved its generalization to novel objects, and generate pseudo-label captions for partially paired images that contain novel objects.

A group of cows on dirt area with trees in background.
+ Novel Object: zebra
= A group of zebras on dirt area with trees in background.

A blue plate holding a frosted cake and knife.
+ Novel Object: pizza
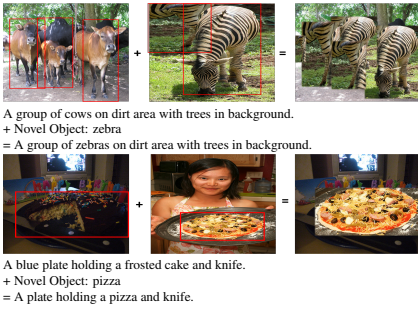= A plate holding a pizza and knife.

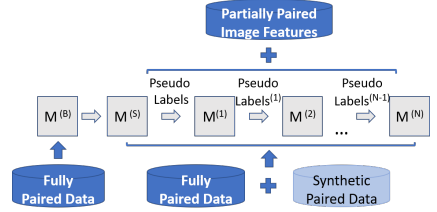Figure 2: Examples of synthetic data generated using our approach.

Figure 3: Overview of our training technique.

This additional partially paired data, on top of fully paired and synthetic data, is used to further fine-tune our model. We repeat Step III a fixed number of times, i.e., N rounds, to generate the final model $M^{(N)}$.

Pseudo-labeling approach: Our pseudo-labeling approach in Step III consists of two key components. 1. We perform offline pseudo-labeling, i.e., we fine-tune the model for several thousands of mini-batch iterations and generate pseudo-labels. This makes it easy to integrate our approach on any existing implementations of captioning models. 2. For each partially paired image, we generate two pseudo-label captions using beam search: one caption using CBS [2], and the other caption without using CBS. We observe that CBS could sometimes negatively impact caption quality (i.e., CIDEr [24] score) while including the novel objects in the caption. On the other hand, we notice that disabling CBS improves the caption quality while missing some novel objects due to the limitations in synthetic data. Using both these captions as pseudo-labels on top of our approach outperforms using CBS caption alone. We provide this ablation study results in Supplemental material.

3.3 SCST-F1

Image captioning models are initially trained using cross-entropy loss, and then using Self-Critical Sequence Training (SCST) [6, 11, 23]. SCST uses policy gradient based RL to directly optimize CIDEr [24], which is a widely accepted captioning quality metric. In general, a NOC dataset is divided into in-domain and out-of-domain splits, wherein the out-of-domain designation is given to an image if there is a mention of any of the novel objects in at least one of its ground-truth captions. The F1-score of novel objects in out-of-domain split [13] captions is a measure of model's ability to accurately include novel objects in predicted captions. It is a widely accepted NOC metric along with out-of-domain CIDEr.

We propose a variant of SCST that directly optimizes F1-score, and refer to it as SCST-F1. Note that our approach facilitates this formulation since we have access to out-of-domain training data, i.e., either synthetic images with their synthetic captions, partially paired images with their pseudo-label captions, or both. We split the SCST reward into three components: in-domain CIDEr, out-of-domain CIDEr and out-of-domain F1-score; these are denoted using $r_1(.)$, $r_2(.)$ and $r_3(.)$ respectively. We use beam search with beam size k during SCST and track top-k most probable captions for each sample image. For each reward component $r_j(.)$, the mean score of these top-k captions [11] is used as its RL baseline $\tilde{r}_j$ during SCST loss calculation. The gradient of SCST-F1 loss for each sample is approximated

as:

$$\nabla_\theta L(\theta) \approx -\frac{1}{k} \sum_{i=1}^{k} (\alpha_1(r_1(y^i) - \tilde{r}_1) + \alpha_2(r_2(y^i) - \tilde{r}_2) +$$
$$\alpha_3(r_3(y^i) - \tilde{r}_3)) \nabla_\theta \log p_\theta(y^i), \qquad (1)$$

where θ denotes the model parameters, y^i is the caption in i-th beam, and $\tilde{r}_j = \frac{1}{k}\sum_{i=1}^{k} r_j(y^i)$ for $j = 1, 2, 3$. We set $r_1(y^i) = 0$ for an out-of-domain sample, and $r_2(y^i) = 0$, $r_3(y^i) = 0$ for an in-domain sample. The hyper-parameters α_1, α_2, α_3 are the weights given to each of the three reward components.

