Transformer-only Show-n-Tell with ResNet Descriptor Context-Tuning

Noctis Yamazaki
dept. of Computer Systems Engineering
Northeastern University
Boston, United States
admin@noctis.work

Ming Luo
dept. of Data Analytic Engineering
Northeastern University
Boston, United States
luo.ming1@northeastern.edu

***Abstract*—This paper aims to create an image captioning novel architecture that infuses Grid and Region-based image caption transformer, ResNet, and BART language model to offer a more detail-oriented image captioning model. Conventional state-of-the-art image captioning models mainly focuses on region-based features. They rely on decent object detector architectures like Faster R-CNN to extract object-level information to describe the image's content. Nevertheless, they cannot remove contextual information, high computational costs, and the ability to introduce in-depth external details of objects presented in the images—the replacement of conventional CNN-based detectors results in faster computation. The experiment can generate image captions comparatively fast with higher accuracy and details with contextual information.**

Keywords*—*image captioning, context tuning, grid features, regional features, image classification

I. Introduction

By infusing Language modeling and Computer Vision techniques, image captioning can generate a semantic description of a given image [18]. To strive for better image captioning, the model must fulfill two critical features: the ability to extract extensive features from a given image and contextually relate these features to detailed textual descriptions.

To achieve the goals mentioned above, researchers have derived two ways to extract potential features from images: region-based feature extraction [15] and grid-based feature extraction [12]. Region-based featured extraction only considers the pixel's local spatial information and maintains the original structure. It is capable of avoiding the risk of false object detection. However, it will fall short when the input images have high spatial resolution, and the object and attribute annotations are large-scale compared to grid-based feature extraction [24]. This is because grid features come from the whole image from a high-layer feature map, unlike region-based, which divides into small local areas.

Due to the perplexity of data collection, it is generally difficult to collect and verify the detailed descriptors in the image captions. In existing datasets such as the Microsoft COCO [11], Flickr [7], and Open Images [8], most images only contain general information without detailed object descriptions, such as the breed of the animals and the model of vehicles. Therefore, many of the datasets that could be utilized to develop object classification models are tiny compared to the popular datasets mentioned previously [4]. Moreover, object detection models have been through multiple iterations of improvement if we can use modern models, such as ResNet [6], with external datasets to capture the detailed description of virtual objects in the image.

The region and grid features are saved in the image captioning part as input from some object detector. A deformable transformer for end-to-end object detection is chosen to acquire the region features, and a Swin Transformer is used to extract grid features [12]. Features coming from the last layer are combined and fed into the self-attention Transformer to generate a sequence of words (i.e., caption).

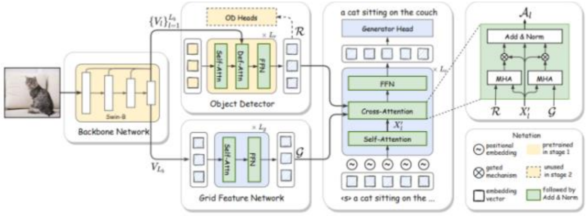


Fig. 1. GRIT architecture [13]

978-1-7281-9523-0/20/$31.00 ©2022 IEEE

Once a caption for an image is generated, if an object matches the existing model's outputs of detailed object classification, the target image will be fed into the object classification model to extract more in-depth information regarding the object represented in the image. A WordNet-based [5] sanity checker is required to ensure that the descriptor given by the external image classification model matches the available object tokens mentioned in the image caption. The additional detailed descriptors combined with the image caption then serve as the inputs for the BART transformer (sequence to sequence modeling) to generate a refined image caption using the context-tuning technique [1].

II. Reraled Work

A. *Feature Extraction for Image Captioning*

Most image captioning models use an encoder-decoder architecture to extract visual features and generate a sequence of words to describe an image. As they rely on Convolutional Neural Networks (CNNs) to extract global characteristics of the input image, their performances suffer due to information loss and insufficient granularity. These shortcomings cannot be alleviated by adding attention mechanisms quickly. Anderson et al. attempted to utilize object detectors, commonly known as Faster R-CNN [20], to extract region features, which leads to higher performance and better accuracy as they are well-known grid-based object detectors and outputs features focusing on prominent objects, but the computational cost remains high. Object detector architecture over vanilla CNNs is that CNNs only convolve through all images, whereas object detectors are end-to-end architecture outputting segmented objects. Also, it has been demonstrated that by switching to the grid feature extraction, an object detector will perform equally well on the visual question-answering tasks [13]. Then, RSTNet [25] was developed to apply grid features to the task of image captioning.

