The Impact of Synthetic Data on Image Captioning Models

Jifeng Wu
Department of Computer Science
25952896
jifengwu2k@gmail.com

Bowen Zhang
Department of Electrical and Computer Engineering
49541030
zbowen12@student.ubc.ca

Ziyu Wang
Department of Electrical and Computer Engineering
44599561
w23z26y25@gmail.com

Abstract

Significant improvements in the quality and availability of generative models have led to widespread synthetic data across the internet. Motivated by previous work on synthetic data for image classification, this paper addresses the challenging task of image captioning, bridging computer vision and natural language processing. The influence of synthetic data on image captioning models is investigated through systematic experiments conducted with the “Show and Tell” model. The experiments found that combining real and generated data can hurt the performance and robustness of the model. However, models trained solely on synthetic images outperform those on an all-real training set when fine-tuned, highlighting the potential of using AI-generated images in pretraining datasets for image captioning. These findings contribute valuable insights for advancing the field, providing a foundation for future exploration into factors influencing this paper’s conclusions, alternative methods for image generation, and the broader impact of synthetic data on diverse image-based tasks.

1 Introduction

In recent years, people have been paying more attention to using synthetic data in training image processing models. This is important because it has been proven that AI-generated images from an “in-the-wild” Stable Diffusion model can make image classification models better by improving accuracy and robustness Bansal and Grover [2023]. This exciting result and the growing stable diffusion models motivate this project. This paper focuses on using synthetic data in image captioning, which is a task involving both computer vision and language understanding that is much more challenging than image classification Vinyals et al. [2015]. The project is designed to have three parts: (1) creating AI images using stable diffusion; (2) training image captioning models with synthetic data and real data; (3) comparing models that were trained with different ratios of real and fake data. The study uses the classic "Show and Tell" image classification model Vinyals et al. [2015] for experiments. The results reveal that models trained only with synthetic images, when fine-tuned on the Flickr8k dataset, perform better than those trained only with real images. This research provides useful information for improving the use of fake data in image captioning models. It opens the door for future exploration into different factors, methods, and impacts on various image-related tasks.

The rest of the paper is organized as follows. Initially, an exploration of the literature on image captioning and diffusion models will be presented. Subsequently, the clarification of models such as “Show and Tell” and Stable Diffusion will be undertaken, incorporating theoretical frameworks and graphical representations. Following this, the paper will provide a concise overview

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

of the experiment design and a comprehensive discussion of the results. Finally, the paper will be concluded, and potential directions for future research will be pointed out.

2 Related Work

2.1 Image Captioning

Image captioning involves describing the visual content of an image using meaningful, syntactically correct sentences. Early approaches to image captioning included description retrieval Pan et al. [2004], Farhadi et al. [2010], Ordonez et al. [2011], Frome et al. [2013], Kiros et al. [2014], Karpathy et al. [2014] and template filling with hand-crafted language generation techniques Yao et al. [2010], Aker and Gaizauskas [2010], Yang et al. [2011], Li et al. [2011], Gupta et al. [2012], Mitchell et al. [2012], Kulkarni et al. [2013], Kuznetsova et al. [2014]. These approaches have been surpassed by deep learning-based generative models, where image inputs are encoded as feature vectors before using a language model to decode a sequence of words from the feature vectors.

Encoding images as feature vectors is the first challenge of an image captioning pipeline. Current visual encoding methods are categorized into four types Stefanini et al. [2022]: non-attentive methods, additive attentive methods, graph-based methods, and self-attentive methods. Non-attentive methods, such as Show and Tell Vinyals et al. [2015] and SCST (FC) Rennie et al. [2017], utilize high-level representations from CNN layers, offering simplicity and a comprehensive grasp of the image context but lacking in specificity and granularity. Additive attentive methods can be further divided into methods that use attention over a grid of CNN features, including Show, Attend and Tell Xu et al. [2015] and SCST (Att2in) Rennie et al. [2017], and methods that use attention over visual regions, including Up-Down Anderson et al. [2018]. Using attention over a grid of CNN features enhances granularity by using 2D activation maps and additive attention, allowing the model to selectively focus on image elements while using attention over visual regions and utilizing object detectors proposing regions for focused attention. Graph-based Encoding, represented by SGAE Yang et al. [2019] and MT Shi et al. [2020], uses graphs over image regions to encode semantic and spatial connections, enhancing the representation but potentially limiting interactions between visual features. Lastly, Self-Attention Encoding supported by the Transformer architecture, including AoANet Huang et al. [2019], X-LAN Pan et al. [2020], DPA Liu et al. [2020], connects each set element (such as each patch of an image) with all others for refined representation, with various adaptations like geometry-aware encoding and memory-augmentation enhancing its capabilities.

