Novel Data Augmentation for resource constrained Image captioning

Stanford CS224N Custom Project

Anirudh Sriram
Department of Electrical Engineering
Stanford University
sanirudh@stanford.edu

Parth Dodhia
Department of Electrical Engineering
Stanford University
pdodhia@stanford.edu

Abstract

Image captioning is an important task for improving human-computer interaction as well as for a deeper understanding of the mechanisms underlying the image description by humans. In recent years, this field of research has rapidly developed and a number of impressive models have come out. However, even the best models have a limit in the quality of the captions generated with a lack of training data. In this work, we come up with a novel data augmentation technique using text-to-text and text-to-image generative models to create good-quality augmented datasets for developing robust image captioning models in data-constrained settings. Our preliminary findings reveal that our augmentation technique allows the model to perform similar to models trained on about 3 to 5 times of original data we train on. We analyze the individual text augmentation techniques used (like synonym replacement or word deletion) and identified that synonyms and embedding word swap methods helps the image captioning model the most to generate better captions. We demonstrate the benefits of using augmentation techniques on both modalities (text and image) instead of focusing on text-only or image-only augmentations. We also show the improvement in model predicting capabilities in using an text-to-image generative model to induce new visual features instead of opting for the commonly used image augmentation techniques like random rotation, random zoom, etc which makes use of existing visuals only.

1 Key Information to include

- Mentor: Rishi Desai (CA)
- External Collaborators (if you have any): No
- Sharing project: No

2 Introduction

Over the past few years, there have been massive strides in the field of natural language processing and a major contributor for that has been the transformer (Vaswani et al., 2017) model. Large language models (GPT models in general (Radford et al., 2019a)) has been the talk of the town and almost all state-of-the-art image captioning models are transformer based, but what we fail to understand is the amount of data that is required to train these models. Obtaining quality training data is still a hassle for many domain specific tasks and there is a pressing need to develop models which are capable of performing well even in data scarce scenarios.

In this work, we come up with a simple and efficient data augmentation technique using text-to-text and text-to-image generative models to improve the quality of predicted captions. Unlike previous works like Atliha and eok (2020), we use five diverse text augmentation techniques like synonym replacement, word deletion, embedding swap, etc. which improves the quality of augmented data and helps with improvement in prediction capabilities. Through basic augmentation techniques like synonym replacement, we start to move more closer towards how a human infers the image captioning task. For a given image, different individuals might come up with different captions, which mean the same but has different set of words describing them and this is exactly what we want to induce into the model.

We first demonstrate the loss in performance of the trained model with decrease in number of samples available for training. We then illustrate the utility of our proposed technique by comparing the three baseline-model's (defined in section 4) performance with and without augmentation in a data scare setting. We also thoroughly investigate the use of each individual text augmentation type and determine which technique contributes the most towards the improvement in performance. We also demonstrate of the benefit of having multi-modal augmentations instead of commonly used text-only or image-only augmentation methods.

3 Related Work

Image captioning has been a task that has received a lot of interest in the recent years primarily because it is at the intersection of both computer vision and natural language processing. There has been a lot of innovative solutions that has come up up in the recent times but LSTM and Bi-LSTM (Wang et al., 2016) based solutions have been the standard basic approach for this task. In recent years works using Transformer models such as Kumar et al. (2022) and Li et al. (2023) have taken over as the state-of-the-art approaches. Even though the models provide significant improvement over existing baselines, the results were obtained under conditions with large and sufficient amount of training data. In order tune these models to be able to perform well under data scarce scenarios works such as Atliha and Šešok (2020) and Li et al. (2021) have tried to different innovative methods such as synonym based word replacement and retrieval methods. Most of these works have restricted themselves only to synonyms based augmentation techniques and single modality (only text).

Generating transformed text for a given input caption has been used a lot in recent times to develop robust NLP models that are able to withstand adversarial attacks. Textattack (Morris et al., 2020) is a commonly used framework that is used for generating the augmented outputs. We also make use of this framework to generate our transformed captions. Apart from the augmentation techniques that we are using (listed in Section 4 other commonly used augmentation techniques include PWWS (Ren et al., 2019), TextFooler (Jin et al., 2019) methods.

