RnD Project Report
Cross domain image captioning and Style Transfer

Sandeepan Naskar (200050126)

April 2024

Contents

1 Introduction

1.1 Image captioning

The task was to generate descriptive textual captions for images. In this work, we implement the following using a transformer-based model (GPT-2 here). We explore the integration of GPT-2-based image captioning with image-to-image diffusion models to achieve cross-styled image generation through style transfer. The goal is to generate visually appealing images by combining the rich contextual understanding provided by the GPT-2 transformer with the image manipulation capabilities of diffusion models.

1.2 Cross styling of the image

The process involves generating a textual description of an image using the GPT-2 transformer. This caption, along with the original image and a style change prompt, is then fed into the image-to-image diffusion model. The diffusion model utilizes the information from the caption and the style change prompt to generate a new image that retains the content of the original image while adopting the desired stylistic characteristics specified by the prompt.

1.3 Advantages

This approach offers several advantages. Firstly, by leveraging the contextual understanding of the image provided by the GPT-2 model, we aim to generate more accurate and contextually relevant captions. Secondly, by incorporating the image and the style change prompt into the diffusion model, we can achieve finer control over the style transfer process, allowing for more diverse and creative image manipulations over GANs

2 Papers reviewed

2.1 Attention is All you Need

The paper introduces the Transformer, a novel network architecture solely based on attention mechanisms, eliminating the need for recurrent or convolutional layers. This model connects encoder and decoder through attention, resulting in superior performance, greater parallelizability, and faster training times compared to traditional models. Experiments on machine translation tasks demonstrate its effectiveness, achieving state-of-the-art results with significantly lower training costs. The Transformer outperforms existing models, including ensembles, achieving 28.4 BLEU on English-to-German translation and setting a new state-of-the-art single-model BLEU score of 41.8 on English-to-French translation. Additionally, the Transformer generalizes well to other tasks, such as English constituency parsing, with both large and limited training data.

2.2 Multi-Modality Cross Attention Network for Image and Sentence Matching

The key to image and sentence matching is accurately measuring the visual-semantic similarity between them. However, most existing methods focus solely on intra-modality relationships within each modality or inter-modality relationships between image regions and sentence words. In contrast, our work proposes a novel MultiModality Cross Attention (MMCA) Network, which jointly models both intra-modality and inter-modality relationships in a unified deep model.

In MMCA, we introduce a cross-attention mechanism that leverages intra-modality relationships within each modality and inter-modality relationships between image regions and sentence words. This allows for mutual enhancement between the two modalities, enhancing image and sentence matching.

Experimental results on standard benchmarks like Flickr30K and MS-COCO show that our proposed model outperforms state-of-the-art methods for image and sentence matching.

2.3 CAT: Cross Attention in Vision Transformer

The widespread use of Transformers in NLP has sparked interest in their potential applications in computer vision (CV). However, the computational requirements for Transformer-based models in CV, particularly those replacing word tokens with image patches (e.g., ViT), are considerable, leading to training and inference bottlenecks.

In this paper, we propose a new attention mechanism in Transformer called Cross Attention. This mechanism focuses attention within image patches to capture local information, reducing computation compared to processing the entire image. Additionally, attention is applied between image patches to capture global information, further reducing computational load. By alternately applying attention within and between patches, we implement cross attention, maintaining performance while lowering computational costs.

We introduce a hierarchical network named Cross Attention Transformer (CAT) for various vision tasks. Our base model achieves state-of-the-art results on ImageNet-1K and improves performance on COCO and ADE20K datasets compared to other methods. This illustrates the potential of our network to serve as a general backbone for vision tasks.

2.4 Deep Reinforcement Learning-Based Image Captioning with Embedding Reward

Image captioning is a complex task due to the challenge of understanding image content and describing it in natural language. Recent advancements in deep neural networks have greatly enhanced performance in this area. While most state-of-the-art methods use an encoder-decoder framework with sequential recurrent prediction, we introduce a novel decision-making framework for image captioning.

