figure 1

Image Segmentation

22.09.2022

Andrew Sinout Shenouda
900182668
https://github.com/andrewsinout

Youssef Ashraf Kandil
900182405
https://github.com/kanndil

Introduction and Motivation Statement

Image processing has been one of the hottest research fields in the past decade since the continuing advances in hardware capabilities, in both aspects of sensors and processing. In the current decade, there is a rising need not only to extract data from images but also, to learn unprecedented knowledge from patterns in images. The research in Panoptic Segmentation using Deep machine learning is having high demand since it is becoming essential in autonomous driving, medical examination, and many others.

Problem Statement

Creating a Panoptic segmentation model for street environments, that can be embedded in autonomous driving vehicles, as well as traffic control and surveillance systems.

Goals of Panoptic Segmentation

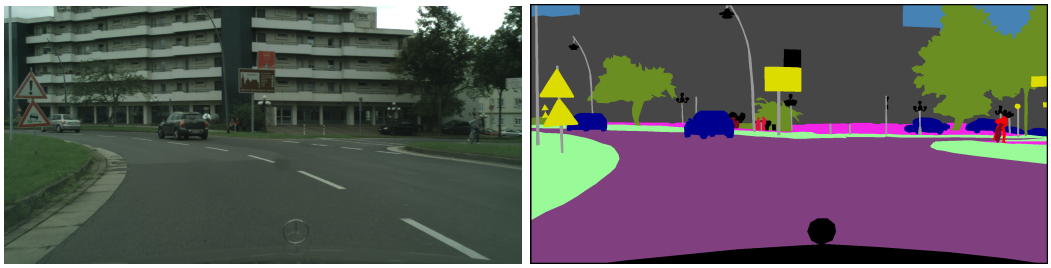
As per [Panoptic Segmentation | Papers With Code]

Panoptic image segmentation unifies between

1. Semantic segmentation (labeling each pixel to a predefined class)
2. Instance segmentation (Objects segmentation and detection)

Input/Output examples

Example 1:



Example 2:

Evaluation metrics

According to the article titled "Semantic vs. Instance vs. Panoptic Segmentation - PyImageSearch"

There are two things to be considered in image segmentation (things and stuff). Things are well-defined shapes objects. Stuff are thebackground regions.

The main metric in panoptic segmentation is called Panoptic Quality (PQ). This metric assesses the quality of the model using predicted masks and instance identifiers on both things and stuff. Panoptic Quality (PQ) multiplies segmentation quality (SQ), which is the score for the matched segments, and recognition quality (RQ) terms, which is the score of the recall values of predicted masks and their precision.

Also, as per the article, Metrics to Evaluate your Semantic Segmentation Model | by Ekin Tiu | Towards Data Science, we have these other evaluation metrics to use:

1. Pixel Accuracy
2. Intersection-Over-Union (Jaccard Index)
3. Dice Coefficient (F1 Score)

Current State-Of-The-Art model results

According to the evaluation metrics discussed above, the main measure of Panoptic segmentation performance is the PQ index, accordingly, it is used to rank the highest performing models.

Mask DINO

Reference paper:

https://arxiv.org/pdf/2206.02777v1.pdf

Source code:

GitHub - IDEA-Research/MaskDINO: Official implementation of the paper "Mask DINO: Towards A Unified Transformer-based Framework for Object Detection and Segmentation"

Mask DINO is the current state-of-the-art model trained over the COCO dataset, scoring 59.4 in the PQ index. The model is a collaboration between the following institutions and was released in June 2022.

- The Hong Kong University of Science and Technology
- Dept. of CST., BNRist Center, Institute for AI, Tsinghua University
- International Digital Economy Academy (IDEA).
- The Hong Kong University of Science and Technology (Guangzhou)

Mask dino follows the recent movement of unifying the detection and segmentation models with one convolution-based model, where it uses a base model named DINO which is a leading model in the field of image processing and segmentation.

The model does the following modifications on DINO:

- Adds mask segmentation branch
- Construct a pixel embedding map that is obtained fromTransformer encoder features in the backbone layer
- Unified query selection for mask
- Unified denoising for mask

Model Results

Here are some comparisons of the results of the highest performing models in image segmentation.

