

# ReCap Pro: Caption Correction using Meta Learning

Sakshi Birthi<sup>1</sup>, Sanjana Mahesh<sup>2</sup>, Sanskriti Mathuria<sup>3</sup>, Sarang J Chilkund<sup>4</sup>, and Bhaskarjyoti Das<sup>5</sup>

<sup>1</sup> Department of Computer Science and Engineering, PES University, Bengaluru, Karnataka, India

sakshibirthi19@gmail.com, sanjanamahesh2002@gmail.com, sanskritimathuria@gmail.com,

sarangchilkund@gmail.com

<sup>2,3,4,5</sup> Department of Computer Science and Engineering - AI & ML, PES University, Bengaluru, Karnataka, India

bhaskarjyoti01@gmail.com

## ARTICLE INFO

## ABSTRACT

Received: 29 Dec 2024

Revised: 15 Feb 2025

Accepted: 24 Feb 2025

This article presents ReCap Pro, a framework that corrects auto-generated captions by dealing with the possible errors in nouns and verbs in the caption. While caption correction has been attempted earlier, it is observed that it has never been tried as a meta-learning-based approach. The work described in this article offers few-shot learning enabling faster learning with fewer samples of images, solving one of the critical limitations of the traditional data-intensive caption generation models. An object detection model trained using Reptile Meta-Learning is employed to detect the correct nouns and a human object interaction (HOI) detection model trained using Prototypical Networks is used to detect the verbs in the image. The proposed method addresses a long-standing limitation of existing caption generation models that rely on large amounts of training data and can be used as an extra layer of performance enhancer with existing caption generators. The suggested technique can be applied as an additional performance enhancer layer over current caption generators to overcome a long-standing shortcoming of those models.

**Keywords:** Few-Shot learning, Generalization, Reptile Meta-Learning, Prototypical Network, Caption Correction

## 1 INTRODUCTION

Caption correction deals with correcting the captions generated by a caption generation model. Caption generation is an important task in activities like visual question answering, assisting the visually impaired by describing images, searching and retrieving social media images and videos, and many more. An ideal image captioning system should be able to correlate new tasks to tasks previously trained on in a human-like way. However, modern caption generation models like BLIP by Junnan Li et al. [1] and CLIP by Alec Radford et al. [2], have limitations in their generalization ability. This means that when a model trained on an extensive dataset encounters an image from its learned classes but in a different orientation or an image of an unseen class, it has difficulty in correctly analyzing the image.

To address the limitations mentioned above of a traditional deep learning model, this study offers a framework based on a meta-learning based strategy. Meta-learning is learning new concepts in the data without sufficient labeled data. Mike Huisman et al. [3] discusses three kinds of meta-learning strategies in deep learning i.e., optimization-based, metric-based, and model-based. The work depicted in this article adopts the first two in the proposed framework. The Prototypical Network by Jake Snell et al. [4], an improved version of the Siamese network (as shown by Gregory Koch et al. [5]), is a metric-based strategy that relies on learning a similarity metric capable of distinguishing between similar and dissimilar pair of samples. Reptile, as shown by Alex Nichol et al. [6], is an optimization-based approach that improves performance over the Model Agnostic Meta-Learning (MAML) algorithm. The framework proposed uses Reptile and Prototypical Networks for correcting nouns and verbs respectively in the generated captions. The proposed method in this article has the potential to be used along with caption generation models to add an extra layer of error checking.

The following sections of this paper are structured as follows. Section 2 provides an outline of relevant research related to the problem statement or the technology used. Section 3 describes the datasets utilized in the study. Section 4 offers a detailed explanation of the proposed solution. Section 5 delves into the experimental setup used to evaluate

the model. Section 6 presents the results obtained from testing the model, along with a discussion of its limitations and underlying assumptions. Finally, Section 7 outlines potential avenues for further research and concludes the article.