For the image augmentation part, we make use of an generative model that is capable of generating new images for given input prompts. In recent years, the field of image generation has seen so many innovative solutions like Mini-Dalle (Dayma et al., 2021), Swin-Imagen (Li et al., 2022) and Stable diffusion (Rombach et al., 2021). In this work we make use of the stable diffusion model.

4 Approach

The pipeline of our proposed method contains of a Text augmentation module, Image augmentation module and a Image captioning model. Our aim is to take a data constrained setting, make use of these individual augmentation technique to procure more data and train an image captioning model on top that data.

4.1 Text Augmentation module: Generate augmented caption prompts

For a given input caption, we create five different modified output captions based on the following word-level transformations: (i) **Word Deletion** - Transforms an input by removing a randomly selected word (Feng et al., 2018), (ii) **Synonym replacement** - Transforms an input by replacing its words with synonyms provided by WordNet (iii) **Embedding swap** - Transforms an input by replacing its words with similar words from embedding space (Mrksic et al., 2016), (iv) **Inner swap random** - Transforms an input by reordering the words in the sentence (Pruthi et al., 2019), (v) **BERT-Masked LM swap** - Generate potential replacements for a word using a masked language model.

We use the popular TextAttack framework (Morris et al., 2020) for implementations of these attacks. TextAttack is a python framework for adversarial attacks, adversarial training, and data augmentation in NLP. For any input sentence, we have restricted our model to only modify one word. The portion of the input sentence to be modified is a hyperparameter which in practice is set based on the length of each input. While creating the augmented outputs, we impose constraints to (i) avoid making transforms on stopwords, (ii) avoid modifying the same word in case of composite transforms and (iii) we make use of GPT-2 (Radford et al., 2019b) model to ensure that difference in log-probability of input and transformed sentence is less than 2.0 (the ensure plausibility and grammaticality of generated text). The transformation model takes the sentence to transform, constraints, transformation type as input and outputs a transformed sentence.

Transformation Type	**Modified Text**
Original Text	**a little girl climbing into a wooden playhouse.**
Word Deletion	a little girl climbing into a _ playhouse
WordNet Synonym swap	a minuscule girl climbing into a wooden playhouse
Embeddings swap	a meagre girl climbing into a wooden playhouse.
Inner swap random	a little girl wooden into a climbing playhouse
BERT-Masked LM swap	a drunk girl climbing into a wooden playhouse

Table 1: Examples of the different transformations used for data creation.

4.2 Image Augmentation module: Generate images from input caption prompts

For a given input caption, we generate an output image making use of pre-trained off-the-shelf **Stable diffusion** (Rombach et al., 2022) model. It is a Latent Diffusion Model that uses a fixed, pretrained text encoder (CLIP ViT-L/14) based on the Imagen paper (Saharia et al., 2022).

We present an example of generated images for the input prompt : **A black dog is playing in the snow**.

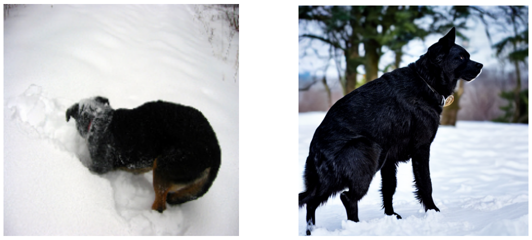


Figure 1: (a) Original Image from Flickr8K dataset, (b) Image generated by Stable diffusion model

The components of a stable diffusion model include a text encoder, image information creator and image decoder (the last two together can be referred as an image generator).

Text encoder: This is the text-understanding component that translates the text information into a numeric representation that captures the ideas in the input prompt. The stable diffusion model makes use of a special Transformer language model - CLIP (Radford et al., 2021). It takes the input text and outputs a list of numbers representing each word/token in the text (a vector per token). That information is then presented to the Image Generator.