Model	Epochs	Query type	PQ	PQ^{Th}	PQ^{St}	Box AP^{Th}_{pan}	Mask AP^{Th}_{pan}
DETR [2]	$500 + 25$	100 queries	43.4	48.2	36	–	31.1
Panoptic Segformer [19]	24	353 queries	49.6	54.4	42.4	–	41.7
Mask2Former* [4]	50	100 queries	51.9/51.5†	57.7	43.0	–	41.7
Mask DINO (ours)	50	100 queries	52.3	58.3	43.2	47.7	43.7
Mask DINO (ours)	50	300 queries	**53.0**(+1.1)	**59.1**(+1.4)	**43.9**(+0.9)	**48.8**	**44.3**(+2.6)
Mask DINO (ours)	**24**	300 queries	51.5	57.3	42.6	46.4	42.8
Mask2Former [4]	12	100 queries	46.9	52.5	38.4	–	37.2
Panoptic Segformer [19]	12	353 queries	48.0	52.3	41.5	–	–
Mask DINO (ours)	12	300 queries	**49.0**(+1.0)	54.8	40.2	43.2	**40.4**(+3.2)

Mask DINO: Towards A Unified Transformer-based Framework for Object Detection and Segmentation

Method	Params	Backbone	Backbone Pre-training Dataset	Detection Pre-training Dataset	val	
					w/o TTA	w/ TTA
Instance segmentation on COCO					AP	
Mask2Former [4]	216M	SwinL	IN-22K-14M	–	50.1	–
Soft Teacher [36]	284M	SwinL	IN-22K-14M	O365	51.9	52.5
SwinV2-G-HTC++ [23]	3.0B	SwinV2-G	IN-22K-ext-70M [23]	O365	53.4	53.7
MasK DINO(Ours)	**223M**	SwinL	IN-22K-14M	O365	**54.5**(+1.1)	–
Panoptic segmentation on COCO					PQ	
Panoptic SegFormer [19]	–M	SwinL	IN-22K-14M	–	55.8	–
Mask2Former [4]	216M	SwinL	IN-22K-14M	–	57.8	–
MasK DINO (ours)	223M	SwinL	IN-22K-14M	O365	**59.4**(+1.6)	–
Semantic segmentation on ADE20K					mIoU	
Mask2Former [4]	215M	SwinL	IN-22K-14M	–	56.1	57.3
Mask2Former [4]	217M	SwinL-FaPN	IN-22K-14M	–	56.4	57.7
SeMask-L MSFaPN-Mask2Former [14]	–M	SwinL-FaPN	IN-22K-14M	–	–	58.2
SwinV2-G-UperNet [23]	3.0B	SwinV2-G	IN-22K-ext-70M [23]	–	59.3	59.9
MasK DINO (ours)	**223M**	SwinL	IN-22K-14M	O365	**59.5**	**60.8**(+0.9)

Mask DINO: Towards A Unified Transformer-based Framework for Object Detection and Segmentation

Mask Dino clearly shows the highest performance with the least number of epochs, which is a dominant advantage above the other models. Moreover, it does not overfit by scaling the number of queries like other models such as Mask2Former, whose performance highly degrades when using 300 queries.

Training Dataset:

COCO (Microsoft Common Objects in Context)

Available datasets

- COCO (Microsoft Common Objects in Context)
    - Mask DIno, the State-of-the-art model, is trained for it (best results)
    - The dataset combines 164K images of numerous contexts, it is not limited to street environment
    - The dataset has labels for the following topics
        - object detection
        - captioning: natural language descriptions
        - keypoints detection
        - stuff image segmentation
        - panoptic: full scene segmentation
        - dense pose

- Cityscapes
    - Has available versions of 5000, or 20000 images sets
    - The dataset limited to street environment, thus suitable for our intended topic
        - Traffic control
        - Traffic surveillance
        - Autonomous driving
    - Many available projects are trained on it
    - To get the dataset:
        - Register https://www.cityscapes-dataset.com/login/
        - Confirm account from your email
        - Login and click download
        - Choose 5 000 annotated images to download (11 GB)
- KITTi
    - Contains hours of traffic recordings

Dataset description (Cityscapes Image Pairs)

This dataset is used in semantic segmentation for improving autonomous driving. There are different important aspects that made us choose this particular dataset to use with the model.

