Exposé - Study Project

Denoising Diffusion Probabilistic Models as image encoders for vision-language tasks in remote sensing

Mark Spitzner
spitznem@informatik.hu-berlin.de
Humboldt University Berlin, Germany

Supervisors:
Prof. Dr. Pedram Ghamisi
p.ghamisi@gmail.com
Helmholtz-Zentrum Dresden-Rossendorf, Germany
Institute of Advanced Research in Artificial Intelligence, Austria

Prof. Dr. Ulf Leser
leser@informatik.hu-berlin.de
Humboldt University Berlin, Germany

November 2, 2023

Contents

1 Introduction

Recent advancements in deep learning, particularly within the fields of natural language processing (NLP) and computer vision (CV), coupled with the emergence of multi-modal approaches, have initiated significant research efforts in the area of vision-language (VL) tasks. These tasks, e.g. image captioning (given an image, describe the content in natural language), change captioning (given images of the same region at different times, describe the changes if any occurred) and visual question answering (VQA) (given an image and a question in natural language, answer the question in natural language in the context of the image content), hold substantial academic significance due to their capability to express a higher degree of domain comprehension compared to relatively simpler tasks, such as image classification, which can be reformulated and solved by a vision-language model (VLM).

Furthermore, vision-language tasks enable the interaction of non-experts with data through intuitive natural language interfaces. Especially in the domain of remote sensing and post-disaster analysis this can be crucial since the ability to evaluate satellite or drone imagery with respect to the needs of first responders e.g. finding a traversable road after a flood event or describing the changes in a region influenced by wild-fire or landslide scenarios.

However, training a model that performs well in this domain is hard due to the scarcity of labeled high-quality data. One major explanation is that the vast amounts of unlabeled data would require significant amounts of manual efforts in order to be useful in a supervised learning task. This emphasizes the development of label efficient approaches that produce meaningful representations based on the visual input and by that enabling language decoders to produce accurate solutions for the downstream task. One way to address this challenge is by the means of generative models that work on the image domain. These models can learn strong representations without labeled data e.g. generative adversarial networks[1] (GANs) or denoising diffusion probabilistic models (DDPMs)[2]. Especially the latter saw recent breakthroughs and where able to show high-quality generated images besides also providing high scalabilty and parallelizability.

The goal of the associated study project is to provide a overview of the necessary theoretical background and the current state of the art as well as evaluate the capabilities of these DDPMs to provide meaningful representations for VLMs applied to a small-scaled proof of concept.

2 Related work

As gathering a comprehensive understanding of the current state of the art will be a major part of the pending study project, we will only give a brief overview considering the most relevant approaches for the study project in this exposé.

In recent developments within the field of vision-language tasks, there has been a notable shift from highly specialized expert models to bigger, generalistic foundation models predominantly built on transformer-based[3] encoder-decoder architectures. This evolution is caused by the increased availability of vast datasets, enhanced computing resources, and the advancements in semi-supervised or self-supervised training techniques. Consequently, this paradigm shift has led to the creation of model architectures containing billions of parameters, exemplified by GIT[4] and Flamingo[5], which have significantly advanced the state-of-the-art in tasks such as image classification, image captioning, and visual question answering.

In contrast to these monolithic architectures, there is a growing trend of leveraging specialized pretrained expert models that possess domain-specific knowledge. One of the most recent exemplars is the Prismer[6] architecture. The integration of pretrained expert models empowers the architecture to require significantly less training data and resources for domain adaptation.

However, a common limitation of these models is their training on conventional RGB images. This may not always align with the requirements of remote sensing environments, where factors such as distribution, perspective, spatial resolution, and spectral attributes can significantly differ. Furthermore, remote sensing tasks often involve multi- or hyper-spectral data, which adds complexity to the problem. To address these challenges, various approaches have been proposed. For instance, Lobry et al.[7] formulated the VQA task tailored to remote sensing and introduced a benchmarking dataset. Their proposed network architecture incorporates a ResNet-152[8] for visual feature extraction and a skip-thoughts RNN[9] to extract relevant information from questions, aggregated by a learnable feature fusion module and fed into a prediction head for answering questions. Building upon this foundation, Siebert et al.[10] enhanced the approach by incorporating VisualBERT to fuse visual and textual features. Progress in VL tasks extends beyond VQA and also includes innovations in change captioning. For example, Chang and Ghamisi [11] introduced a self-attentive encoder for fusing visual features extracted from bi-temporal satellite images, propagating the resulting representations to a transformer decoder responsible for caption generation.