Image information creator: This component is the secret sauce of Stable Diffusion. The image information creator works completely in the image information space (or latent space). This property makes it faster than previous diffusion models that worked in pixel space. This component is made up of a UNet neural network (Ronneberger et al., 2015) and a scheduling algorithm. The word "diffusion" describes what happens in this component. It is the step by step processing of information that leads to a high-quality image being generated in the end.

Image Decoder: The image decoder paints a picture from the information it got from the information creator. It runs only once at the end of the process to produce the final pixel image.

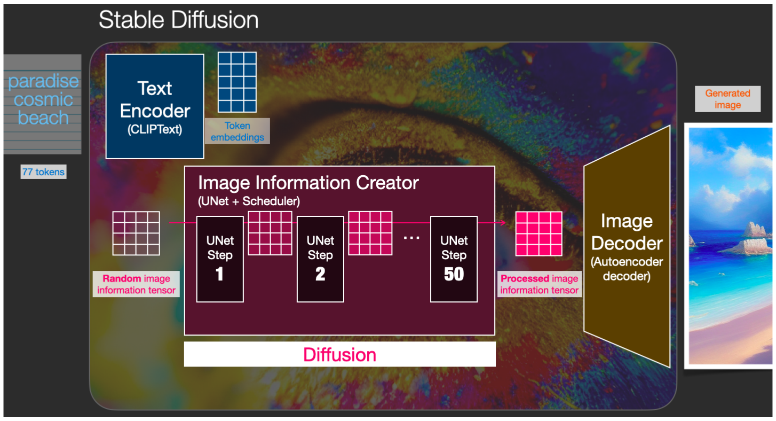


Figure 2: Steps in a stable diffusion model (with sample prompt and generated image) Source: (sta)

4.3 Image captioning model: Generates the caption for given visual input

Our primary research in this work is to understand the capabilities of the augmentation technique proposed. Hence we build three different custom image captioning models and test the effect of using the proposed data augmentation.

- **CNN- LSTM based Image captioning model (CL):** We make use of a pretrained EfficientNet (Tan and Le, 2021) architecture for image feature extraction. The LSTM layers are used for generating the output text sequence give the input vector from the image feature extractor. The input image is converted into a 2048 dimensional feature vector and then passed into a linear layer to bring the size down to the size of the word embedding (512 dim). This passed as input to the LSTM decoder and seeds the decoding process where we make use of a greedy decoder to predict the best token at each time step.
- **CNN - Bi-LSTM based Image captioning model (CBL):** This model is identical to the CNN-LSTM model except for the fact that this uses a Bi-LSTM layer instead of LSTM layer in the decoder. In general, studies have shown using a Bi-LSTM layer in a decoder is not as useful as using it in an encoder instead of a LSTM layer and we observe the same in our experiments as well (which we present in next section).
- **Transformer based Image captioning model (Tr):** We make use of a transformer based encoder-decoder architecture which takes the output of the EfficientNet model as input to the encoder model. We set the number of encoder and decoders to 3 and the embedding size used remains 512. As expected the transformer model (with attention capabilities) outperforms the vanilla LSTM based models.

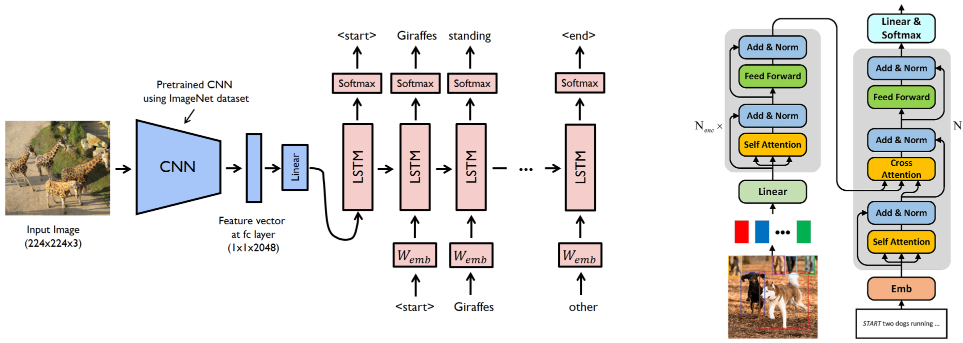


Figure 3: (a) Architecture of our CNN-LSTM model (b) Architecture of a transformer based image captioning model (Kumar et al., 2022)

5 Experiments

5.1 Data

All experiments are performed on the Flickr8K dataset (fli). Flickr 8K dataset consists of 8,091 images that are each paired with five different captions which provide clear descriptions of the salient entities and events in the picture. The images were chosen from six different Flickr groups and were manually selected to depict a variety of scenes and situations.