Afterward, language models are used to predict the probability of a given sequence of words occurring in a sentence. LSTM-based models are predominant due to their sequential processing capability, with single-layer LSTMs (as used in such as Show and Tell Vinyals et al. [2015] and Show, Attend and Tell Xu et al. [2015]) being straightforward in their approach, and two-layer LSTMs (as used in Up-Down Anderson et al. [2018]) offering more complexity by separating visual attention and language modelling tasks. Some LSTM models are further enhanced with self-attention (as used in AoANet Huang et al. [2019], X-LAN Pan et al. [2020], DPA Liu et al. [2020], and AutoCaption Zhu et al. [2020]) for improved performance. Convolutional language models, though less common, use convolutions to process global image features and word embeddings, offering parallel training advantages but lacking in popularity. Transformer-based architectures (as used in ORT Herdade et al. [2019] and M2 Transformer Cornia et al. [2020]) have revolutionized the field with their fully-attentive paradigm, employing masked self-attention and cross-attention operations for more effective language generation. BERT-like architectures (as used in Unified VLP Zhou et al. [2020] and VinVL Zhang et al. [2021]) merge visual and textual modalities early on, utilizing pre-trained text layers for more efficient learning and are often pre-trained on extensive image-caption pairs. Lastly, non-autoregressive language models aim to reduce inference time by generating all words in parallel, often involving multiple generation stages and reinforcement learning for enhanced results.

2.2 Diffusion Model Applications

Diffusion models have been effectively utilized in a variety of fields, such as image Ho et al. [2020, 2022b], Ho and Salimans [2022], speech Chen et al. [2020], Kong et al. [2020], and video Ho et al. [2022a], Singer et al. [2022], Villegas et al. [2022] generation, and image processing tasks Song et al. [2020], Saharia et al. [2022a,c], Wang et al. [2023] including image colorization, super-resolution, inpainting, and semantic editing. A key application area for diffusion models is in producing high-resolution images through large-scale text-to-image generation. Models like Stable Diffusion Rombach et al. [2022], DALL-E Ramesh et al. [2022], Imagen Saharia et al. [2022b], eD-iff Balaji et al. [2022], and GLIDE Nichol et al. [2021] have been notable for creating detailed images.

Although the use of large-scale diffusion models to support downstream tasks is still evolving, recent developments have shown their potential to enhance training data for machine learning. For example, He et al. [2022] demonstrated that using synthetic data generated by GLIDE Nichol et al. [2021] can significantly improve the performance in zero-shot and few-shot image classification tasks. Additionally, a customized dataset created by adapting GLIDE to CIFAR-100 images has been shown to increase the accuracy in CIFAR-100 image classification notably. In addition, Brooks et al. [2023] fine-tuned a stable diffusion model for image editing with creative image-text pairs generated with a combination of GPT-3 and Stable diffusion. Finally, Bansal and Grover [2023] employed zero-shot generated data from an "in the wild" stable diffusion model to train robust image classifiers against natural distribution shifts.

3 Model/Method

3.1 Show and Tell

Show and Tell Vinyals et al. [2015] is a classical neural and probabilistic framework for image captioning. At the core of Show and Tell is the concept of treating the process of generating descriptions from images as a translation problem. Traditionally, statistical machine translation focuses on converting a sentence from a source to a target language. Show and Tell extends this concept by considering the image as the source language and its descriptive text as the target language. It involves a two-step process: encoding the visual information in an image and decoding it into a coherent textual description, as shown in Figure 1.

The first step of Show and Tell involves using Convolutional Neural Networks (CNNs) to process and encode visual inputs. CNNs have demonstrated exceptional performance in image recognition and classification tasks. In Show and Tell, the CNN acts as a feature extractor that converts an image into a vector. This vector serves as a condensed representation of the visual information, capturing essential details required for generating an accurate caption.

Once the image is encoded into a vector, Show and Tell employs Recurrent Neural Networks (RNNs) in the form of Long-Short Term Memory (LSTM) networks to decode it into text. Specifically, the LSTM network takes the image vector as input and generates a sequence of words that form a coherent caption. The LSTM's memory cell encodes information about the image and the sequence of words generated thus far, enabling the generation of contextually relevant and grammatically coherent sentences.

