Project plan, Bachelor's project, s194119

Project title

EyeFormer: Learning to Detect Objects from Eye-Tracking Data using a Vision Transformer

Initial Project Description

In this project, we will revisit the idea of using a ”hands-free” way to annotate objects in images and training object detectors using eye-tracking data[1]. Instead of carefully marking every training image with accurate bounding-boxes, the annotators only need to find the target object in the image, look at it, and press a button. By tracking the eye movements of the annotators while they perform this task, we obtain valuable information about the position and size of the target object. Unlike bounding-box annotation, the eye tracking task requires no annotation guidelines and can be carried out by completely naive viewers. Furthermore, the task can be performed in a fraction of the time it takes to draw a bounding-box (about one second per image).

The goal of this project is to build upon the idea of [1] and build a transformer-based deep learning model that can infer the location of a target object (i.e., bounding box) given the eye fixations on the image. Given the recent success of the attention mechanism and the text transformers [2] in the computer vision community for various image recognition tasks using Vision Transformers [3], we will propose a vision and eye-tracking Transformer that can learn to predict the location of the target object given only human eye fixations on the images.

Initial project-plan

Start-date: 01/09/2022
End-date: 24/11/2022
Credits: 15 ECTS
Councellor: Dimitrios Papadopoulos
Institute: DTU Compute

Revised project description

Annotating image datasets is a tedious and often expensive task in deep learning workflows, especially when domain-expertise is required. Depending on the task at hand, the large variation in image-data makes training deep learning models a big and sometimes unfeasible undertaking when the goal is production-grade computer vision models for very specific tasks, due to the need of sufficient large datasets for training models to acceptable performance. In an effort to overcome this difficulty, this Bachelor's project proposes a transformer-based image annotation pipeline based on eye-tracking data alone from experiment participants in [1] for inferring object bounding-boxes in images, which could be used downstream for annotation in other deep learning pipelines. Hopefully, this may provide a faster alternative to manual image labelling.

Using the eye-tracking data from [1], a variation of a subset of the PASCAL VOC 2012 data-set[4], it may be possible to train vision-transformer models which can generate bounding boxes, segment, or classify class instances. The data-set consists of eye-tracking data from 5 different participants and contains 10 classes. The proposed data-pipeline is to reshape the data-set and corresponding eye-tracking coordinates to a squared format, feeding it through a pretrained Vision-Transformer-encoder [5] in order to obtain a latent space representation of fixation-points per image. Subsequently, a bilinear interpolation is performed for obtaining a coordinate-space which represents the input-coordinates to the ViT. Fixations in this latent space representation are used analogously to words in a Natural Language Processing Transformer architecture. Thus, a series of fixations in the latent space representation form an input-sequence for an additional not-pretrained Transformer-model. Depending on the decoder-architecture, the output could be a bounding-box-prediction, a verdict of object class or a complete segmentation mask. The goal of the project is to investigate whether this general idea is feasible for the given dataset, concretized in the section: Concrete goal definition, to some degree of success compared to baselines of: 1. More traditional image-analytic methods applied on the PASCAL VOC2012 Dataset[4] and 2. A Transformer-model which only uses spatial coordinates of fixations (ie. without latent-space-representation) to infer bounding-boxes. If time allows it, additional baselines may be used for comparison. The provided dataset which is used consists of 6270 (10/20 classes of the PASCAL VOC2012 set) annotated images. For a start, models are trained class-wise for a subset of classes. Thus, a small subset of class-instances containing fixation-data and ground-truth bounding boxes will be used for model-training. In order to choose classes, a minor exploratory data analysis will be applied for choosing classes which intuitively may be considered ”easy”. This analysis will not be based on the image-data itself as this is considered out of scope for the project, but rather on the hardness of the object-detection task given to experiment participants in [1] - e.g. how long time does it take to find and identify the given object, how many tracking-fixations are inside object-boundaries, etc. Ideally, the applied test/train-split will contain all classes in the dataset. This will be implemented if progress allows it, but is not considered a main-goal, as the projects focus is about implementing the pipeline and investigating the novel possibilities which may arise of pair-applied eye-tracking and Vision Transformer-technologies.