In addition to advancements in vision-language tasks, recent developments in Denoising Diffusion Probabilistic Models (DDPMs) have demonstrated their ability to generate high-quality and realistic synthetic visual data[2]. The fundamental concept underlying DDPMs involves learning to predict noise iteratively added to an input image. Upon completing the training process, noise sampling yields high-fidelity synthetic outputs. This indicates a profound understanding of the underlying domain which we aim to exploit to provide beneficial representations for current VL challenges. Up until now, denoising diffusion models have been used for various other task besides general image generation such as super-resolution[12], music synthesis[13] and even change detection in remote sensing images [14] or cloud removal in sentinal-2 images[15]. Especially, the latter show promising performance on remote sensing data which provides additional motivation for this study project.

3 Goal and Approach of the study project

3.1 Goal

The focus of the study project will be to evaluate the performance of DDPM models as encoders for VLMs in the domain of remote sensing datasets. Therefore, the study project will first provide an insight into the state of the art literature for VLMs, DDPMs and foundation models in general. It will also provide an overview of the theoretical background, the used datasets and used evaluation metrics used for common VL tasks. Furthermore, a DDPM will be implemented and trained on the FloodNet[16] VQA dataset as well as the LEVIR-CC[17] dataset. Finally, a RoBERTa[18] model will be used as a language decoder. We will apply this decoder as a frozen model as well as fine-tuned on the respective datasets. The performance of the resulting architecture will be evaluated using the corresponding metrics provided in the original papers to provide comparability. To furthermore provide a sense for the performance of the underlying encoder we will also replace the DDPM with a fine-tuned ViT[19] evaluate and compare the performance of both systems.

3.2 Approach

During the study project we will integrate a DDPM as encoder for visual information for VL tasks. An RoBERTa-Model will be used as encoder for natural language inputs and as decoder in order to provide natural language output. The base idea is to train a DDPM on the target dataset in order to learn meaningful representations of the underlying distribution and to extract key features which would be characteristic during the sampling of similar images.

Usually, DDPMs consist of a Diffusion process which gradually adds noise to an input image. The obstructed input image is then fed through an U-Net which learns to predict the added noise. After infering the noise it is than subtracted from the obstructed image and the reconstruction error is measured in order to provide a learning signal to the model.

During inference and fine-tuning of the full VLM, images are diffused and reconstructed. During this process, the hidden states of the diffusing U-Net are aggregated and fused in order to create a global representation that can be fed into the RoBERTa-Model e.g. by the means of cross-attention or early fusion. The major challenges that need to be solved during the study project are the design of the aggregation and fusion operators and the integration with an existing large language model (LLM).

The architecture is depicted in figure 1. It displays the basic architecture as well as information flow inside the model. We omitted the detailed visualization of the RoBERTa-Model as it is an Transformer encoder as described in the original transformer paper trained in a specific way.

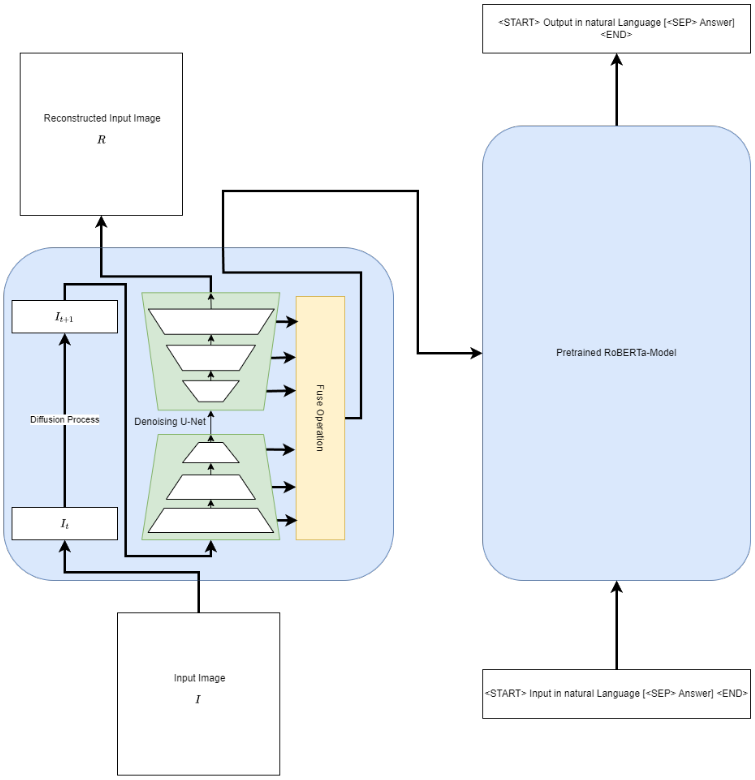
<START> Output in natural Language [<SEP> Answer] <END>
Reconstructed Input Image
R
I_{t+1}
Diffusion Process
Denoising U-Net
Fuse Operation
Pretrained RoBERTa-Model
I_t
Input Image
I
<START> Input in natural Language [<SEP> Answer] <END>

Figure 1: Visualization of the planned architecture. The schematic description of the RoBERTa-Model is omitted, since this is essentially the standard Transformer encoder as described in Vaswani et al.[3]. An image is encoded by aggregating the internal representations of the U-Net during the inverse diffusion process. It is than passed through an fuse layer which is responsible for the aggregation of multiple hierarchical states during different timesteps of the inverse diffusion process.