The dataset's comprehensiveness in terms of the number of classes and their definitions, how diverse they are, the size suitability of the dataset, and the metadata included in the dataset.

First, the features of the dataset and its polygonal annotations:

- Dense semantic segmentation: which indicates that it has great pixel labeling.
- Instance segmentation for vehicles and people: the focus of the data in the needed direction.

Second, the dataset comprehensiveness:

- Has a ground truth labeling for 30 distinct classes.
- The Data points are gathered from various periods of the years (spring, summer, fall)
- Here are the Class Definitions
- Classes are comprehensive, include most commonly seen objects in a street environment
- Many classes

Third, the dataset diversity, contains data with different:

- Day/time
- Good/medium weather conditions
- Manually selected frames
    - a Large number of dynamic objects
    - Varying scene layout
    - Varying background

Note: The data set is relatively large so to avoid running out of resources one of the following tackles can be done:

- Use a comprehensive well-balanced subset of the data set.
- Use small batch sizes.

Literature review

Mask DINO

(current state-of-the-art model): described in detail in the state-of-the-art section.

Panoptic SegFormer:

Reference paper:

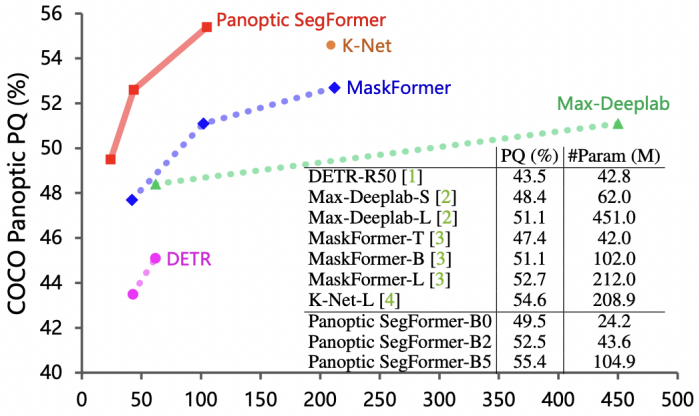
https://arxiv.org/pdf/2109.03814v4.pdf

Source code:

https://github.com/zhiqi-li/Panoptic-SegFormer

This model was considered state of the art till 2017 scoring 56.2 points in the PQ index, check the following two diagrams for comparison with prior models.

Method	Backbone	Epochs	PQ	PQ^{th}	PQ^{st}	#P	#F
Panoptic FPN [7]	R50	36	41.5	48.5	31.1	-	-
SOLOv2 [28]	R50	36	42.1	49.6	30.7	-	-
DETR [1]	R50	325	43.4	48.2	36.3	42.9	248
Panoptic FCN [21]	R50	36	43.6	49.3	35.0	37.0	244
K-Net [4]	R50	36	47.1	51.7	40.3	-	-
MaskFormer [3]	R50	300	46.5	51.0	39.8	45.0	181
Panoptic SegFormer	R50	**12**	48.0	52.3	41.5	51.0	214
Panoptic SegFormer	R50	24	**49.6**	**54.4**	**42.4**	51.0	214
DETR [1]	R101	325	45.1	50.5	37.0	61.8	306
Max-Deeplab-S [2]	Max-S [2]	54	48.4	53.0	41.5	61.9	162
MaskFormer [3]	R101	300	47.6	52.5	40.3	64.0	248
Panoptic SegFormer	R101	24	**50.6**	**55.5**	**43.2**	69.9	286
Max-Deeplab-L [2]	Max-L [2]	54	51.1	57.0	42.2	451.0	1846
Panoptic FCN [36]	Swin-L†	36	51.8	58.6	41.6	-	-
MaskFormer [3]	Swin-L†	300	52.7	58.5	44.0	212.0	792
K-Net [4]	Swin-L†	36	54.6	60.2	46.0	208.9	-
Panoptic SegFormer	Swin-L†	24	**55.8**	**61.7**	**46.9**	221.4	816
Panoptic SegFormer	PVTv2-B5†	24	55.4	61.2	46.6	104.9	349