## 2 LITERATURE REVIEW

### 2.1 Caption Correction

Caption correction has been attempted earlier and two caption correction models in particular deserve mention in the context of our work. CAPTION by Leonardo Ferreira et al. [7] aims to correct wrong nouns in a given caption. Since neural networks fail to understand the semantics of the words used to describe an image, natural language processing adds meaning to the terms used in the caption. Phrase Critic by Lisa Anne Hendricks [8] is a caption generator and corrector combined into a single model, which uses an off-the-shelf localization model known as Visual Genome (by Ranjay Krishna et al.) [9] to obtain descriptions of parts of the given image. Thus, to verify the caption generated by an LSTM model, it uses the description generated by Visual Genome.

Both CAPTION and Phrase Critic only detect errors in nouns, not verbs, unlike the proposed ReCap Pro. ReCap Pro differs in its approach as well i.e. it uses two models to detect the nouns/verbs in the image. It verifies the caption with the help of NLP techniques to see if the noun/verb differs from the one detected and works on replacing that particular word. Also, unlike the previous research, it is a few-shot learning approach.

### 2.2 Object Detection model

Object detection in caption correction plays a crucial role in determining whether the nouns mentioned in the caption exist in the image. Notable off-the-shelf models for object detection include Faster-RCNN, DenseNet, and ResNet. Faster-RCNN by Shaoqing Ren et al. [10] enhances traditional RCNN models (by Ross Girshick et al.) [11] by incorporating a Region Proposal Network, resulting in improved speed and accuracy. DenseNet by Gao Huang et al. [12] introduces direct connections between feature maps, increasing model complexity and accuracy but requiring longer processing times. Skip connections used in ResNet-50 help in adding the output of the previous layers to that of the stacked layers, which makes it possible to train much deeper neural networks faster. According to Kaiming He et al. [13], the ResNet architecture also provides good generalization on object detection tasks.

Zhong Ji et al. [14] proposes that using a pre-trained model with meta-learning has many advantages, such as increased flexibility, better learning, and faster convergence. This improves the model's accuracy drastically. According to Alex Nichol et al. [6], Reptile meta-learning is an optimization-based meta-learning algorithm that can be implemented with fewer computational resources. Hence, ReCap Pro uses ResNet-50 with Reptile meta-learning to train the model for object detection and classification.

### 2.3 Human Object Interaction

Human-Object Interaction (HOI) is a multifaceted field of study that examines the dynamic and intricate relationships between humans and objects in various contexts. Along the lines of few-shot learning for HOI, multiple approaches are taken. Metric-based route has been attempted by Xiyao Liu et al., Zhong Ji et al. and Jiale Yu et al. [15–17]. More specifically, Zhong Ji et al. [16] used two models, one to direct the class prototypes to emphasize discriminative regions within Human-Object Interaction (HOI) and the other to employ a new decision method during training, mitigating discrepancies in patterns generated by the same action. Xiyao Liu et al. [15] employed a graph-based methodology where the prototypes are developed by embedding a visual sub-graph to a dynamic graph-metric space. Jiale Yu et al. [17] recognize the fine-grained interactions by employing a hierarchical relation reasoning which incorporates relations among body parts and a contrastive learning mechanism. In another generation-based work by Zhong Ji et al. [18], the prototypes comprise of noun-aware features, verb-aware features, and visual features in each episode. This is done by using two modules, i.e., one that generates the verb-aware and noun-aware features, and the other has the Prototypical Network to find the interaction in the image of the query set. Oytun Ulutan et al. [19] discuss the importance of including spatial features along with the visual features, which shows an increase of 3 mAP over just the visual features. The work for HOI in this study is mainly inspired by [18] and [19].

## 3 DATASET

For the object detection model, a widely recognized dataset, namely CIFAR-10 proposed by Alex Krizhevsky [20] has been used for training which has images of size 32X32. It is a collection of 60000 images belonging to 10 classes.

Each class has 6000 images, further split as 5000 for training and 1000 for testing.

The HICO dataset, seen in Yu-Wei Chao et al. [21], the benchmark dataset for HOI recognition, has been used for the training of the HOI model. It covers 600 categories of human-object interactions (i.e. verb-object pairs such as “ride-bike”) over 117 everyday actions performed on 80 familiar objects. This dataset is made suitable for the few-shot HOI task to include the support and query set images. The authors have followed the common Novel Noun (NN) split strategy where 45 nouns are a part of the meta-train set and 20 nouns are a part of the meta-test set, both of the sets being disjoint. The authors have also removed all the images in the varieties associated with the verb ‘no interaction’, and selected the main label for images.

