Feature Deformation Meta-Networks in Image Captioning of Novel Objects

Tingjia Cao,[1] **Ke Han,**[1] **Xiaomei Wang,**[1] **Lin Ma,**[3] **Yanwei Fu,**[2] **Yu-Gang Jiang,**[1] **Xiangyang Xue**[1]
[1]Shanghai Key Lab of Intelligent Information Processing, School of Computer Science Fudan University
[2]School of Data Science, and MOE Frontiers Center for Brain Science, Fudan University, [3]Tencent AI Lab
{tjcao16, yanweifu, ygj, xyxue}@fudan.edu.cn, forest.linma@gmail.com

Abstract

This paper studies the task of image captioning with novel objects, which only exist in testing images. Intrinsically, this task can reflect the generalization ability of models in understanding and captioning the semantic meanings of visual concepts and objects unseen in training set, sharing the similarity to one/zero-shot learning. The critical difficulty thus comes from that no paired images and sentences of the novel objects can be used to help train the captioning model. Inspired by recent work (Chen et al. 2019b) that boosts one-shot learning by learning to generate various image deformations, we propose learning meta-networks for deforming features for novel object captioning. To this end, we introduce the feature deformation meta-networks (FDM-net), which is trained on source data, and learn to adapt to the novel object features detected by the auxiliary detection model. FDM-net includes two sub-nets: feature deformation, and scene graph sentence reconstruction, which produce the augmented image features and corresponding sentences, respectively. Thus, rather than directly deforming images, FDM-net can efficiently and dynamically enlarge the paired images and texts by learning to deform image features. Extensive experiments are conducted on the widely used novel object captioning dataset, and the results show the effectiveness of our FDM-net. Ablation study and qualitative visualization further give insights of our model.

Introduction

It is one of the long term goals for the AI community to pursue an agent that can automatically and linguistically describe the captured visual signals. Recently, this task is formulated as the task of image captioning, which has made significant progress powered by deep architectures (Xu et al. 2015). Such a task relies on large-scale datasets containing the image and sentence pairs, like MSCOCO (Chen et al. 2015) and Flickr 30K (Plummer et al. 2015). However, the expensive cost of manually annotating limits the richness of visual concepts in image captioning dataset. For example, compared to the recently released object detection dataset Open Images V4 (Rom et al. 2017) with 600 object classes, MSCOCO (Lin et al. 2014) dataset only contains 91 underlying object classes. Thus, novel object captioning (Tran et al. 2016; Anne Hendricks et al. 2016) is recently studied to generate appropriate descriptions for novel classes which have no training instances. To address this problem, extra resources have been investigated. These resources should be easy to be collected and contain abundant visual concepts.

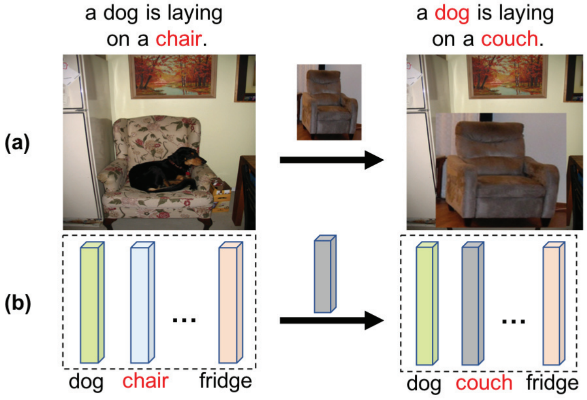


Figure 1: The idea of feature deformation meta-networks (FDM-net). (a) Inconsistent visual problems of directly learning deformed images and the mismatching problem applying deformed methods (Chen et al. 2019a; 2019b) to novel object captioning; (b) Feature deformation meta-networks solve the problems in (a).

Previous works attempt to leverage object detection datasets and text corpora for novel object captioning. Particularly, these methods explored transferring knowledge by designing template-based caption models (Lu et al. 2018; Wu et al. 2018), multi-task models (Anne Hendricks et al. 2016), or creating weakly-annotated data (Anderson, Gould, and Johnson 2018). While these methods have got promising results, we argue that tackling this problem more explicitly from another perspective can lead to better performance.

Humans are good at describing visual scenes with novel objects. Not surprisingly, cognitive evidence (Marr 1982) shows that the visually grounded language generation is not end-to-end, but primarily attributed to the "high-level" symbolic reasoning. That is, once we abstract the scene into symbols, the novel object can be merged into the background automatically by our brains, which inspires us to address novel object captioning task from a completely different perspective. To empower the captioning model with such ability, we synthesize new training instances composed of novel objects and reasonable backgrounds as additional training data. Taking such deformed images (Chen et al. 2019a; 2019b; Satoshi Tsutsui 2019) as augmented images has already shown efficacy to help train image classification models. Although these deformed images might not be visually realistic, they still maintain critical semantic information to help build better classifiers. It can be developed to adapt to the novel image captioning task, as using an original image to keep background content and introducing novel objects by replacing patches. As shown in Fig. 1(a), by replacing the bounding box of "chair" in the original image with a cropped patch "couch" (novel object); we hope to synthesize a deformed image that can be described as "A dog is laying on a couch".