4 Datasets and Evaluation

Once we implemented the described approach, we will train and evaluate the model using two publicly available datasets that contain remote sensing images and incorporate different VLM-tasks. The following sections will describe the datasets and the evaluation method in further detail.

4.1 Datasets

In this project, we conduct an thorough evaluation of our approach utilizing two distinct datasets, each of which presents unique challenges in the domain of VL tasks.

The first dataset, known as FloodNet[16], is a post-disaster UAV dataset captured in after the Hurricane Harvey in Texas, USA. It comprises of high-resolution 4000x3000 pixel images, this dataset serves as a VQA benchmark in the remote sensing domain, providing approximately 11,000 image-question pairs. On average, there are 3.5 questions per image, hand-crafted and categorized into distinct question types, including yes-no queries, simple and complex counting inquiries, and condition recognition. The dataset authors also included segmentation masks and baseline model evaluations for tasks such as classification, semantic segmentation, and VQA. They reported the evaluated accuracy on each type of question.

The second dataset that is used is called LEVIR-CC[17]. It contains change captioning data consisting of an image pair at two different timestamps and five sentences describing any observed changes. Comprising approximately 10,000 256x256 satellite images, the dataset is gathered using the Google Earth API. In addition, the authors developed an architecture to provide a solution to the change captioning problem. The implemented architecture was build of three stages a CNN-based feature extractor, a dual-branch transformer encoder and a transformer decoder. They evaluated the approach and reported BLEU scores at various n-gram levels[20], METEOR[21] , ROUGE-L[22] and CIDEr-D[23] scores.

4.2 Evaluation

In this study project, we will evaluate our approach using the described datasets. Our approach will be trained on the corresponding training splits, with model selection carried out on dedicated validation sets. We will use the test data partitions to compute scores according to the previously reported scores in the dataset literature. Moreover, in order to provide comparability, we substitute the initially described DDPM with a ViT architecture. Subsequently, we apply the same procedure to the modified architecture, thus enabling a meaningful comparative analysis with the reported benchmark scores.

References

[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, *Generative adversarial networks*, Jun. 10, 2014. DOI: 10.48550/arXiv.1406.2661. arXiv: 1406.2661[cs, stat]. [Online]. Available: http://arxiv.org/abs/1406.2661 (visited on 09/30/2023).

[2] J. Ho, A. Jain, and P. Abbeel, *Denoising diffusion probabilistic models*, Dec. 16, 2020. DOI: 10.48550/arXiv.2006.11239. arXiv: 2006.11239[cs, stat]. [Online]. Available: http://arxiv.org/abs/2006.11239 (visited on 07/14/2023).

[3] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, *Attention is all you need*, Aug. 1, 2023. DOI: 10.48550/arXiv.1706.03762. arXiv: 1706.03762[cs]. [Online]. Available: http://arxiv.org/abs/1706.03762 (visited on 10/02/2023).

[4] J. Wang, Z. Yang, X. Hu, *et al.*, *GIT: A generative image-to-text transformer for vision and language*, Dec. 15, 2022. DOI: 10.48550/arXiv.2205.14100. arXiv: 2205.14100[cs]. [Online]. Available: http://arxiv.org/abs/2205.14100 (visited on 09/29/2023).

[5] J.-B. Alayrac, J. Donahue, P. Luc, *et al.*, *Flamingo: A visual language model for few-shot learning*, Nov. 15, 2022. DOI: 10.48550/arXiv.2204.14198. arXiv: 2204.14198[cs]. [Online]. Available: http://arxiv.org/abs/2204.14198 (visited on 09/30/2023).

[6] S. Liu, L. Fan, E. Johns, Z. Yu, C. Xiao, and A. Anandkumar, *Prismer: A vision-language model with an ensemble of experts*, Mar. 11, 2023. DOI: 10.48550/arXiv.2303.02506. arXiv: 2303.02506[cs]. [Online]. Available: http://arxiv.org/abs/2303.02506 (visited on 09/29/2023).

[7] S. Lobry, D. Marcos, J. Murray, and D. Tuia, "RSVQA: Visual question answering for remote sensing data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 12, pp. 8555–8566, Dec. 2020, ISSN: 0196-2892, 1558-0644. DOI: 10.1109/TGRS.2020.2988782. arXiv: 2003.07333[cs]. [Online]. Available: http://arxiv.org/abs/2003.07333 (visited on 09/30/2023).