3.4   Training Loss

In Step I of our training, we use only the cross-entropy loss. In Step II, we first use cross-entropy loss and then use SCST-F1 loss. In Step III, we train using only the SCST-F1 loss.

4   Experiments

4.1   Dataset

We evaluate our approach on *held-out* MS COCO dataset, using the splits provided by DCC [13] for NOC. This dataset is a subset of 2014 MS COCO [10] that originally contains training split with 82,783 images and validation split with 40,504 images. The *held-out* dataset excludes all image-caption pairs that describe any of the 8 selected objects ('bottle', 'bus', 'couch', 'microwave', 'pizza', 'racket', 'suitcase', , 'zebra') from training split, which results in a training split that contains 70,194 images with fully paired data. These 8 object classes are referred to as novel object classes from hereon. In our trials, we consider the remaining 12,589 images excluded from the training split as partially paired data. DCC also provides *held-out* validation and test splits that each contain 20,252 images; they further divide each of these splits into in-domain and out-of-domain splits for NOC evaluation.

4.2   Implementation Details

Synthetic data generation: To generate synthetic data for each novel object, we use the top similar candidate objects in the fully paired data that were identified by FDM-net [8]. For e.g., fully paired images with 'cup', 'glass', or 'vase' are considered as candidates to generate synthetic data for novel object 'bottle'. We exclude four of these candidate object classes ('stove', 'bread', 'box', and 'racquet') as they don't have bounding box annotations in MS COCO dataset. We show our final list in Supplemental material. We generate at most $K = 2400$ synthetic images for each novel object; if there are $m = 3$ candidate objects for a novel object (for e.g., 'bottle' has 'cup', 'glass', 'vase' candidates), we generate at most $K/m = 800$ images for each novel-candidate object pair (for e.g., 'bottle','cup'). We set $A_{min} = 1000$ pixels, $\Delta A_{max} = 200\%$, $AR_{min} = 0.05$, $AR_{max} = 5.0$, and $\Delta AR_{max} = 30\%$. To generate synthetic captions, we use NLTK [18] part-of-speech tagger to identify adjectives and nouns; we set $\Delta R_{adj} = 2$, and $\Delta R_{noun} = 1$. These hyper-parameters were empirically selected by manually inspecting the quality of few synthetic images and their captions, and are by no means the most optimal values.

Model Architecture: We use the popular Bottom-Up Top-Down (Up-Down) captioning model [5] architecture. Please see Supplemental material for details.

Object Detection Labels and CBS: We use the Faster R-CNN model trained on Open Images from *no-caps* baseline [1] to generate object detection labels for the partially paired images. CBS in Step III of our model training uses these labels to generate one pseudo-label caption for each of these images. We use the CBS implementation from *no-caps* baseline [1].

Evaluation Metrics: We use SPICE [3], METEOR [12] and CIDEr [24] metrics to evaluate the caption quality; these are denoted as 'S', 'M' and 'C' in our Results Tables. On out-of-domain test set, we also evaluate the F1-score metric of novel objects, denoted as 'F1' in our Results tables. All the metric scores on out-of-domain test set are macro-averaged across the eight novel object classes for consistency with results reported in previous works [4]. We primarily focus on out-of-domain CIDEr and F1-score for comparison with state-of-the-art results as they are widely accepted as the caption fluency and NOC metrics respectively. F_β-**score of CIDEr and F1-score**: We often notice a trade-off between out-of-domain CIDEr and F1-score in state-of-the-art results (see Table 1), which makes the comparison of results non-trivial. In order to represent them into a single metric for comparison, we also measure the F_β-score [21] of out-of-domain CIDEr and F1-score with β deciding the relative importance of CIDEr. $\beta = 1$ gives equal importance to both metrics, while $\beta > 1$ gives more importance to caption fluency (i.e, CIDEr) than including the novel objects in caption and degrading its fluency. We evaluate the models with β set to 1 and 1.5, and denote them as 'CF1' and 'CF1.5' in our Results Tables.