However, two critical problems existing here should be identified. Firstly, due to the spatial overlapping of different objects, the novel object that blocks all information in the original region may cause the loss of discriminative information. For example, the "couch" in Fig. 1(a) completely blocks the "dog" in the original image, though the generated sentence still emphasizes the "dog." Another problem shown in Fig. 1(a) is the strong noise by the side of the "couch" border, caused by different sizes and ratios of the patches. As shown in Fig. 1(a), on the edge of "couch", we can still find the apparent residual visual information of the "chair". Thus, due to the gap between expected images and actual generated ones, the direct exploitation of deformed images is inefficient to the novel object captioning task. This method is not ready to create appropriate augmented data for image captioning models because of the severe mismatching problem between images and texts caused by the information loss of the context.

Inspired by the "high-level" symbolic reasoning theory in (Marr 1982), this paper proposes feature deformation meta-networks (FDM-net) to overcome problems mentioned above. In particular, we employ features of region-of-interest (RoI) from the image detector as "high-level" symbols. We first remove one RoI feature from the feature map of the original image, which is treated as the candidate to be replaced. On the feature level, there is less overlapping of important visual information and less complex spatial relationships between objects. The remaining features are able to present the background with less information loss or residue. Then, we synthesize new feature maps composed of features from novel objects and backgrounds. Essentially, nowadays popular captioning decoders (Anderson et al. 2018) are trained for a projection from feature maps to sentences. It means that we can synthesize feature maps to represent non-existent images as extra training data, so our method can be seamlessly integrated into the encoder-decoder architecture. Particularly, the FDM-net synthesizes deformed feature maps instead of deformed images to help train better image captioning models. As illustrated in Fig. 1(b), after replacing the feature of the object in original images, the potent visual information of novel objects is added without blocking other discriminative features. Consequently, the mismatching problem is remitted.

To generate corresponding sentences for new feature maps, we improve a sentence reconstruction network with the scene graph in (Yang et al. 2018). This subnet can reduce the semantic bias caused by the discrepancy in attributes and relationships between source and target objects. Considering a sentence containing "coffee cup" ("cup" is to be replaced with "bottle"), we are more likely to get "wine bottle" rather than "coffee bottle" after the network. Benefited from the universally existed structure of sentences, we are able to modify both attributes and objects on scene graph level to get correct descriptions for generated feature maps. Thus, we implement an unsupervised scene graph based sentence reconstruction network in (Yang et al. 2018).

To sum up, we have several contributions. Firstly, we propose feature deformation meta-networks (FDM-net) that aim to help image captioning models adapt to the novel objects by generating deformed training data. It contains feature deformation and scene graph sentence reconstruction sub-nets. Secondly, the feature deformation subnet executes generating deformed feature maps with novel objects. It can avoid the mismatching problem by replacing region-of-interesting (RoI) features instead of creating deformed images. Thirdly, scene graph strategy is applied to reconstruct sentences corresponding to the augmented feature maps. Finally, extensive results on MSCOCO and Open Image illustrate the efficacy of our structure, and we have achieved state-of-the-art results on several metrics for novel object captioning task.

Related Work

Novel Object Captioning

General image captioning aims to describe images with sentences. For increasing scalability of diversified objects, recently novel object captioning (Anne Hendricks et al. 2016; Lu et al. 2018; Wu et al. 2018) has attracted lots of attention. However, most proposed methods are architectural in essence. Researchers have designed template-based caption models (Lu et al. 2018), multi-task models (Anne Hendricks et al. 2016) and novel sampling algorithms (Koehn 2016). These novel structures are disjointed from normal image captioning task to varying degrees, which causes poor performance on in-domain scores. Inspired by the newfangled deformation strategies (Chen et al. 2019b; 2019a; Satoshi Tsutsui 2019), we design deformed meta-networks for the novel object image captioning task which can be seamlessly integrated into popular encoder-decoder captioning models. However, different from image deformation methods, we deformed RoI features in our proposed feature deformation sub-net.

Scene Graph

The task of converting a sentence or an image into a structured meaningful representation has received considerable