5.2 Evaluation metrics

We will evaluate the model performance on standard metrics for image captioning - BLEU score (Papineni et al., 2002), METEOR score (Banerjee and Lavie, 2005). BLEU score (0-100, larger is better) is a metric measuring the overlap of n-grams upto size 4 of the generated captions with the reference ground truths. The METEOR score (0-100, larger is better) is based on a harmonic mean of precision and recall calculated by mapping unigrams from the output to the ground truth.

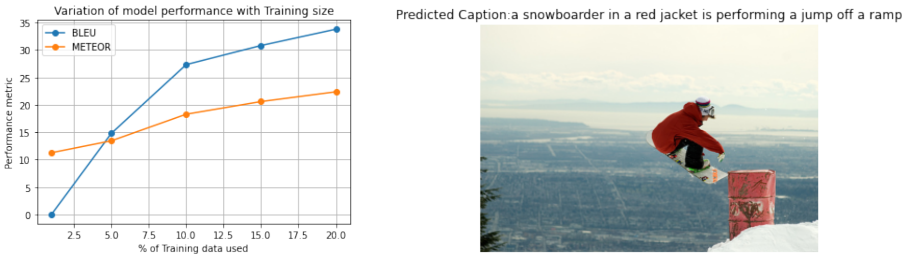
5.3 Experimental details and performance evaluation

The Flickr dataset is split into 6473 train images and 1618 test images (each with 5 captions). In total we have 32365 train samples and 8090 test samples. We train and report the performance metrics (in Table 2) for each of the baseline model when different amounts of training data is available. (Note: x% of training data implies the baseline models are trained on x% of total available training samples instead of the whole training set - For example 10% of training data would correspond on training each model with 3236 samples). The models make use of Adam optimizer with custom learning rate schedule. We also use Early stopping (patience = 5), Model checkpoints and Learning rate warm up to avoid overfitting. All the models are trained on a Tesla K80 GPU. The run time (for the entire training data) is less than two hours for all the baseline models.

% of data	B-CL	B-CBL	B-Tr	M-CL	M-CBL	M-Tr
100	26.46	29.71	44.13	18.82	23.74	32.20
50	21.70	24.15	38.11	15.19	19.78	27.85
20	16.39	19.03	33.79	14.00	16.91	22.38
10	10.57	12.82	27.34	11.31	13.59	18.29
5	4.08	6.49	14.80	10.74	11.52	13.44
1	0.01	0.01	0.01	9.41	10.65	11.27

Table 2: Results on various portions training sets. **Legend:** B-CL: Bleu for CNN-LSTM, B-CBL: Bleu for CNN-BiLSTM, B-Tr: Bleu for Transformer, M-CL: Meteor for CNN-LSTM, M-CBL: Meteor score for CNN-BiLSTM, M-Tr: Meteor for Transformer model

From table 2 we can observe that the transformer model outperforms others. We also see the expected decrease in Bleu and meteor scores as the number of samples the models are trained on decreases. We observe that as size of training data decreases, the difference in performance on state-of-the-art models like transformers and the vanilla LSTM models decreases aligning with our hypothesis in the beginning that even the state of the art models don't do much better in data scarce scenarios.


Variation of model performance with Training size

Predicted Caption:a snowboarder in a red jacket is performing a jump off a ramp

Figure 4: (a) Effect of training size on performance for transformer model (b) Sample prediction from Transformer model trained on 50% training data

We then dig deeper into the transformer models and observe the model prediction capabilities for this models in data scarce scenarios (less than 20% of training data available - Fig:4(a)). From the plot we can observe that when we go below 10% of actual training data size, the image captioning model's performance starts taking a serious hit. We now make use of our proposed data augmentation technique on these subsets of data to see the improvement in model performance.