Training Details: For all the steps in our model training, we use a learning rate (LR) schedule that linearly decreases from initial rate to 0. We use a batch size of 100 for training with cross-entropy loss. During SCST-F1 training, we use a batch size of 50, beam search with beam size $k = 5$ and set $\alpha_1 = 1$, $\alpha_2 = 1$, $\alpha_3 = 1$. In **Step I**, we train our baseline model for 40,000 iterations with an initial LR of 0.01. In **Step II**, we first fine-tune our model with cross-entropy loss for 15,000 iterations with initial LR of 0.005. We then fine-tune it using SCST-F1 for 10,000 iterations with initial LR of 0.0125. In **Step III**, we further fine-tune our model using SCST-F1 for $N = 4$ rounds with 6,000 iterations in each round. We use an initial LR of 0.002 for first round, and scale it by 0.8 across rounds. We performed manual hyper-parameter tuning using CIDEr and F1-scores of validation split to set these hyper-parameter values; so, there may be scope to further improve these settings.

Inference Details: We use beam search with beam size $k = 5$. We do not use CBS during inference to generate the final captions from our PS-NOC.

Caption Post-processing: We notice that using SCST during training sometimes results in captions that end with the words 'in', 'a', etc. To improve caption quality, we perform post-processing during SCST-F1 training and inference, and remove such words at the end of the caption. We refer to our approach without such post-processing as Sol-1, and the one with the post-processing as Sol-2. We used slightly different training schedule for PS-NOC Step III in our Sol-2, compared to Sol-1; these details are provided in Supplemental material.

4.3   Results

Comparison with previous works: In Table 1, we compare our PS-NOC results on the test split with previous works and state-of-the-art results. The column C-RL indicates if the approach uses any form of CIDEr based RL. For PS-NOC, we provide the results for both Sol-1, Sol-2 (discussed in Section 4.2). Sol-1 does not include any caption post-processing, while Sol-2 includes it.

Table 1: Results of our PS-NOC approach on test-split and comparison against state-of-the-art results. PS-NOC provides the highest scores for both out-of-domain CIDEr and F1-scores, which is also seen in their CF1 and CF1.5 scores. Sol-2 includes additional post-processing during SCST-F1 training and inference to improve the caption quality.

ID	Approach	C-RL	Out-of-domain						In-domain		
			S	M	C	F1	CF1	CF1.5	S	M	C
1a	PS-NOC (Sol-1)	Yes	**19.7**	**27.2**	**101.5**	**86.1**	**93.2**	**96.2**	19.2	26.9	110.1
1b	PS-NOC (Sol-2)	Yes	**20.8**	**28.0**	**103.8**	85.9	**94.0**	**97.6**	20.5	27.7	110.9
2	PS3 [■]	No	17.9	25.4	94.5	63	75.6	81.9	19.0	25.9	101.1
3a	FDM (no CBS) [■]	No	19.4	25.9	84.8	64.7	73.4	77.4	20.2	27.2	109.7
3b	FDM (CBS) [■]	No	19.6	25.6	85.3	85.7	85.5	85.4	19.7	26.2	105.5
4	NBT (CBS) [▣]	No	17.4	24.1	86.0	70.3	77.4	80.5	18.0	25.0	92.1
5a	Reg. Sel. [■]	No	18.3	24.9	78.2	75.0	76.6	77.2	19.2	26.2	97.0
5b	Reg. Sel. (DGBS) [■]	Yes	19.4	26.3	88.5	75.1	81.3	83.9	21.0	27.9	115.3
6	ANOC [■]	Yes	18.2	25.2	94.7	64.3	76.6	82.7	-	-	-
7	ECOL-R (CBS) [▣]	Yes	19.1	25.7	99.1	71.8	83.3	88.7	20.8	26.8	112.6

Table 2: Results of our PS-NOC approach on test split by setting different values of α_3 during our SCST-F1 in Step II. We train our model using only Steps I and II for these trials. The F1-score increases as α_3 increases demonstrating that our SCST-F1 technique is effective.

Row	Weight factors			In-domain	Out-of-domain	
	α_1	α_2	α_3	C	C	F1
1	1	1	0	113.4	98.6	70.8
2	1	1	1	111.3	99.2	76.4
3	1	1	5	108.9	99.5	81.7
4	1	1	10	107.4	98.7	82.4
5	1	1	20	104.4	97.3	84.8

PS-NOC results (Rows 1a, 1b) consistently set new state-of-the-art results for both out-of-domain CIDEr and F1-score. They are also the most balanced in terms of CIDEr and F1, which is reflected in their highest F_β-scores in columns CF1, CF1.5. Our results outperform the FDM-net and PS3 results, which are the existing partially supervised solutions for this dataset. Also, PS-NOC does not use the lexical constraints (471 most common adjectives, verbs, noun forms generated by additional training of MS COCO) that PS3 uses during CBS.