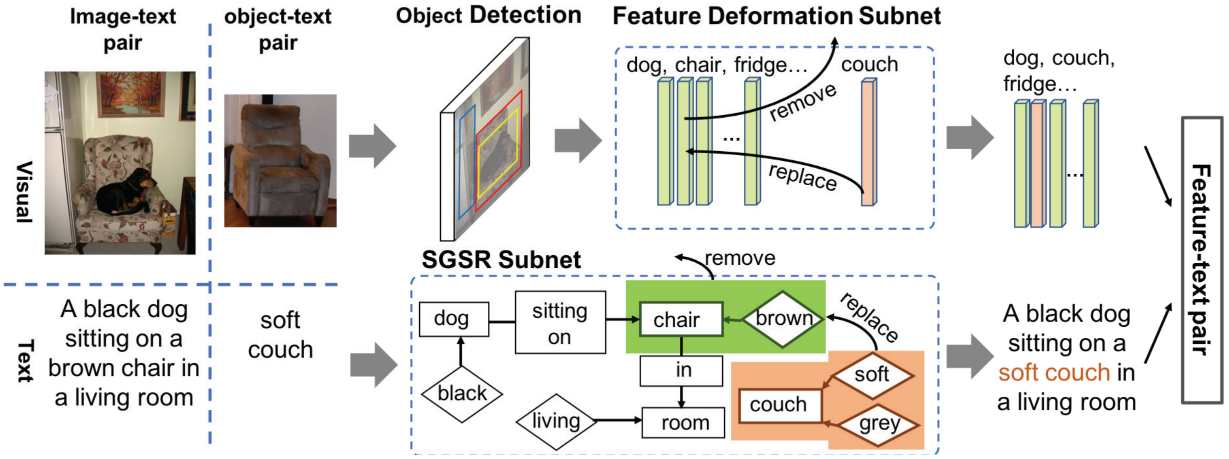


Figure 2: The framework of our proposed feature deformation meta-learning networks. It contains two sub-nets: feature deformation sub-net and scene graph sentence reconstruction sub-net. The former focuses on generating deformed visual features with novel objects and the latter is responsible for reconstructing sentences by replacing key objects and attributes based on scene graph strategy.

attention. The scene graph is such a general representation existing in several vision datasets (Krishna et al. 2016), which contains the structured semantic information including objects, attributes and relationships. In the visual and linguistics community, scene graphs have been used to a number of tasks like image retrieval, visual question answering, and image captioning (Yang et al. 2018). Enlightened by (Yang et al. 2018) that parses captions into scene graphs, we reconstruct deformed sentences similarly by parsing sentences to graphs. Differently, our sentence reconstruction network targets to replace attributes and objects and generate corresponding descriptions for augmented feature maps.

Method

We propose a framework that combines feature deformation sub-net and scene graph sentence reconstruction sub-net (*SGSR*) to caption images with novel objects. Instead of replacing patches of novel objects in images directly, feature-level visual information is applied to introduce the novel objects to the model. Feature deformation sub-net aims to augment the visual training information with novel objects. To fit the augmented visual information, we also reconstruct matched texts containing novel objects synchronously by our proposed *SGSR* sub-net. The pipeline of generating training instances with visual and text pairs is shown in Fig. 2. In the following, both the feature deformation sub-net and scene graph sentence reconstruction sub-net are introduced in detail.

Feature Deformation Subnet

To augment image features with novel objects, we firstly utilize mis-labelled probability strategy (*MLS*) for choosing a set of objects from the seen object set that are easy to confuse. In other words, each novel object corresponds to a group of confused objects. Then we replace candidate features with novel object features. Such feature maps with novel objects and backgrounds are added into the training set accordingly.

Mis-labelled Probability Strategy. Following the setting of novel object captioning, we have a set of training image-sentence pairs with seen classes $\mathcal{C}_s$ and testing images with unseen classes (novel objects) $\mathcal{C}_u$, which $\mathcal{C}_s \cap \mathcal{C}_u = \emptyset$. The purpose of our Mis-labelled Probability Strategy (*MPS*) is to select a group of easily-confused objects $\mathcal{C}_{con}$ for a novel object $\boldsymbol{c}_u \in \mathcal{C}_u$. We have $\mathcal{C}_{con} \subset \mathcal{C}_s$. In advance, we pre-train a general captioning model on image-sentence pairs that only contain seen classes.

The validation images with novel objects are then fed into the pre-trained model. From generated sentences, we denote the number of occurrences of each confused seen object $\boldsymbol{c}_{con} \in \mathcal{C}_{con}$ as $\boldsymbol{n}$. Then, k objects in $\mathcal{C}_{con}$ with largest numbers of occurrences $\boldsymbol{n}$ are preserved and constructed to a mis-labelled group $\mathcal{C}_{con} = \{\boldsymbol{c}_{con}^1, \boldsymbol{c}_{con}^2, ..., \boldsymbol{c}_{con}^k\}$ and their occurrence time $N = \{\boldsymbol{n}_1, \boldsymbol{n}_2, ..., \boldsymbol{n}_k\}$.

Then, for each novel object $\boldsymbol{c}_u \in \mathcal{C}_u$, we normalize occurrence times N to obtain similarity scores between $\boldsymbol{c}_u$ and $\boldsymbol{c}_{con}^i$: $P = \{\boldsymbol{p}_1, \boldsymbol{p}_2, ..., \boldsymbol{p}_k\}$ by $\boldsymbol{p}_i = \frac{\boldsymbol{n}_i}{\sum_j \boldsymbol{n}_j}$. Suppose we would like to generate M augmented training instances: $M \times \boldsymbol{p}_i$ images are applied to generate augmented features with confused object $\boldsymbol{c}_{con}^i$.