Panoptic SegFormer: Delving Deeper into Panoptic Segmentation with Transformers

SegFormer Is a Panoptic segmentation model that introduced the following components to improve the Panoptic segmentation performance:

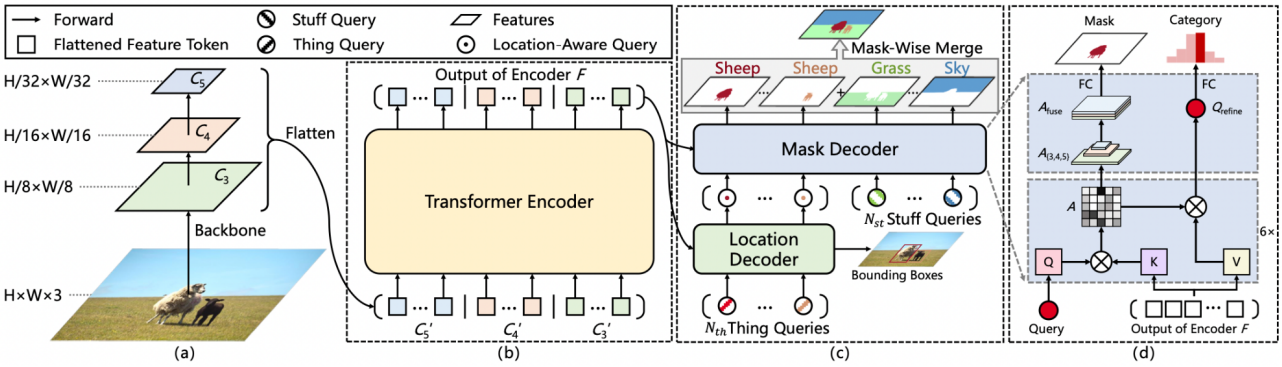
- An efficient deeply-supervised mask decoder
- A query decoupling strategy

  Decouples the responsibilities of the query set and avoids mutual interference between things and stuff.

- An improved post-processing method

  Resolves conflicting mask overlaps.

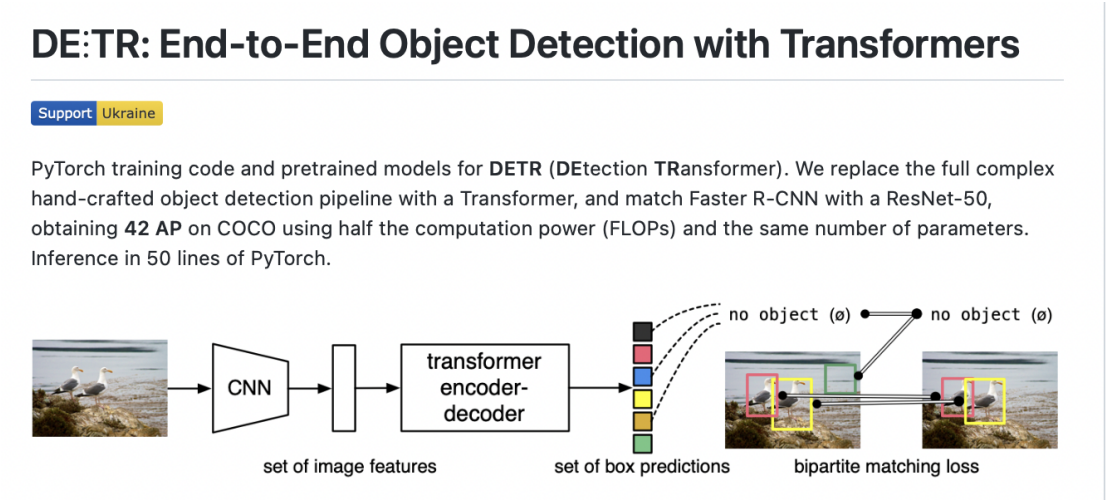
  Increased the PQ index points by 6.2% over the original DETR.