Our approach involves two networks: a "policy network" and a "value network," working collaboratively to generate captions. The policy network provides local guidance by predicting the next word based on the current state's confidence. The value network, on the other hand, provides global and lookahead guidance by evaluating all possible extensions of the current state. Essentially, it adjusts the goal of predicting correct words towards generating captions similar to ground truth captions.

Both networks are trained using an actor-critic reinforcement learning model, with a unique reward defined by visual-semantic embedding. Extensive experiments on the Microsoft COCO dataset demonstrate that our framework outperforms state-of-the-art methods across various evaluation metrics.

2.5 Towards bridging the semantic gap between image and text: An empirical approach

The "semantic gap," which refers to the disparity between low-level features and semantic meanings in images, has been recognized for decades. However, resolving this gap remains a longstanding challenge in the field. In this work, we review the semantic gap problem and survey recent efforts to bridge it.

We emphasize that supervised learning plays a crucial role in bridging the semantic gap today. To illustrate this, we draw experiences from two application domains: object detection and metric learning for content-based image retrieval (CBIR).

Initially, this paper provides a historical retrospective on supervision, transitioning gradually to modern data-driven methodologies and introducing commonly used datasets. Subsequently, it summarizes various supervised learning methods used to bridge the semantic gap, particularly in the context of object detection and metric learning.

2.6 SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations

Guided image synthesis allows users to create and modify photo-realistic images effortlessly. The main challenge lies in balancing faithfulness to user input, such as hand-drawn strokes, and realism of the synthesized image. Current GAN-based methods struggle with this balance, often requiring additional data or loss functions for specific applications.

To overcome these challenges, we propose Stochastic Differential Editing (SDEdit), a new method based on a diffusion model generative prior. SDEdit synthesizes realistic images by iteratively denoising through a stochastic differential equation (SDE). By adding noise to the input image and then denoising it, SDEdit enhances realism without task-specific training or inversions, naturally achieving the balance between realism and faithfulness.

In a human perception study, SDEdit outperforms state-of-the-art GAN-based methods by up to 98.09% in realism and 91.72% in overall satisfaction

scores across various tasks, including stroke-based image synthesis, editing, and image compositing.

3 Implementation

3.1 Image Captioning

Firstly implemented an image captioning mechanism using a modified model architecture that incorporated features of the GPT-2 model. This involved adapting the traditional image captioning framework to integrate aspects of GPT-2's architecture.

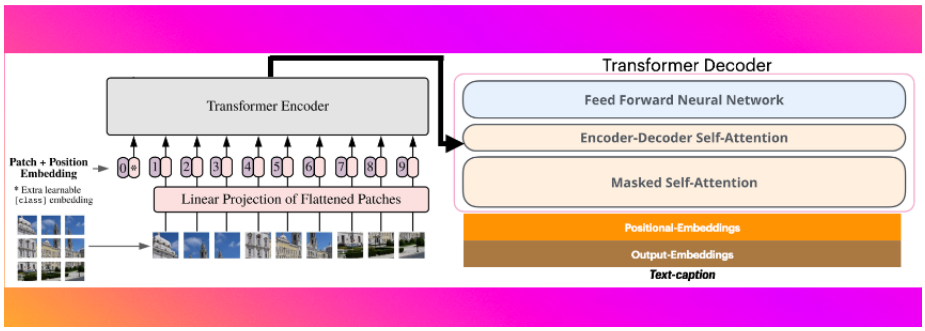


Figure 1: Image captioning GPT-2 transformer Architecture

Here's a more detailed breakdown:
- Traditional Image Captioning Framework: Image captioning typically involves an encoder-decoder architecture. The encoder processes the image to extract features, while the decoder generates captions based on these features.
- Modification for GPT-2 Features: In our approach, we modified this framework to incorporate features from the GPT-2 model. GPT-2 is a language model designed to generate human-like text, capturing long-range dependencies in sequences.
- Integration of GPT-2 Features: The modified architecture allowed us to integrate GPT-2's capabilities into the image captioning process. This integration involved several steps:

- **Encoding Images:** Images were encoded using a pre-trained model, such as a convolutional neural network (CNN), to extract features.
- **Incorporating GPT-2:** GPT-2's features were incorporated into the decoder part of the model. This could involve various strategies, such as initializing the decoder with GPT-2 weights or using GPT-2 layers alongside traditional decoder layers.
- **Generating Captions:** The model generated captions based on the combined image and GPT-2 features. This process often involves a softmax

layer to predict the next word in the caption sequence.