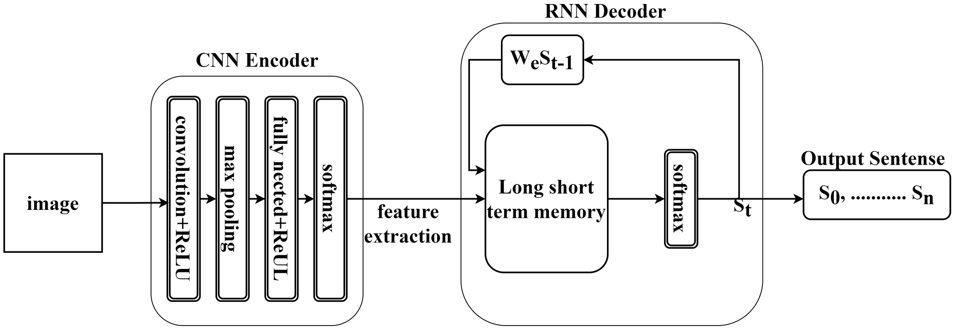

Figure 1: Show and Tell Model Architecture with CNN Encoder

The training of Show and Tell involves maximizing the probability of generating the correct caption for a given image. This is achieved by modelling the conditional log-likelihood of each word in the caption, given the image and the preceding words. If we denote the input image as I and a ground-truth caption describing this image as $S = (S_0, \ldots, S_N)$, Show and Tell's loss for this image-caption pair is as follows:

$$L(I, S) = -\log p(S|I) = -\sum_{t=0}^{N} \log p(S_t | I, S_0, \ldots, S_{t-1})$$

3.2 Diffusion

The diffusion model is a class of generative models that have gained significant attention in recent years for its ability to produce high-quality, diverse samples, particularly in the domain of image generation. Unlike traditional generative models such as Generative Adversarial Networks (GANs) or Variational Autoencoders (VAEs), diffusion models employ a unique mechanism based on the principles of stochastic diffusion processes.

The essence of diffusion models lies in their two-phase process: the forward diffusion phase, which gradually corrupts the data, and the reverse diffusion phase, which aims to recover the original data. This process is analogous to first adding noise to an image and then learning to remove this noise to retrieve the original image.

In the forward diffusion phase, the model incrementally adds Gaussian noise to the original data over a series of steps. This process is typically defined over a discrete time scale, starting from $t = 0$ and incrementally increasing to a final time step $t = T$. At each step, the data moves further from its original state, becoming progressively noisier until it reaches a point where the original information is almost entirely obscured.

The reverse diffusion phase is where the model demonstrates its generative capabilities. Starting from a heavily noised state, the model learns to reverse the noise addition process. Specifically, the sample at the current time step x_{t-1} is drawn from a Gaussian distribution $N(\mu_\theta(x_t), \Sigma_\theta(x_t))$, whose mean $\mu_\theta(x_t)$ is calculated with the sample from the previous time step x_t with a trained neural network μ_θ, and whose variance $\Sigma_\theta(x_t)$ follows a fixed schedule. The model iteratively refines this process, ultimately reconstructing an approximation of the original data or creating new samples.

Expanding upon the basic diffusion model framework, conditional diffusion models incorporate external conditioning variables such as class labels or textual descriptions, as shown in Figure 2. This allows for generating targeted content, making diffusion models versatile for applications requiring specific types of outputs, such as class-specific image generation or text-to-image synthesis.

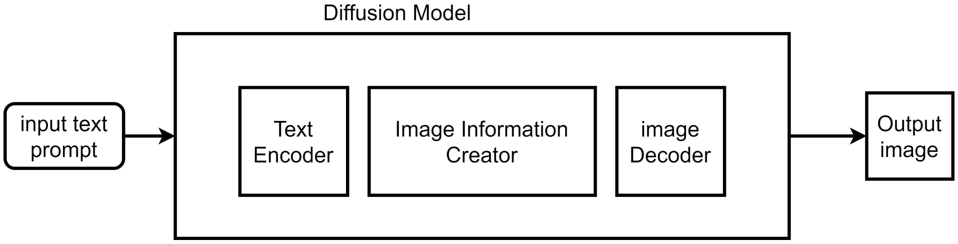

Figure 2: Text to Image Diffusion Model

4 Experiments

4.1 Research Questions

Inspired by the conclusions in Bansal and Grover [2023], the paper proposes the following research questions:

- Can training with a mix of real and synthetic data improve Show and Tell's performance?
- Can training with a mix of real and synthetic data improve Show and Tell's robustness?
- Can pre-training with synthetic data and fine-tuning with real data improve Show and Tell's performance?
- Can pre-training with synthetic data and fine-tuning with real data improve Show and Tell's robustness?