Finally, a custom dataset has been created to evaluate the overall performance of the caption correction model, as no dataset specifically for caption correction is available for the experimental setup. This dataset comprises 400 image links, each paired with an incorrect and correct caption. The images cover a diverse range of classes that the sub-models have been trained on, and the image links were sourced from the internet, while the inaccurate and accurate captions were manually assigned. The custom data set is made available online.<sup>1</sup>

## 4 METHODOLOGY

In this work, the approach taken to build a caption correction model involves the usage of two sub-models i.e. object detection model to detect the nouns and the human-object interaction model to detect the verbs.

1. An optimization based meta-learning technique; Reptile is used for correcting the nouns by performing object detection. Reptile helps in increasing the generalization of the model, but at the same time does not use as many computing resources as it uses only the first-order derivative of the loss function.
2. The recognition of verbs is done using a human-object interaction model by implementing a metric-based meta-learning technique called Prototypical Networks using a similarity metric to see if the new task is similar to the embedding of any previously learned class.

Finally, the entire caption correction model is tested on the custom dataset with images and captions associated with it.

### 4.1 Training the sub-models

ReCap Pro’s object detection sub-model utilizes a ResNet-50 base model pre-trained on ImageNet, serving solely as a feature extractor by removing its classification layer. New layers, including a classification layer, are added to create the meta-model. The complete object detection model undergoes training using Reptile meta-learning on CIFAR-10. Algorithm 1 reveals that the base model weights remain static, while additional layer weights continuously update. The base model extracts features, while meta-layer weights facilitate effective initialization for learning new classes. Through meta-training, the model converges meta-weights to a more generalized value, refining with each episode. Training involves a 10-way 5-shot setting, totaling 1700 episodes.

**Algorithm 1** Object Detection model using Reptile Meta-Learning

---

**Input:** Image and the corresponding labels of the support and query set for each episode

**Output:** Learned Model

**for every episode do**

train the model on the support set obtain the loss by evaluating the query set

update the learning rate using the formula

$learning\ rate \leftarrow (learning\ rate * (1 - (i/num\ meta\ iterations))) * loss$  - (1) update the model weights

**end**

---

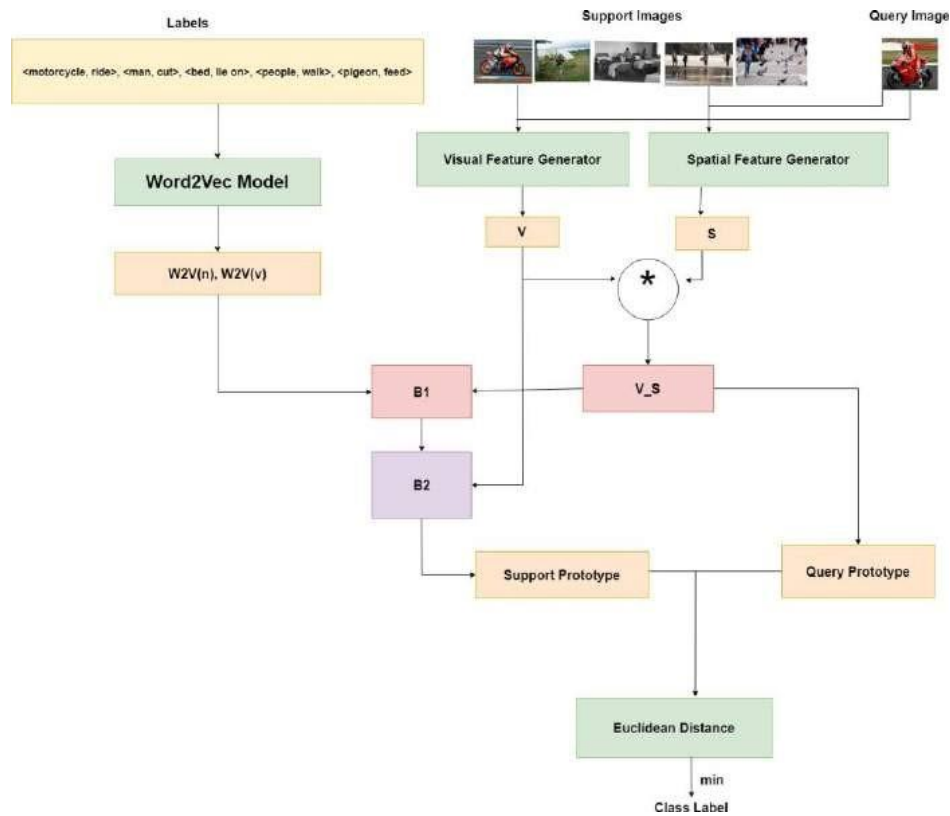
The second sub-model in this study is the human object interaction detection model, referred to as the HOI model. To tackle the challenge of few-shot learning, an episodic training strategy in a 5-way 1-shot configuration has been adopted. In this training approach, each episode is composed of a single image belonging to each class for five classes, resulting in a support set of 5 images, along with a single query image. These support images for each class are selected randomly. Consequently, each episode represents a few-shot learning task, and meta-training is conducted accordingly.