- Training and Fine-Tuning: The model was trained and fine-tuned on a dataset of images with corresponding captions. This involved optimizing the model's parameters to minimize the difference between predicted and ground truth captions.
- Evaluation: The performance of the model was evaluated using metrics such as BLEU (Bilingual Evaluation Understudy) score, which measures the similarity between generated and reference captions.

Overall, this approach leveraged the strengths of both image processing and natural language understanding, resulting in an image captioning mechanism that could potentially generate more contextually relevant and human-like captions.

3.2 Prompt generation

We have stored the captions of each of these images in the a `caption-output-files` directory. We now tokenized a set of styling prompts into styling category, actions and effects and by using a random shuffling function we throw randomly generated prompts. We then run the `prompt-generator.py` file to have the `style-prompt` directory populated.

We will now use these styled prompts and captions and run it on out Diffusion model.

3.3 Stable Diffusion Generator

In the code we first resized all out input images to be of size=(769, 512) resolution in order to make the input dimensions consistent across all input images the result of which we can see in our output images too that are stored in the `output-images` directory.

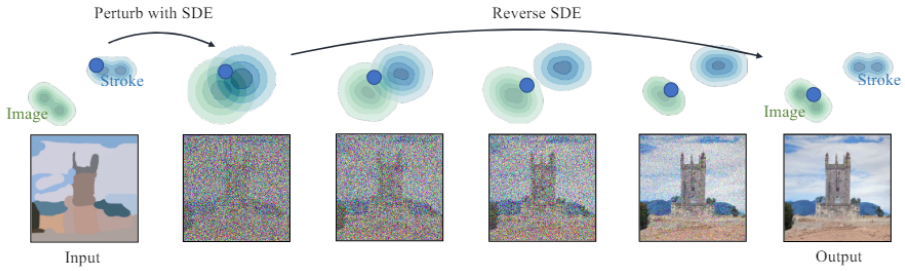


Figure 2: SDEdit Image to Image Diffusion

SDEdit, or Stochastic Differential Editing, is an architecture designed for guided image synthesis and editing. It operates on the principle of iteratively denoising images through a stochastic differential equation (SDE) to enhance

realism while incorporating user guidance. At its core, SDEdit consists of the following steps:

1. **Initialization**: Start with an input image.
2. **Noise Addition and Denoising**:Add noise to the input image based on diffusion model parameters and then iteratively denoise the image through the SDE.
3. **Incorporating User Guidance**: Incorporate user guidance (using styling prompts in this case) into the denoising process to enhance faithfulness to user input.
4. **Iterations**:Repeat the noise addition and denoising process for a predefined number of iterations.
5. **Output**:Return the synthesized image.

This architecture allows SDEdit to generate and edit photo-realistic images while balancing realism and faithfulness to user guidance. It is state-of-the-art in terms of realism and overall satisfaction scores, making it suitable for various tasks, including stroke-based image synthesis, editing, and image compositing.

Algorithm 1 Guided image synthesis and editing with SDEdit (VE-SDE)

Require: $\mathbf{x}^{(g)}$ (guide), t_0 (SDE hyper-parameter), N (total denoising steps)
  $\Delta t \leftarrow \frac{t_0}{N}$
  $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
  $\mathbf{x} \leftarrow \mathbf{x} + \sigma(t_0)\mathbf{z}$
  for $n \leftarrow N$ **to** 1 **do**
    $t \leftarrow t_0 \frac{n}{N}$
    $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
    $\epsilon \leftarrow \sqrt{\sigma^2(t) - \sigma^2(t - \Delta t)}$
    $\mathbf{x} \leftarrow \mathbf{x} + \epsilon^2 \boldsymbol{s_\theta}(\mathbf{x}, t) + \epsilon\mathbf{z}$
  end for
  Return $\mathbf{x}$

Figure 3: SDEdit algorithm

4 Experiments

- **Cross Attention Transformer**
  - **Description**: Implemented a Cross Attention Transformer architecture to explore the potential of Transformer models in computer vision tasks.
  - **Challenges**: The Cross Attention Transformer struggles with handling long-range dependencies in images due to its attention mechanism, leading to difficulty in capturing contextual information effectively.