B. *ResNet Image Classification*

Convolutional Neural Networks (CNNs) are often used for image classification. For example, dog and cat breed classification can achieve high prediction accuracy when applying CNN-based pre-trained models, such as Visual Geometric Group (VGG-16) [18] and residual network-50 (ResNet-50). VGG-16 is a convolutional neural network 16 layers deep, while ResNet-50 [22] is a deep residual network 50 layers deep. It applies Batch Normalization and ReLU activation. Although both models are famous for identifying dog and cat breeds, we use the ResNet-50V2 [17] method for the following reasons: first, compared with VGG-16, ResNet-50V2 requires less memory and allows deeper networks to be trained quickly and fast. In addition, VGG-16 tends to cause unpredictable loss due to its constantly learning and relearning nature. However, ResNet-50V2 does not have this problem due to its residual-based learning nature. Therefore, we apply ResNet-50V2 as the pre-trained model.

C. *WordNet Descriptor Validation*

WordNet [5] is an extensive lexical database of English in which words, nous, and verbs are related based on their hierarchy, synonyms, and other vital relations. It is essential in image captioning as a source of knowledge and validator. It is used to validate the relationships between species and organisms, focusing on supporting better cooperation between models. Due to an excessive number of species, the classification classes increase with the number of species. As a result of huge features due to many courses, it causes the model to make predictions with high variance, which is time-consuming and leads to less accurate results.

D. *BART Transformer Model for Text Summarization*

Unlike BERT, which is only an encoder, Bi-Directional Auto Regressive Transformer (BART) [9] is a sequence-to-sequence de-noising autoencoder and decoder architecture. It is widely used in problems of sequence-to-sequence modeling. Text summarization is one of the application areas where BART is used. Text summarization is mainly composed of extractive text summarization and abstractive text summarization. Extractive summarization makes a summary by shortlisting the most critical lines from the text, whereas abstractive text summarization summarizes semantically as humans do.

III. Proposed Model

This section describes the object classification used in the model to offer detailed descriptors to image captioning, caption generator, and BART for contextual tuning.

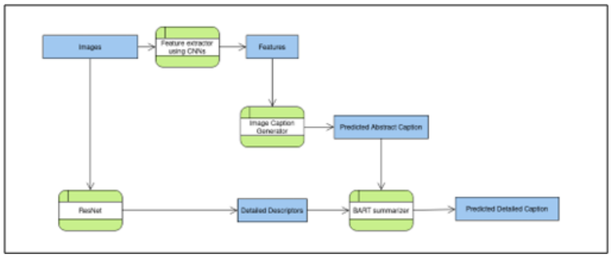

Images
Feature extractor using CNNs
Features
Image Caption Generator
Predicted Abstract Caption
ResNet
Detailed Descriptors
BART summarizer
Predicted Detailed Caption

Fig. 2. Proposed methodology

Providing detail descriptors consists of two parts, one for extracting the features from the input image and the other for integration checks and training caption generator for generalized captions using WordNet.

A. *Extracting Detailed Visual Features from Images*

First, all images were encoded into the numeric format and stored in RGB format in the resized shape of 244x244. Moreover, all the pixel values were normalized in the range [-1, 1]. Second, the position and value of pixels were changed to increase the modified versions. By this step, the model can be more generalized. Therefore, the classification accuracy can be improved. Hence, we build a ResNet50V2 [17] model with 80% of the dataset. Next, the model was trained with the RMSprop optimizer [27] for 100 epochs. In the end, the outputs can be used as part of the inputs for the WordNet integration checker.

B. *WordNet Species and Biological Relationship Validation*

A taxonomic unit is a group within a taxonomic hierarchy, from the broadest taxonomic "domain" to the most basic taxonomic "species." We combine Wordnet and image classification into a learning framework in our work, mainly extracting words about "species" from Flicker30K. The word at each vertex represents a category under animal taxonomy, and the upward arrow for each word means it has a hypernym. The structure can be visualized in Fig. 2. Each synset can have multiple hypernyms [2], so the intergenerational hypernyms (the hypernyms of the searched word) may not be related to the searched words. Our WordNet graph is a branching tree to solve this problem. Each tree branch is a single directed path and leads to a unique root. "Animal" will be the endpoint to search for roots of different species. Find the shortest ancestral path from the most basic "species" to the animal "Animalia" with WordNet hypernyms.