The meta-testing phase follows a similar procedure, where images are sampled from the meta-test set. This approach allows for the evaluation of the HOI model's performance in a few-shot learning context, helping to assess its ability to generalize to new, unseen data.

Algorithm 2 is used for training the HOI model with the help of Prototypical Networks. Figure 1 illustrates the high-level architecture of the model showcasing the usage of the four main components, i.e, visual, spatial, word embeddings, and Block1 and Block2 pair implemented in the training of the model.

The four major components used in the training of the model are:

1. **Visual characteristics:** The initial step involves the loading of a pre-trained ResNet-50 model [13]. Subsequently, the image undergoes pre-processing, and the extraction of visual features is carried out.



**Figure. 1.** The proposed architecture for the HOI model

---

#### Algorithm 2 Training strategy for the HOI model

---

**Input:** Image and the corresponding labels of the support and query set for each episode

**Output:** Learned Model

**for every episode do**  $V = \text{Visual}(\text{image})$   $S = \text{Spatial}(\text{image})$

$W_n, W_v = W2V(\text{noun}), W2V(\text{verb})$

---

```

V_S=V*S
B1N = Block1(V_S,Wn)
B1V = Block1(V_S,Wv)
B2N = Block2(V, B1N) B2V = Block2(V, B1V)
prototypes = λ*(V) + (1-λ)*(B2N + B2V)
loss=PrototypicalLoss(query_V_S, prototypes)
loss.backward()
optimizer.step()
end

```

---

2. **Spatial characteristics:** This approach focuses on image processing through the utilization of a pre-trained YOLOv5 model by Glenn Jocher et al. [22] with the objective of extracting pertinent information for the generation of a spatial configuration map. This map serves to visually represent the positions of identified humans and objects within the image.

The bounding boxes associated with humans and objects are determined using a pre-trained YOLOv5 model, denoted as 'yolov5s'. This information is then used to create a binary spatial configuration map, where a value of 1 signifies the presence of a human or object within the respective region of the image, while a value of 0 denotes absence as done in [19]. To further analyze this spatial configuration map, Convolutional Neural Network (CNN) layers and Global Average Pooling (GAP) are applied.

3. **Word2Vec model:** The Word2Vec model, which is invoked through the Gensim library, is engineered to operate on a word list and generate a transformed vector, encompassing Word2Vec word embeddings. These embeddings are instrumental in extracting semantic information from class labels such that the model acquires semantic awareness of the image.

4. **Block1 and Block2 pair:** Block1 encompasses a network tailored to acquire task representations that encompass both semantic and spatial awareness. This is achieved through the concatenation of spatial and semantic vectors, followed by their passage through a series of fully connected layers. Block2, on the other hand, constitutes a network to analyze the output of block1 along with the visual features and derive the ultimate embeddings and prototypes for the respective classes.