% of data	B-CL	B-CBL	B-Tr	M-CL	M-CBL	M-Tr
20	20.30(+23.86%)	23.08(+21.28%)	36.47(+7.93%)	19.94(+42.43%)	21.76(+28.68%)	28.58(+27.70%)
10	16.49(+56.01%)	19.26(+50.23%)	33.21(+21.47%)	16.31(+44.21%)	17.38(+27.89%)	23.35(+27.67%)
5	14.25(+249.26%)	17.85(+175.04%)	30.10(+103.38%)	13.59(+26.54%)	15.45(+34.11%)	20.72(+54.17%)
1	7.02(x702)	10.48(x1048)	15.84(x1584)	10.51(+11.69%)	11.10(+4.23%)	13.29(+17.92%)

Table 3: Results with Augmentation **Legend:** B-CL: Bleu for CNN-LSTM, B-CBL: Bleu for CNN-BiLSTM, B-Tr: Bleu for Transformer,M-CL: Meteor for CNN-LSTM, M-CBL: Meteor score for CNN-BiLSTM, M-Tr: Meteor for Transformer model

From Table 3 we can observe the significant improvement in both metrics when our proposed augmentation technique is used. We can observe that the improvement is much more significant when the scarcer the available data is (lower the amount of available training data, more useful is our method). On average the model's performance after augmentation resembles models trained on almost three to five times more data samples.

5.4 Significance of augmenting using both modalities

In our proposed technique, we augment the data with both generated text and image samples. As explained in Section 4, we make use of the Textattack framework for text augmentation and the Stable diffusion model for image augmentation. The question here is, "does it really help to augment using both text and image, instead of text-only or image-only". To evaluate this, we take a data scarce setting where the model is trained on 10% of original training data and compare the performance of the captioning model under different augmentation settings.

Augmentation method	Captioning Model	BLEU score	METEOR score
No augmentation	CNN-LSTM	10.57	11.31
No augmentation	Transformer	27.34	18.29
Text only	CNN-LSTM	14.72	14.18
Text only	Transformer	30.91	21.84
Image only	CNN-LSTM	13.40	12.71
Image only	Transformer	29.62	20.66
Text and Image	CNN-LSTM	16.49	16.31
Text and Image	Transformer	33.21	23.35

Table 4: Performance of Transformer and CNN-LSTM Captioning model under data scarce setting using different augmentation scenarios

From Table 4 we can observe that both text only and image-only augmentations improves model performance, with text augmentation resulting in a larger increase. We also observe that using both modalities together gives that largest jump in captioning performance aiding us to conclude that using both modalities for augmentation helps the model extract and utilize more useful information and perform better.

5.5 Analyzing the text transformation techniques

We now move onto analyze the importance of each text transformation technique used in helping the model predict robust captions. We take a data scarce setting (with 20% of training data used) and augment the model only using a single transformation technique at a time and test each model on the same test dataset (in Fig:5(a)). From Figure 5(a), we observe that using Synonyms and Embedding based word replacement aids the model the most to come up with better captions. We also pointed earlier that these two transformations are vital in helping us understand how humans caption images, thus by using these transformations we aid the model to learn that feature and make better predictions on the test data.

As expected the deletion transformation helps the least, this can be attributed to the fact that the sentence formed after deletion may not always be grammatically correct and we try to understand more about this with another experiment. Now we take a model and train it on a 20% subset from original training data (no augmentation). Instead of testing on the usual test data, we now test on transformed test data of one transformation type (for example all test samples are now transformed with Synonym based word replacement). We observe in Figure 5(b) that the model trained on non augmented captions is able to perform well in well-aligned transformation's test set like synonym and embedding swap, but fail to do well on deletion and Inner swap transformations. Hence in a constrained resource setting it is beneficial to use more of synonym, embedding and MaskLM based transformations but deletion and Inner swap like transformations adds diversity to the captions and make them more robust.