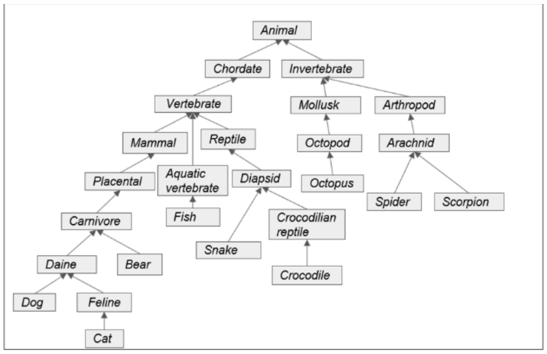


Fig. 3. Overview of finding taxonomy of species through WordNet hypernyms

In addition to using taxonomic methods to group and categorize species of living things, we also use brands to classify species further. Breeds refer to specific populations with morphological and characteristic features in common with species, and species are the largest group that can produce fertile offspring through reproduction. To reduce the confusion of features caused by too many breeds, we use WordNet to classify populations of different species as the same species, which is based on DNA barcodes to help us identify and describe species [19]. Wordnet can identify the hypernym of each breed and locate it in the species to which it belongs.

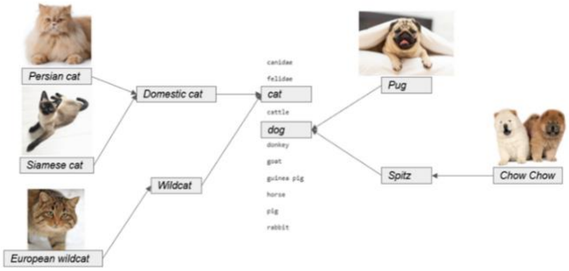


Fig. 4. Perplexity of the tracing of our models

To reduce the training time of breed image classification and improve the accuracy. Using WordNet helps reduce the time-consuming and low accuracy of the model caused by too many features. We set a text document consisting of 200 species names as our verification points and then performed ancestral paths for various breeds through WordNet hypernyms. This helps us to reduce the path to finding common breed names for common animals. This allows each model to concentrate on specialization and a more efficient collaborative environment between models.

By replacing breeds and species with their generalized animal category, like a dog, the caption generator model was trained to produce generalized captions with less variance, assuring language semantics.

Before the detailed descriptors generated from the ResNet object classification model are sent to the BART module for context-tuning [1], we trace the breed names offered by the object classification model to ensure that they match the categories of animals presented in the image caption. By doing so, we can significantly avoid the false classification error generated by the ResNet model.

C. BART Context-Tuning

To ensure that we can infuse detailed descriptors into existing image captions, we need to ensure that our model can produce readable and plausible text in human languages. However, it is difficult for us to achieve the goal by relying on the existing image caption datasets due to their vocabulary and grammar variation limitations. To address the shortcomings, we rely on Context-Tuning on the current pre-trained model developed based on an enormous corpus [1]. In this way, we will be able to provide enough semantic information.

To train the model to fill the corresponding detailed descriptors into the correct locations, in the conventional BERT model [3], "[MASK]" tokens combined with general texts will be used as inputs. However, in our model, we construct the input sequence in the following manner, as shown in Fig. 4:

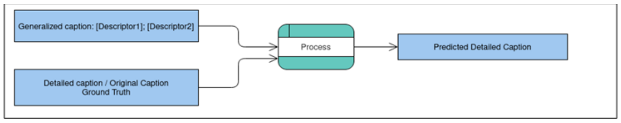

Generalized caption: [Descriptor1]; [Descriptor2]
Detailed caption / Original Caption Ground Truth
Process
Predicted Detailed Caption

Fig. 5. BART text summarization input streams

Problem can be modeled as sequence-to-sequence generation. BART is used for this task, as it has promising results in summarization tasks and can produce captions taking in context detailed descriptors. The caption with detailed descriptors can not only be as simple as the original image caption with abstract object word replaced, but it can also be tuned with grammatical changes. We are trying to output detailed caption with these descriptors and maintain language semantics. As shown in Fig.4, our model has an input of a sequence of captions followed by descriptors. By meddling with the input streams, we can make our captions more accurate to life and easy to read for the audience.

IV. EXPERIMENTS

A. Datasets

1) **Object Detector** In our model, we trained our object detectors in two steps. In the first step, the show-and-tell model (caption generator model) is prepared based on COCO datasets. We also exclude the duplicated samples in the testing and validation splits of the COCO dataset.