Finally, the prototype for each class in the support set of the episode is obtained using Formula 2, where  $\lambda$  is the hyper parameter used:

$$\text{prototype} = V * \lambda + (1 - \lambda) * (B_{2N} + B_{2V}) \quad (2)$$

The authors employ a custom prototypical loss function, which calculates the Euclidean distance between the generated support set prototypes and the embedding of the query image which is calculated by extracting the visual features and multiplying it with the spatial features. Subsequently, LogSoftmax operation is applied to the distances. It then predicts the class labels based on the probabilities and compares them with the ground truth, thus calculating the accuracy. Finally, the loss is calculated as the log probability of the actual query label and the model is updated with the loss and the optimizer.

## 4.2 Working of the model

---

### Algorithm 3 Working of the Caption Correction Model

---

**Input:** A csv file consisting of three columns, image link, incorrect caption and the correct caption

---

```

Output: Corrected caption Load the csv file
for every row in the csv file do
  image link, caption = read the row
  clean the image url to remove any white spaces if present browse the url and download the image
  image=load the downloaded image tokenized caption=tokenization of caption
  pos tagged caption=pos tagging the tokenized caption
  if pos tagged caption has no verbs or has no nouns indicating humans then
    corrected caption=Object|detection model(image, caption)
  else
    corrected caption=HOI model(image, caption)
  end
return corrected caption
end

```

---

As stated by algorithm 3, firstly, the custom dataset in the csv format is read to obtain the image link and the caption corresponding to the image. The images are then downloaded and saved. The image and the given caption are defined as the input. It has to be made sure that the caption given should contain only one verb and one noun other than the nouns indicating humans such as ‘man’, ‘woman’, ‘female’, ‘male’, ‘couple’, ‘men’, ‘women’, ‘people’, ‘person’, ‘human’. The caption is converted into tokens using the tokenization technique and they are then tagged with the corresponding parts of speech, with the help of POS- tagging using the NLTK library. If a verb is not present in the caption or if a common noun indicating the presence of a person is not present in the caption, the image is only passed to the noun model as it is assumed that there are no verbs to be corrected in the caption. Otherwise, the image is sent to the verb model and the resulting ⟨action, object⟩ pair is detected. Once the labels are detected by the noun/verb model, the incorrect word(s) in the caption, if any, is replaced by the correct word(s), and the correct caption is obtained. The corrected captions are compared with the corresponding captions from the csv file across all images and the accuracy is obtained.

## 5 EXPERIMENTAL WORK

### 5.1 Optimization-Based Meta-Learning for Object Detection

The initial step involves preprocessing the image to suit the ResNet-50 architecture using the TensorFlow library. The labels are then one hot encoded. The base model, ResNet-50, is loaded and the top layers of ResNet-50 are removed. A few layers, namely 2D average pooling, Dense layer, BatchNormalization, and Dropout layer, are added multiple times on top of the model for fine-tuning the base model. This constitutes the Reptile model. The learning rate is set to 0.01, the number of meta-iterations is set to 1700 and the dropout rate is set to 0.2. The optimizer used to compile the model is Adam, categorical cross entropy is used to calculate the loss, and the metrics used to measure the performance is accuracy. The dataset is split into a 10-way, 5-shot setting. The model’s weight and learning rate are updated based on the loss obtained from the evaluation. Using this methodology, we randomly choose one index from the entire training set to add to the query set and repeat it for all ten classes. The model is trained on this support set. The query set has one image from one of the randomly chosen classes among the 10 classes. The model is then evaluated based on this query set.

### 5.2 Prototypical Networks for Human-Object Interaction Detection

The initial step in the image processing pipeline involves preprocessing the input image through resizing, grayscale conversion, and normalization. Subsequently, the visual features are extracted.

Block1 consists of two linear layers, with each having sigmoid as its activation function. Block 2, on the other hand,

has three, each using sigmoid as its activation function. Both blocks have a dropout function which has a dropout rate of 0.2.

The learning rate for the Adam optimizer is set at 0.001, and the regularization parameter is also established at 0.001. Based on experimentation, the value for the hyperparameter lambda is chosen to be 0.5. The model was trained for 500 tasks and tested on 300 tasks.