Feature Deformation. After selecting confused object sets from base dataset by *MLS*, a pre-trained detection model of faster R-CNN with ResNet-101 (Anderson et al. 2018), named as $\mathbf{g_{det}}(\cdot)$, is utilized to extract RoI region features. Given an image $\mathcal{I}_{nov}$ containing a novel object $\boldsymbol{c}_u \in \mathcal{C}_u$ and a training image $\mathcal{I}_{ori}$ containing a confused seen object $\boldsymbol{c}_{con}$ in its confused group $\mathcal{C}_{con}$, the RoI region features (Ren et al. 2015) of the images are obtained by:

$$\boldsymbol{f}_{nov} = \mathbf{g_{det}}(\mathcal{I}_{nov}) \qquad \boldsymbol{f}_{ori} = \mathbf{g_{det}}(\mathcal{I}_{ori}) \qquad (1)$$

where $\boldsymbol{f} = \{f_1, \ldots, f_T\}$ means RoI region features with T regions in an image. Then we classify features of each

Table 1: Top three similar concepts for novel class decided by Mis-labeled Probability Strategy

bottle	bus	couch	microwave	pizza	racket	suitcase	zebra
cup	truck	chair	stove	sandwich	bat	bag	giraffe
glass	car	bed	refrigerator	bread	frisbee	backpack	elephant
vase	train	bench	toaster	cake	racquet	box	cow

Table 2: Evaluation on sentence reconstruction network. Where B@1, B@4, M, R, C and S mean Blue-1, Blue-4, METEOR, Rough, CIDEr-D and SPICE respectively.

Models	B@1	B@4	M	R	C	S
SGSR	91.6	64.9	38.7	71.7	175.0	34.3

region in $\boldsymbol{f}_{nov}$ and $\boldsymbol{f}_{ori}$ and denote f^u_{nov} and f^s_{ori} as the region features that are classified to the novel object $\boldsymbol{c}_u$ and the confused seen object $\boldsymbol{c}_{con}$ respectively. We preserve the background of the training image by excluding the confused seen object $\boldsymbol{c}_{con}$ by $\boldsymbol{f}_{ori} - \{f^s_{ori}\}$ and we introduce novel objects into augmented instance by:

$$\tilde{\boldsymbol{f}} = (\boldsymbol{f}_{ori} - \{f^s_{ori}\}) \cup \{f^u_{nov}\} \tag{2}$$

Scene Graph Sentence Reconstruction Subnet

Despite the effectiveness of *MLS*, the augmented pairs in principle can not be directly utilized to boost the performance of novel object captioning. The key problem comes from the visual difference. To this end, the motivation of *SGSR* is to decrease the differences between the augmented fake pairs and image captioning datasets. Specifically, one important goal of the proposed sub-net in this subsection is to repair possible predicative errors, semantic errors and meaningless sentences, caused by the fake augmented data.

To train the model, we apply scene graph strategy to generate sentences. Several sentences $\mathcal{S} = \{\mathcal{S}_1, \mathcal{S}_2, ..., \mathcal{S}_m\}$ are given in the training dataset to describe $\mathcal{I}_{ori}$. For a augmented feature maps $\tilde{\boldsymbol{f}}$, our target is to generate several sentences that can describe it: $\tilde{\mathcal{S}} = \{\tilde{\mathcal{S}}_1, \tilde{\mathcal{S}}_2 ..., \tilde{\mathcal{S}}_m\}$. As shown in Fig. 2, we modify Auto-Encoding Scene Graphs (Yang et al. 2018) to reconstruct our target sentences,

$$\begin{aligned} \textbf{Encoder} &: \mathcal{G} \leftarrow \mathcal{S} \\ \textbf{Scene Graph} &: \tilde{\mathcal{G}} \leftarrow \mathcal{G} \\ \textbf{Decoder} &: \tilde{\mathcal{S}} \leftarrow \tilde{\mathcal{G}} \end{aligned} \tag{3}$$

So, we should encode the candidate sentences $\mathcal{S}$ to scene graph $\mathcal{G}$ firstly. We further illustrate the process of $\tilde{\mathcal{G}} \leftarrow \mathcal{G}$. We get attributes for the novel object by:

$$\mathcal{A}_{nov} = \mathbf{g_{attr}}(f^u_{nov}) \tag{4}$$

where f^u_{nov} is the feature of novel object, and $\mathbf{g_{attr}}$ is an attribute classifier network with two fully connected layers trained on visual genome dataset (Krishna et al. 2016). Then we replace the object node O_{ori} with O_{nov}. Since the attributes node connected with O_{ori} may be not exactly corresponding to new generated $\tilde{\boldsymbol{f}}$, such attributes should be deleted. To give more specific information for novel objects, $\mathcal{A}_{nov}$ is taking as attribute nodes and we construct the edges to O_{nov}. In this way, we will get sentences containing less noise than simply parsing and replacing a single word of ‘chair’ with ‘couch’ and its attributes. Scene Graph Decoder employs Beam Search (Anderson et al. 2016a) to generate m new sentences $\tilde{\mathcal{S}}$ from $\tilde{\mathcal{G}}$; more reasonable sentences can be generated in such a manner.

Experiments

Dataset Split. We follow the novel object captioning split (NOC split) introduced by (Anne Hendricks et al. 2016) to evaluate our proposed method. It comes from the standard split of MSCOCO 2014 (Chen et al. 2015) that contains 120K images, and each image is labelled with five human-annotated sentences. Eight objects (bottle, bus, couch, microwave, pizza, racket, suitcase and zebra) are selected as novel objects in NOC split. Correspondingly, all image-sentence pairs that include novel objects are removed from the standard training set. In the standard validation dataset, half of the pairs are randomly selected into new validation set, and others are selected into the test set. New validation and test sets are further separated into out-of-domain and in-domain subsets based on whether including positive examples for eight novel objects. For the in-domain validation set and test set, there are no image-sentence pairs containing novel objects while images from out-of-domain sets include novel objects. So in this split, models are required to describe images containing novel objects.