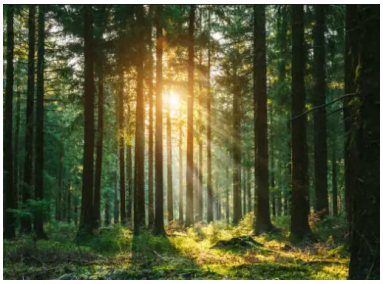
- **Deep RL-Based Image Captioning**
  - **Description**: Developed an image captioning mechanism using a Deep Reinforcement Learning (RL) approach to investigate the effectiveness of RL in generating descriptive captions for images.
  - **Challenges**: Deep RL-based image captioning suffers from instability during training, leading to difficulties in convergence and generating coherent captions.
- **Multi-Modality Cross-Attention Network for Image Captioning**
  - **Description**: Implemented a Multi-Modality Cross-Attention Network for image captioning to leverage cross-modal attention mechanisms for a better understanding of images and their corresponding captions.
  - **Challenges**: The Multi-Modality Cross-Attention Network faces challenges in effectively integrating information from different modalities, leading to inconsistencies between images and captions.
- **Image Caption Matching with CLIP Model**
  - **Description**: Explored image caption matching using the pretrained CLIP model to evaluate the performance of CLIP in matching images and their captions.
  - **Challenges**: The CLIP model struggles with generalization to unseen images or captions, leading to poor performance in caption matching tasks.
- **Image Reconstruction with GAN and Cross Attention Transformers**
  - **Description**: Implemented an architecture combining GAN and Cross Attention Transformers for image reconstruction to explore methods for reconstructing images from captions and input images.
  - **Challenges**: The combination of GAN and Cross Attention Transformers faces difficulties in effectively leveraging both image and text information, leading to poor quality image reconstructions.

5 Results

5.1 Code

The code is in the following github link: `https://github.com/Sandeepan-Naskar/Style-transfer-for-image-captioning`

5.2 Outputs



(Original) Captioned: "a forest filled with lots of trees and bushes"

(Cross-styled) Prompt: "Apply an image into a sketch"

(Original) Captioned: "a large brown cow standing on top of a snow covered field"

(Cross-styled) Prompt: "Turn an image into a painting"

6 References

- Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, Stefano Ermon: SDEdit: Image Synthesis and Editing with Stochastic Differential Equations. CoRR abs/2108.01073 (2021)
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin: Attention Is All You Need. CoRR abs/1706.03762 (2017)

- Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, Feng Wu: Multi-Modality Cross Attention Network for Image and Sentence Matching. CVPR 2020: 10938-10947
- H. Lin, X. Cheng, X. Wu and D. Shen, "CAT: Cross Attention in Vision Transformer," 2022 IEEE International Conference on Multimedia and Expo (ICME), Taipei, Taiwan, 2022, pp. 1-6, doi: 10.1109/ICME52920.2022.9859720.
- Z. Ren, X. Wang, N. Zhang, X. Lv and L. -J. Li, "Deep Reinforcement Learning-Based Image Captioning with Embedding Reward," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 1151-1159, doi: 10.1109/CVPR.2017.128.
- Javed, Hira et al. 'Towards Bridging the Semantic Gap Between Image and Text: An Empirical Approach'. 1 Jan. 2024 : 1 – 13.
- `https://huggingface.co/nlpconnect/vit-gpt2-image-captioning`
- `https://huggingface.co/papers/2108.01073`