**Table 1.** Model accuracy results on the CIFAR-10 dataset

Model	Accu- racy
Deep Learn- ing	30.01%
<b>ours</b>	<b>60.29%</b>

## 6 RESULT ANALYSIS

### 6.1 Optimization Based Meta-Learning for Object Detection

We have conducted the experiments on the CIFAR-10 dataset and promising results are obtained as seen in Table 1. The accuracy obtained is 60.29%, out-performing the deep learning model run with the same architecture. The deep learning model's diminished performance is attributed to its requirement of a larger dataset to learn effectively, in contrast to the meta-learning model, which demonstrates improved learning even with a smaller dataset.

### 1.2 Prototypical Networks for Human-Object Interaction Detection

**Table 2.** Model accuracy results on the test-split of the HICO-FS dataset

Model	Type	5-way 1-shot
SGAP-Net [16]	Metric	38.16 $\pm$ 1.65%
DGIG-Net [15]	Metric	39.13 $\pm$ 1.68%
LGM-Net [23]	Genera- tion	35.14 $\pm$ 1.64%
SADG-Net [18]	Genera- tion	39.01 $\pm$ 1.70%
HRM-CL [17]	Metric	40.14 $\pm$ 1.31%
<b>ReCap Pro</b>	<b>Metric</b>	<b>56.1 <math>\pm</math> 1.5%</b>





Experiments have been conducted on the challenging HICO-FS dataset and the results obtained in a 5-way, 1-shot setting are compared with the existing work as can be seen in Table 2. It is observed that this approach, giving an accuracy of 56% outperforms the previous ones by a remarkable margin of about 15%. One possible reason behind this could be the inclusion of spatial features along with visual features in understanding the interactions. This is important because when a human interacts with an object, the correlation between them is of high significance as studied and discussed in [19] as well.

### 6.2 Caption Correction

The caption correction model has been tested using the custom dataset, wherein the images are sent to the sub-models to predict the noun/verb, and the caption is replaced by the predicted labels. It is then checked with the assigned correct caption and the accuracy of the model is found to be 47.07%. This is due to the fact that the distribution of the images the model is trained on and tested on are completely different and reflect a real-world situation.



**Table 3.** ReCap Pro correcting captions generated by BLIP

Image	BLIP caption	Caption Fed to ReCap Pro	Corrected Caption
	A man sitting at a desk	A man sitting at a desk	A man holding at a laptop
	A woman laying in the grass with her head on her hands	A woman laying in the grass	A woman lying in the couch
	A cow grazing in a field of green grass	A cow	A horse
	A man riding a motorcycle on the road	A man riding on motorcycle	A man sitting on motorcycle

An attempt has been made to rectify the captions generated by BLIP [1], as illustrated in Table 3. This table presents instances wherein the captions produced by BLIP [1] were found to be inaccurate, and ReCap Pro effectively corrected the erroneous segments of the generated captions. Differences between the generated captions and those provided to ReCap Pro arise due to the latter assuming the presence of at most one verb and only one noun in the caption.

Example 1. Consider the caption “A woman laying in the grass with her head on her hands”. When confronted with such examples, ReCap Pro forwards the image and caption to the verb model, due the existence of a verb in the caption. Upon image analysis, the model identifies the labels as ⟨“lie on”, “couch”⟩, constituting the ⟨action, object⟩ pair. However, given the presence of three nouns, namely, “grass”, “head” and “hands”, in the caption, the model encounters ambiguity in determining which noun to replace with “couch”.

A similar challenge arises in cases involving multiple verbs. The observations presented in Table 3 indicate that caption generation models may not consistently produce accurate captions, highlighting the potential of the proposed model in this study to enhance the performance of image-captioning systems.

### 6.3 Assumptions and Limitations

This investigation has been carried out using a set of foundational assumptions that underpin its investigative methodology and subsequent findings :

1. The caption contains only one verb and one noun (other than the common noun) as mentioned above.
2. ReCap Pro does not predict the given caption as right or wrong. It instead runs the model every time



and if the predicted noun/verb is the same as the original caption, it does not make any changes.

3. ReCap Pro completely depends on the NLTK library for POS-tagging; if performed incorrectly, the model might make wrong corrections to the caption.