Effectiveness of SCST-F1: In Table 2, we demonstrate that our proposed SCST-F1 is effective by training Step-II using different values of α_3. We train each of these models using our PS-NOC Steps I and II, and evaluate the results on test split. Row 1 with $\alpha_3 = 0$ corresponds to regular CIDEr based SCST with no F1-score reward. As we increase α_3 from 0 to 20, we see the corresponding increase in out-of-domain F1-score of test split. Note that for a given captioning model, there could be trade-off between in-domain CIDEr and out-of-domain CIDEr or F1-score. So, $\alpha_1, \alpha_2, \alpha_3$ should be chosen accordingly. We give equal importance to all the three scores in this work and therefore set all these three values to 1.

Qualitative examples: We show qualitative examples of the captions generated using our PS-NOC Sol-1 in Figure 4, for each of the 8 novel objects. It correctly generates captions that include these novel objects, compared to the baseline model that does not include them.

Figure 4: Examples of captions generated using our approach for the 8 novel object classes.

Table 3: Test split results from the ablation studies on different components of our PS-NOC approach. The results demonstrate the benefits of using i. Synthetic data, ii. Pseudo-labeling, iii. SCST, iv. SCST-F1, and v. Our overall approach.

Row	C-RL	Training technique	Out-of-domain				In-domain		
			S	M	C	F1	S	M	C
1	No	I	19.6	28.0	69.7	0.0	19.4	27.9	108.0
2	No	I + CBSInf	18.1	26.1	76.6	56.2	17.6	25.8	88.8
3	No	I + II	19.8	28.2	89.0	62.1	19.3	27.6	103.4
4	No	I + II + III	19.9	28.4	96.3	75.8	19.5	27.8	105.9
5	Yes	I + II(SCST)	20.5	27.8	98.6	70.8	20.2	27.6	113.4
6	Yes	I + II(SCST) + III	20.2	28.3	99.8	72.4	19.9	27.9	108.2
7	Yes	I + II(SCST) + III(SCST)	20.1	27.8	101.0	78.8	19.6	27.1	111.0
8	Yes	I + II(SCST-F1)	20.6	28.1	99.2	76.4	19.9	27.6	111.3
9	Yes	I + II(SCST-F1) + III	20.2	28.5	102.2	75.7	20.0	28.2	108.2
10	Yes	I + II(SCST-F1) + III(SCST-F1) Sol-1	19.7	27.2	101.5	86.1	19.2	26.9	110.1
11	Yes	I + II(SCST-F1) + III(SCST-F1) Sol-2	20.8	28.0	103.8	85.9	20.5	27.7	110.9

4.4 Ablation Studies

In Table 3, we provide the test split results from our extensive ablation studies on different components of our PS-NOC approach. We denote the three Steps in our approach as I, II and III respectively. II denotes training using cross-entropy loss, while II(SCST) and II(SCST-F1) denote additional training using regular CIDEr based SCST and our SCST-F1 respectively. III denotes training using cross-entropy loss, while III(SCST) and III(SCST-F1) denote training using regular SCST and our SCST-F1 instead. We also run a trial with CBS decoding at inference on top of Step I, and denote it as 'I + CBSInf'.

This Table demonstrates the following:

(a) Using our synthetic data improves the out-of-domain scores over the baseline, as shown in Row 1 vs. Row 3. It also provides higher scores than running CBS only at inference, as shown in Row 2 vs. Row 3.

(b) Using our pseudo-labeling approach on top of Step II improves the out-of-domain scores, as shown in all the rows with an additional '+ III'; for e.g., Row 4 vs. Row 3.

(c) Using SCST during Steps II or III improves the out-of-domain scores over the ones with cross-entropy loss training, as shown in all rows with an additional '(SCST)'; for e.g., Row 5 vs. Row 3.

(d) Using SCST-F1 during Steps II or III improves the out-of-domain scores over using

regular SCST, as shown in all rows with additional '-F1'; for e.g., Row 8 vs. Row 5, and Row 10 vs. Row 7.