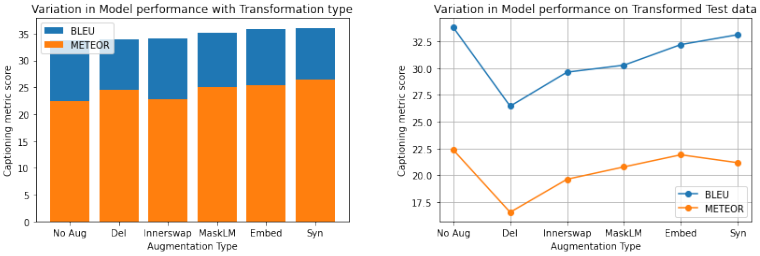


Figure 5: (a) Model performance across different augmentation types (b) Model performance across different augmentation type on transformed test data

5.6 Analyzing Image augmentation technique

Using generative models such Stable diffusion is resource and time consuming in general. There are several commonly used image augmentation techniques like random rotation, random cropping etc. We analyze why using an generative model to develop completely new visual information helps the model to perform better.

We take 20% of original training data. We now develop two training sets from this, one based on conventional image augmentation technique making use of Random rotation, random zoom and random contrast and the other based on our stable diffusion model. We finally train both the data using the transformer model and test on our standard test set consisting of 8090 test samples. The results are added in Table 5.

Augmentation method	Captioning Model	BLEU score	METEOR score
No augmentation	Transformer	33.79	22.38
Conventional Data augmentation	Transformer	33.86	23.69
Stable diffusion based	Transformer	35.82	26.57

Table 5: Comparison of different Image augmentation techniques

From Table 5 we can observe that there is little to no change in using the conventional image augmentation techniques in this case. This can be due to the fact that the representations learnt by most of these sophisticated image feature extractors like EfficientNET are almost invariant to these transformation (like rotation, zoom etc.). In case of the generative model, we create completely new visual features which the captioning model can utilize. Hence the stable diffusion based augmentation technique is preferred over the other standard image augmentation techniques.

6 Analysis

Our results demonstrate the utility of using our proposed augmentation technique in data scarce scenarios. While transformer based models maybe the state-of-the-art, we could observe that they are able to push ahead of vanilla LSTM like models only in presence of sufficient amount of data. Through our experiments we were also able to analyze the importance of each text augmentation technique used, but the analysis was restricted to transformation put in use in this work. We could expand our transformation type to generate more diverse captions. The proposed technique seems to work much better in cases with extreme data scarcity (like 1% data availability), whereas the improvement is not as significant in cases with larger data availability (like 10% and 20% data availability). Beyond a certain threshold, adding more and more data is not solution to improve model performance. The image generation process can be time consuming compared to the conventional image augmentation techniques, we need to decide if the performance gain that we get from the image generation model is worth the time we spend in generating the new images. We also observed that synonyms and embedding based word replacement transformations seem to help the model the most aligned with our inference of how humans do image captioning. Overall with this augmentation technique, most models are able to mimic captioning models trained on 3 to 5 times more training data.