4. If the caption has a noun indicating a human, and the image sent to the verb model does not contain a human, it fails to detect the action. This is because of the fact that the verb model is a *human-object interaction* model.

## 7 CONCLUSION

The work described in this article is the maiden attempt at caption correction using meta-learning. The technique of meta-learning to overcome the drawbacks of a traditional deep learning approach has shown promising results. The article proposes a novel approach to help caption generators perform better. Employing Prototypical Networks for the HOI model and an optimization-based approach for the object detection model have proven to be helpful in achieving a significant 47.07% accuracy. Overall, this article serves as a stepping stone to enhance the performance of a model using lesser samples of data. The authors intend to expand this work by applying the methodology to other parts of speech, including adverbs and adjectives. Additionally, future efforts will address the detection of object-object or human-human interactions.

## REFERENCES

- [1] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.
- [2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.
- [3] Mike Huisman, Jan N Van Rijn, and Aske Plaat. A survey of deep meta-learning. *Artificial Intelligence Review*, 54(6):4483–4541, 2021.
- [4] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.
- [5] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2. Lille, 2015.
- [6] Alex Nichol and John Schulman. Reptile: a scalable metalearning algorithm. *arXiv preprint arXiv:1803.02999*, 2(3):4, 2018.
- [7] Leonardo Ferreira, Douglas De Rizzo Meneghetti, and Paulo Santos. Caption: Correction by analyses, pos-tagging and interpretation of objects using only nouns. 12 2019.
- [8] Sekaran, R., Ramachandran, M., Patan, R., & Al-Turjman, F. (2021). Multivariate regressive deep stochastic artificial learning for energy and cost efficient 6G communication. *Sustainable Computing: Informatics and Systems*, 30, 100522.
- [9] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations, 2016.
- [10] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017.
- [11] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation, 2014.
- [12] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks, 2018.
- [13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.
- [14] Xinge Ma, Jin Wang, Liang-Chih Yu, and Xuejie Zhang. Knowledge distillation with reptile meta-learning for pretrained language model compression. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico

- Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4907–4917, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.
- [15] Xiyao Liu, Zhong Ji, Yanwei Pang, Jungong Han, and Xuelong Li. Dgig-net: Dynamic graph-in-graph networks for few-shot human–object interaction. *IEEE Transactions on Cybernetics*, 52(8):7852–7864, 2021.
  - [16] Zhong Ji, Xiyao Liu, Yanwei Pang, and Xuelong Li. Sgap-net: Semantic-guided attentive prototypes network for few-shot human-object interaction recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11085–11092, 2020.
  - [17] Jiale Yu, Baopeng Zhang, Qirui Li, Haoyang Chen, and Zhu Teng. Hierarchical reasoning network with contrastive learning for few-shot human-object interaction recognition. In *Proceedings of the 31st ACM International Conference on Multi-media*, pages 4260–4268, 2023.
  - [18] Zhong Ji, Ping An, Xiyao Liu, Changxin Gao, Yanwei Pang, and Ling Shao. Semantic-aware dynamic generation networks for few-shot human–object interaction recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.
  - [19] Oytun Ulutan, ASM Iftekhar, and Bangalore S Manjunath. Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13617–13626, 2020.
  - [20] Alex Krizhevsky. Learning multiple layers of features from tiny images. pages 32– 33, 2009.
  - [21] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. Hico: A benchmark for recognizing human-object interactions in images. In *Proceedings of the IEEE international conference on computer vision*, pages 1017–1025, 2015.
  - [22] Glenn Jocher, Alex Stoken, Jirka Borovec, Liu Changyu, Adam Hogan, Laurentiu Diaconu, Jake Poznanski, Lijun Yu, Prashant Rai, Russ Ferriday, et al. ultralytics/yolov5: v3. 0. *Zenodo*, 2020.
  - [23] Huaiyu Li, Weiming Dong, Xing Mei, Chongyang Ma, Feiyue Huang, and Baogang Hu. Lgm-net: Learning to generate matching networks for few-shot learning. In *International conference on machine learning*, pages 3825–3834. PMLR, 2019.