For a detailed object classification model, we trained the ResNet model using the Stanford Dogs Dataset [26], which contains 20,580 dog images of 120 breeds, and the Oxford-IIIT Pet Dataset [16], which includes 2371 cat images with 12 breeds. Since each breed has a similar count of images, the dataset is in balance. We first split 80% of images into training and 20% into the test dataset. To ensure the training and test datasets are balanced, we split the dataset by breed and merge the corresponding proportional images into each dataset. Regarding final benchmarks, the Flickr8K [7] dataset was then used for testing the performance of the trained models.

2) Image Captioning COCO dataset was used to train the language decoder in our experiments. The dataset contains 123,287 images, each annotated with five different captions. As mentioned previously, the Flickr8K dataset is also used for benchmark purposes to evaluate the effectiveness of the trained model on other image captioning datasets.

B. Hyperparameters

In our experiments, all models have each layer's dimension set to 1024. For a fair comparison, we set the number of layers for object detector, grid feature detector, as well as caption generator to 6. Such settings are shared across VGG-16, ResNet, InceptionV3 [21], and GRIT models.

C. Performance Analysis

For our experiment, we mainly focus on the CIDEr-d score [23], n-gram BLEU score [14], and ROUGE score [10] of the VGG-16, InceptionV3, and GRIT models based on the Microsoft COCO dataset. CIDEr-d score is used for object detection where other scores are designed for image caption benchmarks.

The CIDEr-d score [23] measures a sentence's consensus with how most people describe the image. The evaluation protocol is based on human participants judging the sentence similarity between candidate sentences and ground truth sentences of pictures and then comparing the generated sentences by describing the images with human-generated sentences. In comparing two sentences, the CIDEr-d score captures their similarity, grammaticality, salience, and accuracy and then performs a quality score on the sentence generated from the image description. The quality score represents how similar the image description-generated sentence is to the human-generated sentence.

The BLEU score [14] is used to evaluate the quality of the translation work of the natural language processing system. Our evaluation protocol compares the similarity between the n-grams of sentences produced by the model and the n-grams of human-translated sentences. N-gram refers to a series of words selected to appear in the sentence, and the number of words is determined by n. The BLEU score for the model has nothing to do with the position of a series of words selected by n-gram. The range of the BLEU score is between 0 and 1, and the closer the BLEU score is to 1, the more the sentences produced by the model match the human-translated sentences.

ROUGE-N [3] is similar to the BLEU score in that ROUGE-N measures the number of matching n-grams between sentences generated by our model and those translated by humans. In ROUGE-N, N represents the n-gram we are using. ROUGE-N counts the recall rate, while BLEU is mainly based on the precision rate, so ROUGE-N's model scoring is different from Blue score's model scoring, and the results of ROUGE-N and Blue score are complementary.

TABLE I. RESULTS OF EVALUATION BASED ON THE MICROSOFT COCO DATASET

Model	**CIDEr-d**	**BLEU**				**ROUGE**		
		1-gram	*2-gram*	*3-gram*	*4-gram*	*R-1*	*R-2*	*R-l*
VGG-16	115.3	0.547	0.250	0.206	0.293	0.154	0.020	0.137
Model	**CIDEr-d**	**BLEU**				**ROUGE**		
		1-gram	*2-gram*	*3-gram*	*4-gram*	*R-1*	*R-2*	*R-l*
InceptionV3	120.2	0.550	0.377	0.246	0.300	0.197	0.085	0.172
GRIT	142.3	0.738	0.470	0.351	0.294	0.423	0.170	0.327
GRIT+ResNet	142.3	0.726	0.459	0.383	0.299	0.415	0.161	0.343

Based on the results shown in TABLE I, it is evident that our proposed model, while maintaining high object detection capabilities, it also shines in offering better image captions than the conventional models. However, once we combine the detailed descriptors from the ResNet models, the image caption quality will seem inferior on the paper as the added descriptors can sometimes be several tokens longer than the original caption.

D. Testing on Flickr8K Dataset

Next, we offer a more holistic look on the real-world performance of our model against other traditional models, we performed tests on Flickr8K datasets. The average runtime for an image in Flickr8K dataset [7] using each of the four models is shown in TABLE II.