(e) Finally, Row 10 and Row 11 are the results from our complete PS-NOC approach, and these have the highest scores. This is an improvement of 85.9 F1-score and 34.1 CIDEr points over the baseline model (Row 1 vs. Row 11) that does not use partially paired data.

Additional ablation studies, along with limitations and future work, are included in the Supplemental material.

5   Conclusion

In this work, we proposed a model-agnostic Partially-Supervised Novel Object Captioning solution, PS-NOC. PS-NOC primarily focuses on the training approach that uses fully paired image-caption data and partially paired images with object detection labels and bounding box annotations. We first generate paired synthetic data within the context of fully paired data. Since, we generate synthetic images instead of region features, our approach is also applicable to the newer captioning models that use the entire image as inputs. We then propose a training approach to train the captioning model for novel objects using this additional synthetic data and partially paired data. This includes novel training techniques such as SCST-F1 and offline pseudo-labeling for NOC. Using UpDown captioning model as baseline, our PS-NOC approach sets new state-of-the-art results on *held-out* MS COCO out-of-domain test split with 85.9 F1-score and 103.8 CIDEr. This is an improvement of 85.9 F1-score and 34.1 CIDEr points over the baseline model that does not leverage partially paired data.

References

[1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8948–8957, 2019.

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Guided open vocabulary image captioning with constrained beam search. *arXiv preprint arXiv:1612.00576*, 2016.

[3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer, 2016.

[4] Peter Anderson, Stephen Gould, and Mark Johnson. Partially-supervised image captioning. *arXiv preprint arXiv:1806.06004*, 2018.

[5] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[6] Shashank Bujimalla, Mahesh Subedar, and Omesh Tickoo. B-scst: Bayesian self-critical sequence training for image captioning. *arXiv preprint arXiv:2004.02435*, 2020.

[7] Marco Cagrandi, Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cuc-chiara. Learning to select: A fully attentive approach for novel object captioning. *arXiv preprint arXiv:2106.01424*, 2021.

[8] Tingjia Cao, Ke Han, Xiaomei Wang, Lin Ma, Yanwei Fu, Yu-Gang Jiang, and Xi-angyang Xue. Feature deformation meta-networks in image captioning of novel objects. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10494–10501, 2020.

[9] Xianyu Chen, Ming Jiang, and Qi Zhao. Leveraging human attention in novel object captioning. In *International Joint Conference on Artificial Intelligence*, 2021.

[10] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evalua-tion server. *arXiv preprint arXiv:1504.00325*, 2015.

[11] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Confer-ence on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[12] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014.

[13] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, and Trevor Darrell. Deep compositional captioning: Describing novel object categories without paired training data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–10, 2016.

[14] Xiaowei Hu, Xi Yin, Kevin Lin, Lijuan Wang, Lei Zhang, Jianfeng Gao, and Zicheng Liu. Vivo: Visual vocabulary pre-training for novel object captioning. *arXiv preprint arXiv:2009.13682*, 2020.

[15] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Shahab Kamali, Matteo Malloci, Jordi Pont-Tuset, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://storage.googleapis.com/openimages/web/index.html*, 2017.

[16] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Inter-national Journal of Computer Vision*, 123(1):32–73, 2017.

[17] Wei Liu, Sihan Chen, Longteng Guo, Xinxin Zhu, and Jing Liu. Cptr: Full transformer network for image captioning. *arXiv preprint arXiv:2101.10804*, 2021.

[18] Edward Loper and Steven Bird. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, page

63–70, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1118108. 1118117. URL https://doi.org/10.3115/1118108.1118117.

[19] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7219–7228, 2018.

[20] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[21] David Powers. Evaluation: From precision, recall and f-measure to roc, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2(1):37–63, 2011.

[22] 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.

[23] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7008–7024, 2017.

[24] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[25] Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, Raymond Mooney, Trevor Darrell, and Kate Saenko. Captioning images with diverse objects. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5753–5761, 2017.

[26] Yufei Wang, Ian D Wood, Stephen Wan, and Mark Johnson. Ecol-r: Encouraging copying in novel object captioning with reinforcement learning. *arXiv preprint arXiv:2101.09865*, 2021.

[27] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021.

[28] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Incorporating copying mechanism in image captioning for learning novel objects. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6580–6588, 2017.