4.2 Metrics

The metric that this paper used to quantify the performance of Show and Tell is the BLEU (Bilingual Evaluation Understudy) score Papineni et al. [2002], which is a modified n-gram precision score between a generated sentence and a set of reference sentences.

Originally proposed for automatic evaluation of machine translation, the BLEU score is also the most commonly used metric in the image captioning literature. Even though this metric has some obvious drawbacks, it is quick, inexpensive, and has been shown to correlate well with human evaluations.

4.3 Dataset

Several datasets, as described in Table 1, consisting of images and caption sentences describing those images are used during the experiment. In each dataset, each image has been annotated by labellers with five relatively visual and unbiased sentences.

Dataset	Training	Validation	Testing	Total
Pascal_VOC_2008 Farhadi et al. [2010]			1000	1000
Flickr8k Rashtchian et al. [2010]	6000	1000	1000	8000
Flickr30k Young et al. [2014]				30000
COCO2014 Lin et al. [2014]	82783	40504		123287

Table 1: The statistics of the datasets

Due to the high cost of generating synthetic data, the team only generated synthetic data for the smallest dataset with a training and validation/testing set, Flickr8k, and trained Show and Tell on a mixture of real and synthetic images from the Flickr8k dataset. The other datasets are used to evaluate the robustness of the trained model.

4.4 Data Generation

The team utilized Stable Diffusion Rombach et al. [2022] to generate synthetic data conditioned on the captions of the images in the dataset. Specifically, based on Bansal and Grover [2023], the team used the Stable Diffusion-V1-5 implementation and inference settings detailed in the diffusers von Platen et al. [2022] library. For Flickr8k Rashtchian et al. [2010], the team constructed a 6k generated training dataset, 1K validation dataset, and 1K testing dataset from Stable Diffusion by conditioning on the captions for the images.

4.5 Training Details

Supervised learning models require large amounts of data. However, high-quality datasets often comprise fewer than 100,000 images. Thus, to combat overfitting, the team initialized the Convolutional Neural Network (CNN) component of Show and Tell with weights from a pre-trained model (like those trained on ImageNet) with all other weights randomly initialized (initializing word embeddings from a large news corpus led to no significant gains), as described in Vinyals et al. [2015].

The training approach involved stochastic gradient descent with a fixed learning rate of 0.001 and no momentum. The embeddings and LSTM memory were set to 512 dimensions each.

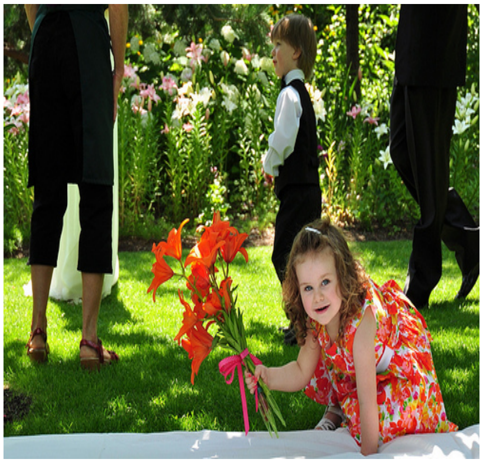

Figure 3: The real image in Flickr8k

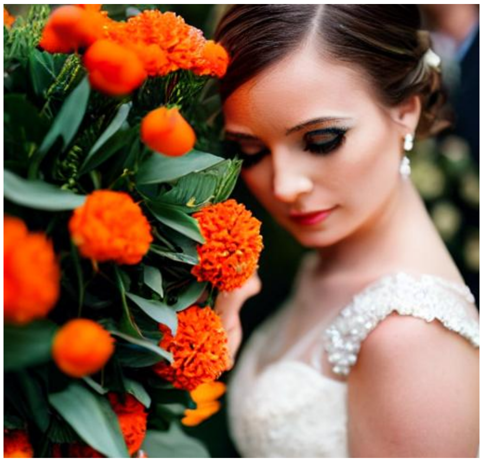

Figure 4: The AI-generated image

Caption of image: A girl at a wedding holding some orange flowers.

Basic tokenization is used for data preprocessing. All words that appeared at least four times in the training dataset are retained.

4.6 Environment Setup

The experiments are all run on the Google Colab platform with T4 GPUs. The code can be found in this GitHub repository[1].