To evaluate the expandability of our method, we also conduct experiments on Open Image dataset – a large-scale dataset. Fifty classes are randomly chosen from Open Image dataset as novel objects. We utilize our FDM-net on MSCOCO training set to help learn the image captioning model.

Implementation Details. In our model, we use the pretrained bottom-up attention model to extract visual features. To make a fair comparison, we use traditional cross-entropy loss during training. The RoI features of novel objects come from OpenImage. Specially, we use mis-labelled probability strategy (*MLS*) to select top similar seen objects for each unseen object. As shown in Tab. 1, the top three similar seen objects are considered to conduct the replacing work with their corresponding novel objects. That means we set $k = 3$. Besides, constrained beam search (*CBS*) algorithm (Koehn 2016) is also applied in the test and validation stage. For ensuring the diversity of our augmented dataset, we extract

Table 3: Results on test dataset of DCC split. Where ○/●, ◇/◆ and □/■ mean whether apply Mis-Labeled Strategy (*MLS*), scene graph sentence reconstruction network (*SGSR*) and Constrained Beam Search (*CBS*) respectively. And No. means the augmented examples number of each novel object.

No.	Strategy			Out-of-Domain Scores				In-Domain Scores		
	MLS	SGSR	CBS	SPICE	METEOR	CIDEr	F1	SPICE	METEOR	CIDEr
0	○	◇	□	13.4	21.1	65.2	0	19.8	26.5	105.1
500	●	◇	□	18.2	24.6	79.9	57.2	19.6	26.4	105.0
1000	○	◇	□	18.0	24.9	79.8	53.5	19.6	26.3	104.0
	●	◇	□	18.6	25.0	80.6	63.7	19.6	26.2	103.3
	●	◆	□	19.4	**25.9**	84.8	64.7	20.2	**27.2**	109.7
	●	◆	■	**19.6**	25.6	**85.3**	85.7	19.7	26.2	105.5
1500	○	◇	□	18.4	24.8	80.3	64.0	19.4	26.3	104.2
	●	◇	□	18.4	25.1	81.4	64.8	19.3	26.2	102.6
	●	◆	□	19.0	25.7	84.0	64.7	**20.3**	**27.2**	**110.7**
	●	◆	■	19.3	25.5	83.9	**87.0**	19.3	25.9	104.7

100 novel object features as resources for the following replacement. To implement the SGSR, we train and evaluate a sentence reconstruction network, using sentences in Visual Genome and the input sentence itself as the ground truth. Tab 2 shows the results tested on Visual Genome, which demonstrates that our *SGSR* subnet can serve as a potent tool to generate high-quality synthetic descriptions.

Baselines and Competitors. We also compared with recent models, such as DCC (Anne Hendricks et al. 2016), NBT (Lu et al. 2018) and PS3 (Anderson, Gould, and Johnson 2018). Some decoding algorithms such as constrained beam search (Koehn 2016) are proved to be helpful to this task. For a fair comparison, when comparing with popular methods, we implement the same setting on FDM-net of whether using *CBS*.

Evaluation Metrics. We use three standard automatic evaluations metrics: CIDEr-D (Vedantam, Lawrence Zitnick, and Parikh 2015), METEOR (Banerjee and Lavie 2005) and SPICE (Anderson et al. 2016b). F1 scores are also reported for evaluating the performance of captioning eight novel concepts. We also report the human evaluation scores in our experiments. Word incorporation and image description proposed in (Venugopalan et al. 2017) are utilized as evaluation metrics. Given two sentences describe one image, people are asked to pick up the better one based on different evaluation points. Word incorporation assesses in which sentence organizes more meaningfully sentences with novel objects and image description focuses on which sentence describes the whole image better.