Panoptic SegFormer: Delving Deeper into Panoptic Segmentation with Transformers

A Deformable DETR (End-to-End Object Detection with Transformers) deep learning model, which is a fast and efficient version of DETR, is utilized by SegFormer for multi-scale

features processing. DETR is their main model, which uses a CNN (Convolutional neural network).



GitHub - facebookresearch/detr: End-to-End Object Detection with Transformers

DETR model zoo:

https://github.com/facebookresearch/detr#model-zoo

Colab notebook:

https://colab.research.google.com/github/facebookresearch/detr/blob/colab/notebooks/detr_attention.ipynb

Training Dataset:

COCO (Microsoft Common Objects in Context)

Maskedattention Mask Transformer (Mask2Former):

Reference paper:

https://arxiv.org/pdf/2112.01527v3.pdf

Source code:

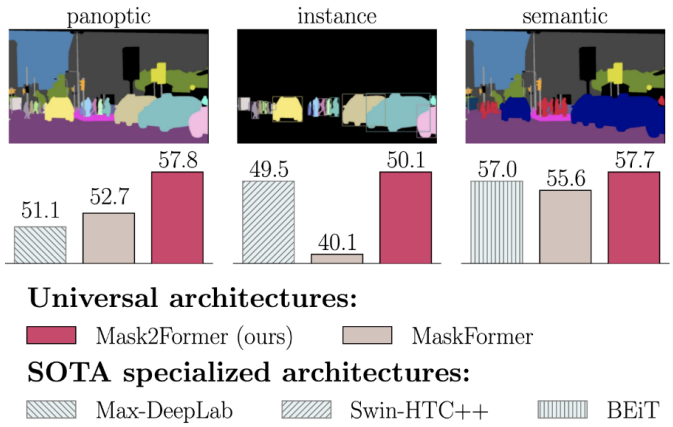
https://github.com/facebookresearch/Mask2Former

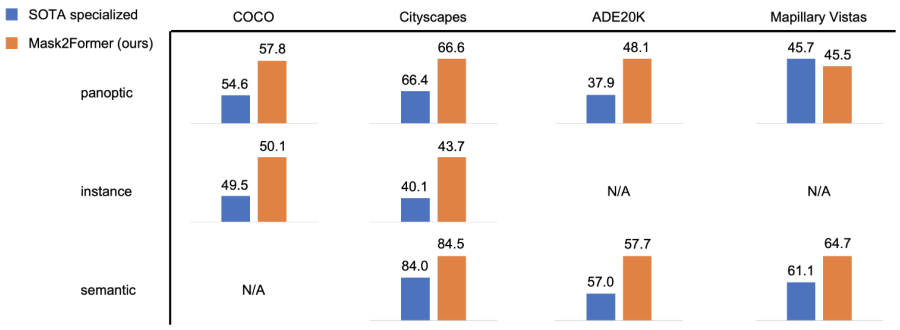
This is a relatively new architecture of image segmentation that is capable of presenting it in all ways (panoptic, instance, or semantic). It is composed of two main items:

- This paper uses masked attention instead of cross attention because the latest studies suggest that the cross attention layer is the reason behind very low-speed conversations, so it implements a mask attention layer instead. “For this, we propose masked attention, a variant of cross attention that only attends within the foreground region of the predicted mask for each query.” (according to the paper mentioned below).

- The paper also suggested a new implementation for the base model provided by the MaskFromer (previous innovation), which utilizes GPU computing more efficiently to speed up the training process which helps train over epochs more efficiently.

-

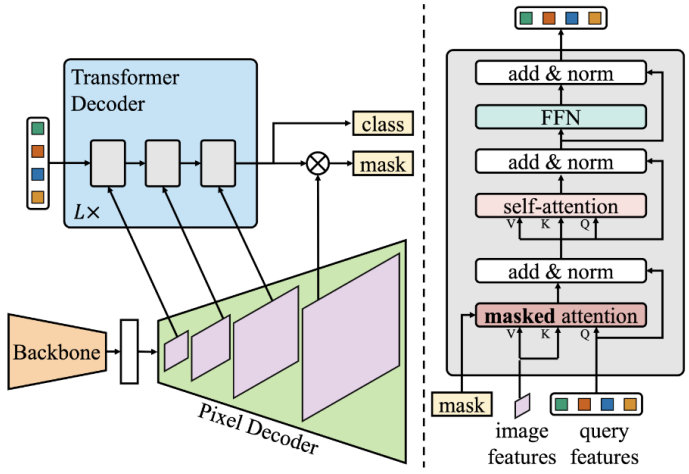




This comparison shows how Mask2former performance in the 3 different elements of image segmentation (panoptic, instance, or semantic) compared to the best models of each element with the different datasets.