[8] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, Dec. 10, 2015. DOI: 10.48550/arXiv.1512.03385. arXiv: 1512.03385[cs]. [Online]. Available: http://arxiv.org/abs/1512.03385 (visited on 10/02/2023).

[9] R. Kiros, Y. Zhu, R. Salakhutdinov, *et al.*, *Skip-thought vectors*, Jun. 22, 2015. DOI: 10.48550/arXiv.1506.06726. arXiv: 1506.06726[cs]. [Online]. Available: http://arxiv.org/abs/1506.06726 (visited on 10/02/2023).

[10] T. Siebert, K. N. Clasen, M. Ravanbakhsh, and B. Demir, *Multi-modal fusion transformer for visual question answering in remote sensing*, Oct. 10, 2022. DOI: 10.48550/arXiv.2210.04510. arXiv: 2210.04510[cs]. [Online]. Available: http://arxiv.org/abs/2210.04510 (visited on 09/30/2023).

[11] S. Chang and P. Ghamisi, *Changes to captions: An attentive network for remote sensing change captioning*, Apr. 3, 2023. DOI: 10.48550/arXiv.2304.01091. arXiv: 2304.01091[cs]. [Online]. Available: http://arxiv.org/abs/2304.01091 (visited on 10/02/2023).

[12] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, *Image super-resolution via iterative refinement*, Jun. 30, 2021. DOI: 10.48550/arXiv.2104.07636. arXiv: 2104.07636[cs, eess]. [Online]. Available: http://arxiv.org/abs/2104.07636 (visited on 10/02/2023).

[13] C. Hawthorne, I. Simon, A. Roberts, *et al.*, *Multi-instrument music synthesis with spectrogram diffusion*, Dec. 12, 2022. DOI: 10.48550/arXiv.2206.05408. arXiv: 2206.05408[cs, eess]. [Online]. Available: http://arxiv.org/abs/2206.05408 (visited on 10/02/2023).

[14] W. G. C. Bandara, N. G. Nair, and V. M. Patel, *DDPM-CD: Remote sensing change detection using denoising diffusion probabilistic models*, Jun. 27, 2022. DOI: 10.48550/arXiv.2206.11892. arXiv: 2206.11892[cs]. [Online]. Available: http://arxiv.org/abs/2206.11892 (visited on 10/02/2023).

[15] R. Jing, F. Duan, F. Lu, M. Zhang, and W. Zhao, "Denoising diffusion probabilistic feature-based network for cloud removal in sentinel-2 imagery," *Remote Sensing*, vol. 15, no. 9, p. 2217, Jan. 2023, ISSN: 2072-4292. DOI: 10.3390/rs15092217. [Online]. Available: https://www.mdpi.com/2072-4292/15/9/2217 (visited on 10/02/2023).

[16] M. Rahnemoonfar, T. Chowdhury, A. Sarkar, D. Varshney, M. Yari, and R. Murphy, *FloodNet: A high resolution aerial imagery dataset for post flood scene understanding*, Dec. 5, 2020. DOI: 10.48550/arXiv.2012.02951. arXiv: 2012.02951[cs]. [Online]. Available: http://arxiv.org/abs/2012.02951 (visited on 09/29/2023).

[17] C. Liu, R. Zhao, H. Chen, Z. Zou, and Z. Shi, "Remote sensing image change captioning with dual-branch transformers: A new method and a large scale dataset," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–20, 2022. DOI: 10.1109/TGRS.2022.3218921.

[18] Y. Liu, M. Ott, N. Goyal, *et al.*, *RoBERTa: A robustly optimized BERT pretraining approach*, Jul. 26, 2019. DOI: 10.48550/arXiv.1907.11692. arXiv: 1907.11692[cs]. [Online]. Available: http://arxiv.org/abs/1907.11692 (visited on 09/29/2023).

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, *An image is worth 16x16 words: Transformers for image recognition at scale*, Jun. 3, 2021. DOI: 10.48550/arXiv.2010.11929. arXiv: 2010.11929[cs]. [Online]. Available: http://arxiv.org/abs/2010.11929 (visited on 06/26/2023).

[20] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318. DOI: 10.3115/1073083.1073135. [Online]. Available: https://aclanthology.org/P02-1040 (visited on 10/29/2023).

[21] S. Banerjee and A. Lavie, "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*, Ann Arbor, Michigan: Association for Computational Linguistics, Jun. 2005, pp. 65–72. [Online]. Available: https://aclanthology.org/W05-0909 (visited on 10/29/2023).

[22] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013 (visited on 10/29/2023).

[23] R. Vedantam, C. L. Zitnick, and D. Parikh, *CIDEr: Consensus-based image description evaluation*, Jun. 2, 2015. DOI: 10.48550/arXiv.1411.5726. arXiv: 1411.5726[cs]. [Online]. Available: http://arxiv.org/abs/1411.5726 (visited on 10/29/2023).