Quantitative Performance

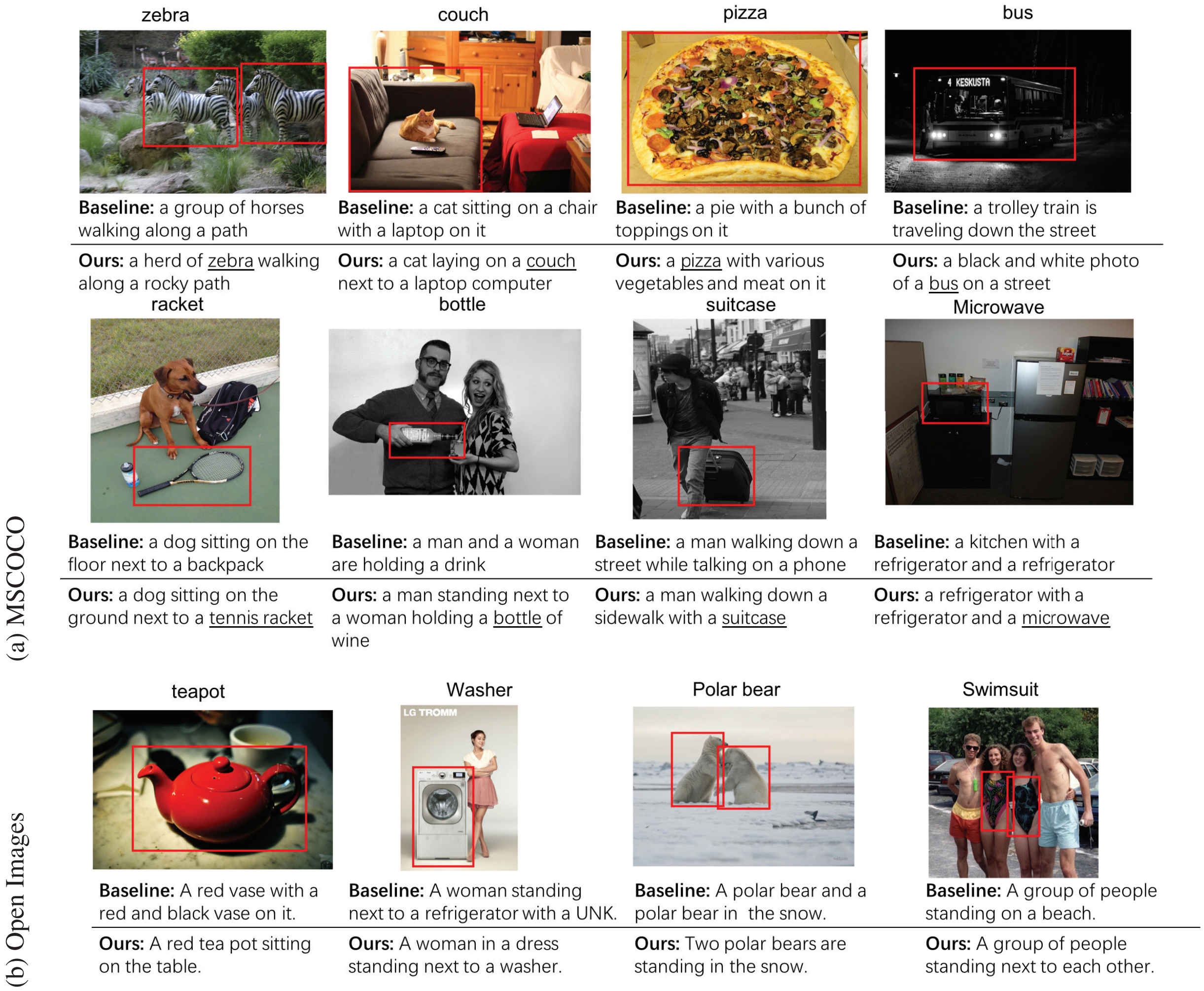
To display the quantitative performance of our proposed FDM-net on novel object image captioning task, we conduct several experiments to analyze our model versatilely by using different strategies (such as *FDM*, *MLS*, *SGSR* and *CBS*) and various numbers of augmented feature-text pairs. Meanwhile, we also compare our results with state-of-the-art competitors. Besides, we report the human evaluation results of our model. All experiments illustrated the quantitative superiority of our model in various aspects.

Results with Different Assignments. Several experiments are conducted by adding different numbers of augmented instances and utilizing kinds of strategies, as shown in Tab. 3. We use the results that are generated by general captioning model on NOC split as our baseline. Furthermore, different numbers of augmented feature-text pairs are added into the NOC training split. The *MLS* strategy denotes that we apply the features of top three nearest objects to the following replacement, while one most similar object is chosen for each novel object based on human common sense whenever *MLS* is not used. Directly, we parse and replace the words of objects and attributes in the original ground truth sentences if our proposed *SGSR* is not utilized. Besides, constrained beam search decoder is an optional strategy in the testing stage.

The results of our model with kinds of strategies and different numbers of augmented instances are shown in Tab. 3. There are several points we want to highlight. Firstly, our FDM-net displays great advantages compared with the general captioning model. No matter how many instances we add, the performance is increased. Just as shown in the Tab. 3, most of the scores are improved after adding augmented feature-text pairs by the FDM-net, especially F1 score increases from 0% to more than 50%. Obviously, the results on out-of-domain split have greater promotion than in-domain because our FDM-net focuses on enlarging the novel object knowledge in the training stage, which has a greater influence on the out-of-domain dataset. Secondly, scores are gradually raised by utilizing several strategies. When we pay attention to scores of the out-of-domain setting, applying *MLS* strategy makes an improvement, and it means our *MLS* method is efficient and useful. Particularly, our *SGSR* makes a great contribution in elevating novel object captioning performance. It states that replacing attributes and objects together on the scene graph level with *SGSR* trained on extra text resources makes significant contribution on novel object

Table 4: Evaluating performance with popular methods.

