Master's Thesis proposal: Data Augmentation Using Latent Diffusion Models for Object Detection

Ahmed Sheta

October 23, 2024

Supervised by

Mathias Zinnen

Introduction

The Object Detection for Olfactory References (ODOR) dataset, a collection of publicly available artwork images, consists a total of 38,116 object-level annotations 4,712 images, spanning an extensive set of 139 fine-grained categories. However, the dataset faces a significant challenge: the number of annotated instances varies greatly between classes. While some classes have as many as 5,800 annotations, others have only three. This extreme shortage of annotations in certain classes makes it difficult to train a robust and effective object detection model. Conventional data augmentation methods may fall short in effectively addressing this issue, this thesis proposes using advanced data augmentation techniques powered by pretrained Latent Diffusion Models (LDMs) and ControlNet models.

Methodology

To achieve the thesis objectives, the following methodology will be employed:

- **Train multiple object detection models:** A baseline object detection model shall be trained using the original ODOR dataset to establish a reference point for evaluating the impact of data augmentation. On top of that, models shall be trained on the synthesized images.
- **Data augmentation with LDMs and ControlNet:** Pretrained LDMs shall be leveraged to perform inpainting, generating synthetic images that augment the dataset. Moreover, levering ControlNet's edge conditioning may help augmenting the data while preserving the primary intrinsic characteristics. This process will focus on classes with few annotated instances to increase the diversity and quantity of training data. The synthesized images' annotations shall be adapted according the the image sizes.

- **Explore different mask creation strategies:** Multiple strategies for creating masks shall be explored, with the goal of optimizing the regions that are inpainted by the diffusion models. The new masking ideas so far include the following:
  - **Saliency-based masking:** Utilizing saliency maps to highlight the most informative regions for inpainting by diffusion models.
  - **Gradient-based Masking:** Employing gradient information to guide mask generation, focusing on areas with significant changes in pixel intensity.
  - **Histogram of Oriented Gradients (HOG)-based masking:** Creating masks based on the distribution of gradient orientations to capture key structural features.
  - **Entropy-based masking:** Leveraging entropy to identify regions with high uncertainty or information content for targeted inpainting.
  - **Guided Masking Strategy:** Attempting to develop a novel approach that combines saliency-based masking with object bounding boxes to pinpoint the most informative regions. This strategy optimizes mask placement around these areas, enabling the model to augment data by focusing on critical parts while maintaining the object's intrinsic characteristics.
  - **Masking random parts of the object:** Masking random parts of the object in the image to varying extents, introducing diversity on different spots of the object.
  - **Masking random parts of the context:** Masking parts of the background or context surrounding the object to shift focus away from non-essential elements and emphasize the object's features.
  - **Masking the border region around objects:** Giving diversity boundary areas around the objects, letting the object detection model focus its efforts on regions that define the object's shape and edge characteristics rather than the surroundings.
- **Fine-tuning the pretrained ControlNet for ODOR images:** The data distribution of the artwork objects may differ from that the real world objects, the pretrained ControlNet shall be fine-tuned on the ODOR images, aiming to induce a bias towards artwork images to better match the artwork visual characteristics.
- **Evaluation of data augmentation**
  - To evaluate the different masking strategies (and potentially their combinations), a smaller training set of 500-5000 synthetic samples will be generated.
  - The detection model will be trained on these subsets and evaluated on a non-synthetic test set.
  - For one or a few of the best-performing strategies or strategy combinations, we will scale up the synthetic training data generation to assess if we can outperform the baseline detection method.
  - Additionally, we will evaluate a training scheme where the model is pretrained with a large amount of synthetic data and then fine-tuned using the real training set.