TABLE II. RESULTS OF EVALUATION BASED ON THE MICROSOFT COCO DATASET

Model	**VGG-16**	**InceptionV3**	**GRIT**	**GRIT+ResNet**
Average Runtime (second)	1.201	0.9762	0.6285	0.7015

While our proposed model not only achieves maximum Runtime in the benchmark, the added ResNet features along with BART context tuning and WordNet validation do not heavily tax the performance and accuracy of the model. Some sample results are listed in Appendix A.

V. SUMMARY AND CONCLUSION

This paper proposes a Transformer-based architecture for GRIT, ResNet, and BART. It integrates the grid- and region-based features extracted from input images to give better visual information. Moreover, the ResNet image classification and WordNet-based validation modules can offer accurate and richer descriptions of the specific object represented in the images. Furthermore, the BART transformer has demonstrated its ability to infuse simple image captions and detailed descriptors to develop detailed human-like captions. The experimental results have validated our approach's performance and accuracy compared to the conventional image captioning model in inference accuracy.

The lightweight portable object classification model combining with pre-trained image caption model can help users to reduce the development and training costs of their applications. The success of our proposed model could enable the end users to develop more suitable custom object detection models based on their industrial needs. For instance, local parking management could establish an object detection model based on different make and models of vehicles, along with a car plate recognition model. The image caption model could then describe the condition of the parked vehicle along with other helpful information from the image to offer documentation recording in the text to reduce the need for higher storage space led by images.

For future development, if users intend to increase the capacity and diversity of detailed descriptors in the model, an object detector model using existing generalized datasets such as Flicker and Microsoft COCO will not help to yield optimal detection accuracy. Therefore, a more robust object detector model is needed for future improvements. Besides, as grammar tuning is still based on a user-defined corpus, experiments on the large external corpus or automated corpus generation for BART context-tuning are crucial for future development.

References

[1] Carnegie, N. B., & Wu, J. (2019). Variable selection and parameter tuning for BART modeling in the Fragile Families Challenge. *Socius: Sociological Research for a Dynamic World, 5*, 237802311982588.

[2] Choi, W.S., On, K., Heo, Y., & Zhang, B. (2020). Toward General Scene Graph: Integration of Visual Semantic Knowledge with Entity Synset Alignment. *ALVR*.

[3] Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv, abs/1810.04805*..

[4] Elhagry, A., & Kadaoui, K. (2021). A Thorough Review on Recent Deep Learning Methodologies for Image Captioning. ArXiv, abs/2107.13114.

[5] Fellbaum, C.D. (2000). WordNet: an electronic lexical database. Language, 76, 706.

[6] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770-778.

[7] Kiros, R., Salakhutdinov, R., & Zemel, R.S. (2014). Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. ArXiv, abs/1411.2539.

[8] Kuznetsova, A., Rom, H., Alldrin, N.G., Uijlings, J.R., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., Duerig, T., & Ferrari, V. (2020). The Open Images Dataset V4. International Journal of Computer Vision, 128, 1956-1981.

[9] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. ACL.

[10] Lin, C. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. *ACL 2004*.

[11] Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C.L. (2014). Microsoft COCO: Common Objects in Context. *ECCV*.

[12] Lu, J., Xiong, C., Parikh, D., & Socher, R. (2017). Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3242-3250.

[13] Nguyen, V., Suganuma, M., & Okatani, T. (2022). GRIT: Faster and Better Image captioning Transformer Using Dual Visual Features. *ArXiv, abs/2207.09666*.

[14] Papineni, K., Roukos, S., Ward, T., & Zhu, W. (2002). Bleu: a Method for Automatic Evaluation of Machine Translation. ACL.

[15] Parameswaran, S.N., & Das, S. (2018). A Bottom-Up and Top-Down Approach for Image Captioning using Transformer. *Proceedings of the 11th Indian Conference on Computer Vision, Graphics and Image Processing*.

[16] Parkhi, O.M., Vedaldi, A., Zisserman, A., & Jawahar, C.V. (2012). Cats and dogs. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 3498-3505.

[17] Phan, A., Nguyen, N., Trieu, T., & Phan, T. (2021). An Efficient Approach for Detecting Driver Drowsiness Based on Deep Learning. *Applied Sciences*.

[18] P You, Q., Jin, H., Wang, Z., Fang, C., & Luo, J. (2016). Image Captioning with Semantic Attention. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4651-4659.

[19] Radev, I., & Kancheva, Z. (2021). Handling synset overgeneration: Sense Merging in BTB-WN. *Proceedings of the Student Research Workshop Associated with RANLP 2021*.