Concrete goal definitions

1. Exploratory data-analysis on the eye-tracking/bounding-box dataset
   - Define intuitive measures, which as a whole model detection-hardness of classes
   - Based on the measures, choose a fitting number of easy classes for the data-pipe
2. Learn about transformer-models, and followingly Vision Transformers (no prior experience)
3. Learn to implement a Vision Transformer model using PyTorch in Python
4. Become familiar with using DTU’s HPC-systems for project-work
5. Build a baseline ”blind” ViT-model which tries to predict bounding-boxes based on eye-tracking data alone. Ie. the input is simply a series of coordinates.
6. Implement and apply a pre-trained Vision-Transformer for feature-extraction
   - Evaluate performance
7. Implement a transformer-model, which both uses eye-tracking-data and ViT-extracted features to infer bounding boxes
   - Evaluate performance
   - Run some experiments:

- Compare using mean-tracking point of both eyes vs. two-signal approach (right eye, left eye)
- Investigate on which types of data performance is better vs. worse in order to propose data-alterations etc. for later projects
- If time allows it, investigate if more data-features could increase model-performance (e.g. fixation duration)

8. Formulate the methodology and results of the project clearly in a self-written report.

Revised project-plan

Start-date: 01/09/2022
End-date: 15/12/2022
Credits: 15 ECTS
Councellor: Dimitrios Papadopoulos
Institute: DTU Compute

References

[1] Dim P. Papadopoulos et al. "Training Object Class Detectors from Eye Tracking Data". Ed. by David Fleet et al. Cham, 2014.

[2] Ashish Vaswani et al. "Attention Is All You Need". 2017. DOI: 10.48550/ARXIV.1706.03762. URL: https://arxiv.org/abs/1706.03762.

[3] Alexey Dosovitskiy et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: (2020). DOI: 10.48550/ARXIV.2010.11929. URL: https://arxiv.org/abs/2010.11929.

[4] M. Everingham et al. "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results". http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[5] Ross Wightman. "PyTorch Image Models". https://github.com/rwightman/pytorch-image-models. 2019. DOI: 10.5281/zenodo.4414861.

[6] Albert Einstein. "Zur Elektrodynamik bewegter Körper. (German) [On the electrodynamics of moving bodies]". In: *Annalen der Physik* 322.10 (1905), pp. 891–921. DOI: http://dx.doi.org/10.1002/andp.19053221004.

[7] Olga Russakovsky et al. "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision (IJCV)* 115.3 (2015), pp. 211–252. DOI: 10.1007/s11263-015-0816-y.

[8] Mark Everingham et al. "The Pascal Visual Object Classes Challenge: A Retrospective". In: *Int. J. Comput. Vision* 111.1 (Jan. 2015), pp. 98–136. ISSN: 0920-5691. DOI: 10.1007/s11263-014-0733-5. URL: https://doi.org/10.1007/s11263-014-0733-5.

[9] Ross Girshick. "Fast R-CNN". 2015. DOI: 10.48550/ARXIV.1504.08083. URL: https://arxiv.org/abs/1504.08083.

[10] Kaiming He et al. "Mask R-CNN". 2017. DOI: 10.48550/ARXIV.1703.06870. URL: https://arxiv.org/abs/1703.06870.

[11] Dim P. Papadopoulos et al. "We don't need no bounding-boxes: Training object class detectors using only human verification". 2016. DOI: 10.48550/ARXIV.1602.08405. URL: https://arxiv.org/abs/1602.08405.

[12] Shaoqing Ren et al. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". 2015. DOI: 10.48550/ARXIV.1506.01497. URL: https://arxiv.org/abs/1506.01497.

[13] Hao Su, J. Deng, and L. Fei-Fei. "Crowdsourcing annotations for visual object detection". In: (Jan. 2012), pp. 40–46.