Model	CNN	CBS	Out-of-Domain Scores				In-Domain Scores		
			SPICE	METEOR	CIDEr	F1	SPICE	METEOR	CIDEr
DCC(Koehn 2016)	VGG-16	-	13.4	21.0	59.1	39.8	15.9	23.0	77.2
NOC(Venugopalan et al. 2017)	VGG-16	-	-	21.4	-	48.8	-	-	-
C-LSTM(Yao et al. 2017)	VGG-16	-	-	23.0	-	55.7	-	-	-
LRCN(Koehn 2016)	VGG-16	+	15.9	23.3	77.9	54.0	18.0	24.5	86.3
LRCN(Anne Hendricks et al. 2016)	Res-50	+	16.4	23.6	77.6	53.3	18.4	24.9	88.0
NBT(Lu et al. 2018)	VGG-16	-	15.7	22.8	77.0	48.5	17.5	24.3	87.4
	Res-101	+	17.4	24.1	86.0	70.3	18.0	25.0	92.1
PS3(Anderson, Gould, and Johnson 2018)	Res-101	-	17.9	25.4	**94.5**	63.4	19.0	25.9	101.1
	Res-101	+	18.2	25.2	92.5	62.4	19.1	25.9	99.5
FDM-net	Res-101	-	19.4	**25.9**	84.8	64.7	**20.2**	**27.2**	**109.7**
	Res-101	+	**19.6**	25.6	85.3	**85.7**	19.7	26.2	105.5



Baseline: a group of horses walking along a path

Ours: a herd of zebra walking along a rocky path

Baseline: a cat sitting on a chair with a laptop on it

Ours: a cat laying on a couch next to a laptop computer

Baseline: a pie with a bunch of toppings on it

Ours: a pizza with various vegetables and meat on it

Baseline: a trolley train is traveling down the street

Ours: a black and white photo of a bus on a street

Baseline: a dog sitting on the floor next to a backpack

Ours: a dog sitting on the ground next to a tennis racket

Baseline: a man and a woman are holding a drink

Ours: a man standing next to a woman holding a bottle of wine

Baseline: a man walking down a street while talking on a phone

Ours: a man walking down a sidewalk with a suitcase

Baseline: a kitchen with a refrigerator and a refrigerator

Ours: a refrigerator with a refrigerator and a microwave

Baseline: A red vase with a red and black vase on it.

Ours: A red tea pot sitting on the table.

Baseline: A woman standing next to a refrigerator with a UNK.

Ours: A woman in a dress standing next to a washer.

Baseline: A polar bear and a polar bear in the snow.

Ours: Two polar bears are standing in the snow.

Baseline: A group of people standing on a beach.

Ours: A group of people standing next to each other.

Figure 3: Visualization on MSCOCO and Open Images datasets

Table 5: Human evaluation results. Where 'coco' means the test set comes from MSCOCO dataset and 'open image' is from Open Images dataset.

Options	Word Incorporation		Image Description	
	coco	open image	coco	open image
FDM-net better	50.0	48.8	55.8	54.0
NBT better	18.0	21.0	16.9	19.4
Equally good	26.5	20.8	22.7	26.3
Neither is good	5.5	9.4	5.5	10.2

captioning than directly parsing and replacing. In the meantime, *CBS* algorithm makes an extra boosting on results, especially on F1 score. Finally, we can sum up that the more augmented instances do not mean the better performance when we compare the results of adding 500, 1000 and 1500 feature-text pairs in training data. Overall speaking, the best novel object captioning performance is achieved by adding 1000 extra instances with all of the *MLS*, *SGSR* and *CBS* strategies as shown in Tab. 3.

Comparison with Popular Methods. In this part, we compare the results of our proposed FDM-net with popular methods. In Tab. 4, comparing with popular methods separately on whether using *CBS*, we can see that our framework with *MLS* and *SGSR* achieves the state-of-the-art score among all the methods in terms of out-domain SPICE, METEOR, average F1 score; and all metrics tested on the in-domain subset. Especially, our F1 score increases 15.4% compared with the previous best score of NBT with *CBS*. Our SPICE score increases 7.6% compared with the previous best score of PS3.

We introduce external knowledge by *SGSR* and *MLS* to assist in generating reasonable pairs. Our proposed FDM-net concentrates on generating valid objects to ensure higher scores on the novel image captioning task. However, existing sentence reconstruction networks cannot generate sentences as commonsensible and fluent as ground truth captions under some circumstances. So it is reasonable to bring a higher score on SPICE which focuses on propositional semantic content, while causes weak performance on CIDEr-D. This is sensitive to the fluency and the correct grammar of sentences.

Human Evaluation Results. To better evaluate the performance on novel object captioning task of our proposed method, we take experiments on marking sentences that are generated by Neural Baby Talk (NBT) and ours by asking several people. We apply two test settings of MSCOCO dataset and Open Image dataset. For MSCOCO dataset, we directly use the out-of-domain testing set in NOC split. While for Open Image dataset, we select 2000 images containing 50 novel classes which do not appear in MSCOCO as the Open Image testing split. Before that, we augment the knowledge of these novel objects to NOC training dataset by FDM-net in the training stage. As shown in Tab. 5, we use Neural Baby Talk (Lu et al. 2018) as our competitor.

The sentences generated by FDM-net achieve better scores over all metrics than NBT on all testing dataset. The scores of our FDM-net are almost two times higher than those of NBT, and it states clearly that our method can generate better sentences than NBT on describing both the whole image and novel objects. Inspiringly, almost 90% of the sentences generated by FDM-net and NBT are easy to read by people. Furthermore, word incorporation scores are lower than image description scores. It reflects a very interesting phenomenon in our work that describing novel objects probably is a bit harder than that of describing images with proper sentences. To sum up, the proposed FDM-net is thoroughly evaluated in this work, and our model on novel object captioning achieves the state-of-the-art results.