[20] Ren, S., He, K., Girshick, R.B., & Sun, J. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*, 1137-1149.

[21] Szegedy, C., Vanhoucke, V.,Process Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2818-2826.

[22] Vatathanavaro, S., Tungjitnob, S., & Pasupa, K. (2018). White Blood Cell Classification: A Comparison between VGG-16 and ResNet-50 Models.

[23] Vedantam, R., Zitnick, C.L., & Parikh, D. (2015). CIDEr: Consensus-based image description evaluation. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4566-4575.

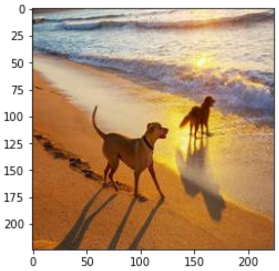
[24] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A.C., Salakhutdinov, R., Zemel, R.S., & Bengio, Y. (2015). Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *ICML*.

[25] Zhang, X., Sun, X., Luo, Y., Ji, J., Zhou, Y., Wu, Y., Huang, F., & Ji, R. (2021). RSTNet: Captioning with Adaptive Attention on Visual and Non-Visual Words. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15460-15469.

[26] Zhao, P., Xie, L., Zhang, Y., & Tian, Q. (2021). Universal-to-Specific Framework for Complex Action Recognition. *IEEE Transactions on Multimedia, 23*, 3441-3453.

[27] Zou, F., Shen, L., Jie, Z., Zhang, W., & Liu, W. (2019). A Sufficient Condition for Convergences of Adam and RMSProp. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11119-11127.

Appendix A. Additional Samples for Image Captions

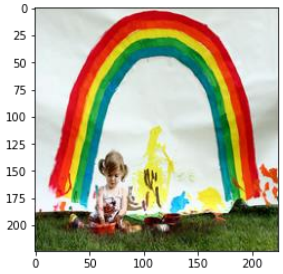


VGG-16: dogs playing together on a beach

InceptionV3: dogs play together on the beach

GRIT: a brown dog and a yellow dog play on the beach

GRIT+ResNet:
A brown Chihuahua and a Cane Corso play on the beach

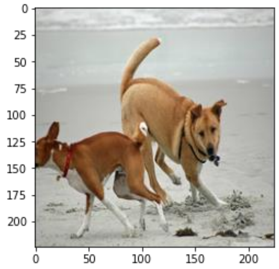


VGG-16: little girl covered in paint sits in front of a painted rainbow with her hands in a bowl

InceptionV3: a boy and blue umbrella up against a large stuffed animal

GRIT: little girl covered in paint sits in front of a painted rainbow and a bowl

GRIT+ResNet: little girl covered
in paint sits in front of a painted rainbow and a bowl

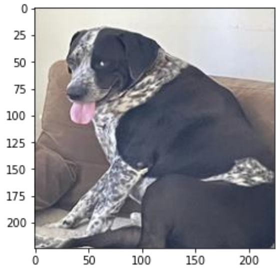


VGG-16: woman in a blue jacket rides a brown pony near a water

InceptionV3: a dog is looking down a beach

GRIT: a brown dog and a black dog are on the beach

GRIT+ResNet:
a brown Labrador Retriever and a black Golden Retriever are on the beach

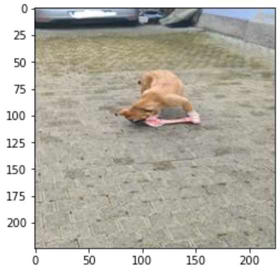


VGG-16:
man playing with a black dog on a white blanket

InceptionV3: a man is on a bed

GRIT:
a dog is sitting on a coach

GRIT+ResNet:
a Brittany is sitting on a coach

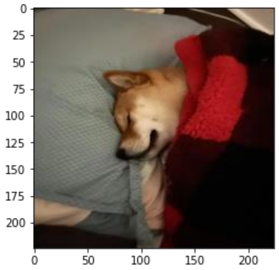


VGG-16: dog is running on the beach

InceptionV3: a young boy with his frisbee

GRIT: a dog is playing on the grass

GRIT+ResNet:
a Labrador Retriever is playing on the grass



VGG-16: small boy putting something in his mouth with both hands

InceptionV3: a black and white kitten on a banana a bird

GRIT: a kitten is sleeping on the bed

GRIT+ResNet: a kitten is sleeping on the bed