4.7 Results

4.7.1 Can training with a mixture of real and synthetic data improve Show and Tell's performance?

The BLEU scores of Show to Tell trained with different ratios of real and synthetic images on the Flickr8k testing set are presented in Table 2

Ratio of Real Images	Ratio of Synthetic Images	BLEU Score
1.00	0.00	0.1054
1.00	0.50	0.1035
1.00	1.00	0.1032
0.50	1.00	0.0941
0.00	1.00	0.0893

Table 2: Testing Show and Tell Trained on the Flickr8k Training Set on the Flickr8k Testing Set

Increasing the percentage of synthetic images in the training set leads to lower BLEU scores. This shows that contrary to the results obtained for training image classifiers, training Show and Tell with a mixture of real and synthetic data does not improve its performance. On the contrary, it leads to performance degradation compared with training with real data only.

4.7.2 Can training with a mixture of real and synthetic data improve Show and Tell's robustness?

The BLEU scores of Show to Tell trained with different ratios of real and synthetic images on other datasets are presented in Table 3.

Increasing the percentage of synthetic images in the training set leads to lower BLEU scores on all other datasets. This shows that contrary to the results obtained for training image classifiers, training

[1] `https://github.com/abbaswu/The-Impact-of-Synthetic-Data-on-Image-Captioning-Models`

Dataset	Ratio of Real Images	Ratio of Synthetic Images	BLEU Score
Pascal_VOC_2008	1.00	0.00	0.0618
Pascal_VOC_2008	1.00	0.50	0.0593
Pascal_VOC_2008	1.00	1.00	0.0618
Pascal_VOC_2008	0.50	1.00	0.0593
Pascal_VOC_2008	0.00	1.00	0.0511
Flickr30k	1.00	0.00	0.1652
Flickr30k	1.00	0.50	0.1568
Flickr30k	1.00	1.00	0.1541
Flickr30k	0.50	1.00	0.1350
Flickr30k	0.00	1.00	0.0811
COCO2014	1.00	0.00	0.0621
COCO2014	1.00	0.50	0.0610
COCO2014	1.00	1.00	0.0607
COCO2014	0.50	1.00	0.0598
COCO2014	0.00	1.00	0.0579

Table 3: Testing Show and Tell Trained on the Flickr8k Training Set on Other Datasets

Show and Tell with a mixture of real and synthetic data does not improve its robustness. On the contrary, it leads to performance degradation compared with training with real data only.

4.7.3 Can pre-training with synthetic data and fine-tuning with real data improve Show and Tell's performance?

The BLEU scores of Show to Tell pre-trained on synthetic data on the Flickr8k testing set are presented in Table 4

Number of Fine-tuning Epochs	BLEU Score
5	0.1086
10	0.1072
20	0.1055
40	0.1064

Table 4: Testing Show and Tell pre-trained on Synthetic Data on the Flickr8k Testing Set

It can be observed that Show and Tell achieved a higher BLEU score when trained on synthetic data then fine-tuned on the Flickr8k training set than when directly trained on the Flickr8k training set only, as depicted in the first row of Table 2. Furthermore, with the increase in the number of fine-tuning epochs, the BLEU score gradually approaches that of training the model on the Flickr8k training set only. This suggests that pretraining and fine-tuning may lead to Show and Tell performing better than directly training on the training set.

4.7.4 Can pre-training with synthetic data and fine-tuning with real data improve Show and Tell's robustness?

The BLEU scores of Show to Tell trained on synthetic data on other datasets are presented in Table 5

Compared with Table 3, we can observe that Show and Tell only achieves a higher BLEU score when pre-trained on synthetic data then fine-tuned on the Flickr8k training set for the COCO2014 dataset, but not for the Pascal_VOC_2008 and the Flickr30k dataset. This suggests that pretraining and fine-tuning do not increase the robustness of Show and Tell.

Dataset	Number of Fine-tuning Epochs	BLEU Score
Pascal_VOC_2008	5	0.0605
Pascal_VOC_2008	10	0.0610
Pascal_VOC_2008	20	0.0606
Pascal_VOC_2008	40	0.0592
Flickr30k	5	0.1034
Flickr30k	10	0.1152
Flickr30k	20	0.1350
Flickr30k	40	0.1547
COCO2014	5	0.0642
COCO2014	10	0.0628
COCO2014	20	0.0612
COCO2014	40	0.0622

Table 5: Testing Show and Tell Trained on the Flickr8k Training Set on Other Datasets

5 Conclusion

In conclusion, the exploration into the impact of synthetic data on image captioning models reveals a subtle relation between performance, robustness and the amount of synthetic data. The experimental findings, conducted with the Show and Tell model, suggest that training with a *mixture* of real and synthetic data does not improve performance and robustness, as in the case of training image classification models Bansal and Grover [2023], but underscore the potential of using AI-generated images in *pretraining* datasets, with models trained solely on synthetic images and fine-tuned on real data outperforming those relying solely on real data.