Qualitative Performance

Visualization on MSCOCO Dataset with Novel Object. We display one instance which is generated by our proposed FDM-net for each novel object in DCC testing dataset, as shown in Fig. 3(a). The trained model employs 1000 novel object feature-text pairs which are produced by our feature deformation and *SGSR* subsets. We can see that compared with baseline without adding augmented data (first line of Tab. 3), captions generated by our methods can capture novel objects and describe the key objects of images correctly. It is impressive that several sentences can express the coherent content of images, such as the content: 'zebra', 'walking along', 'rocky path' in the first instance.

Preliminary Experiments on Open Images Dataset. One of our primary motivation in this work is to leverage very rich visual concepts from object detection datasets to limited visual concepts from image caption models. We further validate the generalization ability of our method to a large-scale open-conceptual dataset - Open Image. We train our model on MSCOCO image caption dataset. Randomly chosen four classes as novel classes from Open Image dataset. For each novel class, we randomly select 20 images to extract the features for FDM-net with the setting of Base+1000+*MLS*+*SGSR* and another 20 images for evaluation. In Fig. 3(b), we provide some images and the descriptions our method generated. We give two out-domain successful examples – the first two instances in Fig. 3(b), one failure out-domain example – the fourth instance in Fig. 3(b) and an in-domain example – the third instance in Fig. 3(b). We can find that in two out-domain successful examples, the descriptions generated by our model are better than the ones from baseline. In some cases, we find that it will be helpful to describe the in-domain samples, since this novel data deformation method will also increase the number of training samples of these objects. Furthermore, even in the failure case, the produced sentence can still give us the main content of the image. This further demonstrates the effectiveness of our model.

Conclusion

We propose FDM-net combined with the prevailing encoder-decoder framework to tackle the novel object captioning problem. In particular, it is a conceptually simple but powerful approach that generates additional training instances on the feature level. Our FDM-net aims to solve the mismatching problem when doing deformation on the spacial level in vision-language tasks. By creating new feature maps with corresponding texts, we bridge the gap between expected visual information and actually generated visual information. Extensive experiments demonstrate that our approach has achieved the state-of-the-art performance on novel object captioning task.

Acknowledge: This work was supported in part by the Shanghai Research and Innovation Functional Program under Grant 17DZ2260900, and Shanghai Municipal Science and Technology Major Project(19511120700, and 2018SHZDZX01) and ZJLab.

References

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016a. Guided open vocabulary image captioning with constrained beam search. *arXiv preprint arXiv:1612.00576*.

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016b. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, 382–398. Springer.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. 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*, 6077–6086.

Anderson, P.; Gould, S.; and Johnson, M. 2018. Partially-supervised image captioning. In *Neural Information Processing Systems*.

Anne Hendricks, L.; Venugopalan, S.; Rohrbach, M.; Mooney, R.; Saenko, K.; and Darrell, T. 2016. Deep compositional captioning: Describing novel object categories without paired training data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–10.

Banerjee, S., and Lavie, A. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollár, P.; and Zitnick, C. L. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Chen, Z.; Fu, Y.; Chen, K.; and Jiang, Y.-G. 2019a. Image block augmentation for one-shot learning zitian. In *AAAI*.

Chen, Z.; Fu, Y.; Wang, Y.-X.; Ma, L.; Liu, W.; and Hebert, M. 2019b. Image deformation meta-networks for one-shot learning. In *CVPR*.

Koehn, P. 2016. Statistical machine translation. *arXiv preprint arXiv:1612.00576*.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; Bernstein, M.; and Fei-Fei, L. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. In *arXiv preprint arXiv:1602.07332*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2018. Neural baby talk. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7219–7228.

Marr, D. 1982. Vision: A computational investigation into the hu- man representation and processing of visual information. *Cambridge, Massachusetts*.

Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J.; and Lazebnik, S. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, 2641–2649.

Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR* abs/1506.01497.

Rom, H.; Uijlings, J.; Popov, S.; Veit, A.; Belongie, S.; Gomes, V.; Gupta, A.; Sun, C.; Chechik, G.; Cai, D.; Feng, Z.; Narayanan, D.; and Murphy, K. 2017. Openimages: A public dataset for large-scale multi-label and multi-class image classification.

Satoshi Tsutsui, Yanwei Fu, D. C. 2019. Meta-reinforced synthetic data for one-shot fine-grained visual recognition. In *Neural Information Processing Systems*.

Tran, K.; He, X.; Zhang, L.; Sun, J.; Carapcea, C.; Thrasher, C.; Buehler, C.; and Sienkiewicz, C. 2016. Rich image captioning in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.

Venugopalan, S.; Anne Hendricks, L.; Rohrbach, M.; Mooney, R.; Darrell, T.; and Saenko, K. 2017. Captioning images with diverse objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5753–5761.

Wu, Y.; Zhu, L.; Jiang, L.; and Yang, Y. 2018. Decoupled novel object captioner. In *Proceedings of the 26th ACM international conference on Multimedia*, 1029–1037.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, 2048–2057.

Yang, X.; Tang, K.; Zhang, H.; and Cai, J. 2018. Auto-encoding scene graphs for image captioning. *arXiv preprint arXiv:1812.08658*.

Yao, T.; Pan, Y.; Li, Y.; and Mei, T. 2017. Incorporating copying mechanism in image captioning for learning novel objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6580–6588.