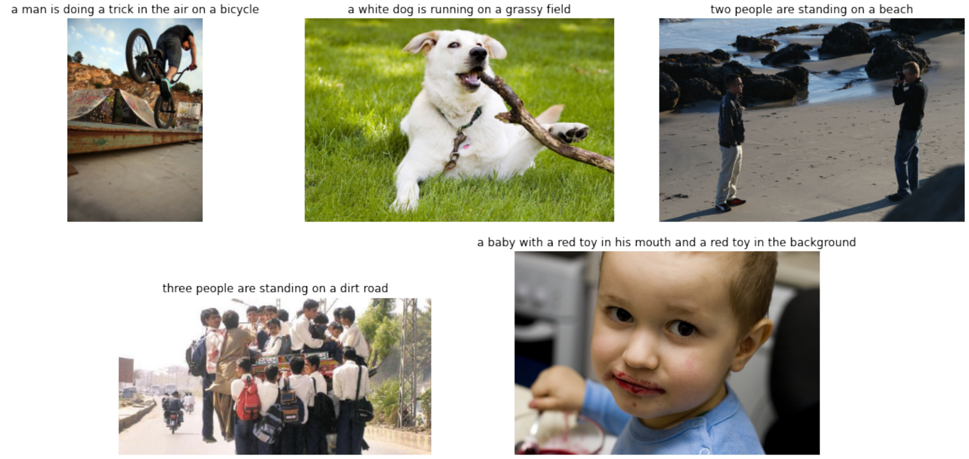

a man is doing a trick in the air on a bicycle
a white dog is running on a grassy field
two people are standing on a beach
three people are standing on a dirt road
a baby with a red toy in his mouth and a red toy in the background

Figure 6: Sample predictions from model trained only on Synonym, embedding, MaskLM, Inner Swap and Deletion augmentation (in order)

7 Conclusion

Our data augmentation technique helps improving the model performance significantly in data scarce scenarios. We were able to replicate results similar to models trained with three times the amount of original samples we used for training. We also demonstrated the benefit of using both text-to-image and text-to-text augmentation techniques instead of restricting to single modality. We also compared several individual text transformation technique and their utilities. Similarly for the image augmentation case, we could show that our generative model outperformed conventional image augmentation techniques. Future work could benefit from doing a thorough analysis with domain specific datasets like medical report generation where the data is actually scarce. We could also extend the type and number of text transformations that we use (include transformations like PWWS, TextFooler). We could also adapt the use of other captioning metrics like CIDEr and ROUGE.

References

Flickr8k. `https://www.kaggle.com/datasets/adityajn105/flickr8k`.

Stable Diffusion. `https://jalammar.github.io/illustrated-stable-diffusion/`.

Viktar Atliha and Dmitrij eok. 2020. Text augmentation using bert for image captioning. *Applied Sciences*.

Viktar Atliha and Dmitrij Šešok. 2020. Text augmentation using bert for image captioning. *Applied Sciences*, 10(17).

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic* *and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Boris Dayma, Suraj Patil, Pedro Cuenca, Khalid Saifullah, Tanishq Abraham, Phúc Lê Khc, Luke Melas, and Ritobrata Ghosh. 2021. Dall·e mini.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728, Brussels, Belgium. Association for Computational Linguistics.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2019. Is BERT really robust? natural language attack on text classification and entailment. *CoRR*, abs/1907.11932.

Deepika Kumar, Varun Srivastava, Daniela Popescu, and Jude Hemanth. 2022. Dual-modal transformer with enhanced inter- and intra-modality interactions for image captioning. *Applied Sciences*, 12:6733.

Guodun Li, Yuchen Zhai, Zehao Lin, and Yin Zhang. 2021. Similar scenes arouse similar emotions. In *Proceedings of the 29th ACM International Conference on Multimedia*. ACM.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models.

Ruijun Li, Weihua Li, Yi Yang, Hanyu Wei, Jianhua Jiang, and Quan Bai. 2022. Swinv2-imagen: Hierarchical vision transformer diffusion models for text-to-image generation.

John X. Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp.

Nikola Mrksic, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gasic, Lina Maria Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve J. Young. 2016. Counter-fitting word vectors to linguistic constraints. *CoRR*, abs/1603.00892.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Danish Pruthi, Bhuwan Dhingra, and Zachary C. Lipton. 2019. Combating adversarial misspellings with robust word recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5582–5591, Florence, Italy. Association for Computational Linguistics.

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. 2021. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019a. Language models are unsupervised multitask learners.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019b. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1085–1097, Florence, Italy. Association for Computational Linguistics.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-resolution image synthesis with latent diffusion models. *CoRR*, abs/2112.10752.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022. Photorealistic text-to-image diffusion models with deep language understanding.

Mingxing Tan and Quoc V. Le. 2021. Efficientnetv2: Smaller models and faster training.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Cheng Wang, Haojin Yang, Christian Bartz, and Christoph Meinel. 2016. Image captioning with deep bidirectional lstms. *CoRR*, abs/1604.00790.