This study contributes to the evolving landscape of the application of synthetic data in image-based deep learning tasks. Future work involves a deeper understanding of underlying factors influencing the experimental results, alternative conditioning methods, and the broader implications for other image-based tasks, as this paper anticipates the growing viability of this strategy with advancements in synthetic data quality and cost-effectiveness.

References

Ahmet Aker and Robert Gaizauskas. Generating image descriptions using dependency relational patterns. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 1250–1258, 2010.

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

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

Hritik Bansal and Aditya Grover. Leaving reality to imagination: Robust classification via generated datasets. *arXiv preprint arXiv:2302.02503*, 2023.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023.

Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. Wavegrad: Estimating gradients for waveform generation. *arXiv preprint arXiv:2009.00713*, 2020.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10578–10587, 2020.

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*, pages 15–29. Springer, 2010.

Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems*, 26, 2013.

Ankush Gupta, Yashaswi Verma, and C Jawahar. Choosing linguistics over vision to describe images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 26, pages 606–612, 2012.

Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? *arXiv preprint arXiv:2210.07574*, 2022.

Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image captioning: Transforming objects into words. *Advances in neural information processing systems*, 32, 2019.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022a.

Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *The Journal of Machine Learning Research*, 23(1):2249–2281, 2022b.

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4634–4643, 2019.

Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. *Advances in neural information processing systems*, 27, 2014.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020.

Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. Babytalk: Understanding and generating simple image descriptions. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2891–2903, 2013.

Polina Kuznetsova, Vicente Ordonez, Tamara L Berg, and Yejin Choi. Treetalk: Composition and compression of trees for image descriptions. *Transactions of the Association for Computational Linguistics*, 2:351–362, 2014.

Siming Li, Girish Kulkarni, Tamara Berg, Alexander Berg, and Yejin Choi. Composing simple image descriptions using web-scale n-grams. In *Proceedings of the fifteenth conference on computational natural language learning*, pages 220–228, 2011.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

Fenglin Liu, Xuancheng Ren, Xian Wu, Shen Ge, Wei Fan, Yuexian Zou, and Xu Sun. Prophet attention: Predicting attention with future attention. *Advances in Neural Information Processing Systems*, 33:1865–1876, 2020.

Margaret Mitchell, Jesse Dodge, Amit Goyal, Kota Yamaguchi, Karl Stratos, Xufeng Han, Alyssa Mensch, Alexander Berg, Tamara Berg, and Hal Daumé III. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 747–756, 2012.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011.

Jia-Yu Pan, Hyung-Jeong Yang, Pinar Duygulu, and Christos Faloutsos. Automatic image captioning. In *2004 IEEE International Conference on Multimedia and Expo (ICME)(IEEE Cat. No. 04TH8763)*, volume 3, pages 1987–1990. IEEE, 2004.

Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-linear attention networks for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10971–10980, 2020.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk*, pages 139–147, 2010.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024, 2017.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022a.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022b.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022c.

Zhan Shi, Xu Zhou, Xipeng Qiu, and Xiaodan Zhu. Improving image captioning with better use of captions. *arXiv preprint arXiv:2006.11807*, 2020.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. From show to tell: A survey on deep learning-based image captioning. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):539–559, 2022.

Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*, 2022.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models, 2022.

Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J Fleet, Radu Soricut, et al. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18359–18369, 2023.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10685–10694, 2019.

Yezhou Yang, Ching Teo, Hal Daumé III, and Yiannis Aloimonos. Corpus-guided sentence generation of natural images. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 444–454, 2011.

Benjamin Z Yao, Xiong Yang, Liang Lin, Mun Wai Lee, and Song-Chun Zhu. I2t: Image parsing to text description. *Proceedings of the IEEE*, 98(8):1485–1508, 2010.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588, 2021.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13041–13049, 2020.

Xinxin Zhu, Weining Wang, Longteng Guo, and Jing Liu. Autocaption: Image captioning with neural architecture search. *arXiv preprint arXiv:2012.09742*, 2020.