method	backbone	query type	epochs	PQ	PQ^{Th}	PQ^{St}	AP^{Th}_{pan}	$mIoU_{pan}$	#params.	FLOPs	fps
DETR [5]	R50	100 queries	500+25	43.4	48.2	36.3	31.1	-	-	-	-
MaskFormer [14]	R50	100 queries	300	46.5	51.0	39.8	33.0	57.8	45M	181G	17.6
Mask2Former (ours)	R50	100 queries	50	**51.9**	**57.7**	**43.0**	**41.7**	**61.7**	44M	226G	8.6
DETR [5]	R101	100 queries	500+25	45.1	50.5	37.0	33.0	-	-	-	-
MaskFormer [14]	R101	100 queries	300	47.6	52.5	40.3	34.1	59.3	64M	248G	14.0
Mask2Former (ours)	R101	100 queries	50	**52.6**	**58.5**	**43.7**	**42.6**	**62.4**	63M	293G	7.2
Max-DeepLab [52]	Max-L	128 queries	216	51.1	57.0	42.2	-	-	451M	3692G	-
MaskFormer [14]	Swin-$L^{\dagger}$	100 queries	300	52.7	58.5	44.0	40.1	64.8	212M	792G	5.2
K-Net [62]	Swin-$L^{\dagger}$	100 queries	36	54.6	60.2	46.0	-	-	-	-	-
Mask2Former (ours)	Swin-$L^{\dagger}$	*200 queries*	*100*	**57.8**	**64.2**	**48.1**	**48.6**	**67.4**	216M	868G	4.0

Until the recent release of Mask-Dino the current state of the art, Mask2Former was the leading model with a score of 57.8 in the PQ index.



Mask2Former adopts the same meta architecture as MaskFormer, with our proposed Transformer decoder replacing the standard one. The key components of our Transformer decoder include a masked attention operator, which extracts localized features by constraining cross-attention to within the foreground region of the predicted mask for each query, instead of attending to the full feature map. To handle small objects, we propose an efficient multi-scale strategy to utilize high-resolution features. It feeds successive feature maps from the pixel decoder's feature pyramid into successive Transformer decoder layers in a round robin fashion. Finally, we incorporate optimization improvements that boost model performance without introducing additional computation.

Model Zoo:

https://github.com/facebookresearch/Mask2Former#model-zoo-and-baselines

Training Dataset:

COCO (Microsoft Common Objects in Context): scored 57.8

Cityscapes: scored 66.6

Comparison on cited papers:

Panopto segmentation CNNs U-Net	Summary	Results (COCO)	Feasibility	Pros	Cons
Mask Dino	(mentioned above in details)	scored 59.4	Very complex	State of the art Uses	N/A
SegFormer	Introduced: An efficient deeply-supervised mask decoder A query decoupling strategy An improved post-processing method	scored 56.2	Most feasible since has collab	Uses DETR that has a collaboratory notebook	Trained only on COCO
Mask2Former	introduced masked attention instead of cross attention	scored 57.8		Trained on COCO and cityscapes	N/A

Here is a comparison between the different features provided by top-ranking models.

Models	Detection	Segmentation	End-to-End	Feature Extraction	Denoising	Box Helps Mask	Mask Helps Box
Mask-RCNN [12]	✓	Instance		RoI pooling	No	✓	
DETR [2]	✓	Panoptic and instance	✓	Standard attention	No	✓	
DINO [37]	✓	No	✓	Deformable attention	Contrastive DN for box		
HTC [3]	✓	Instance		RoI pooling	No	✓	
Mask2former [4]		Panoptic, instance, and semantic	✓	Masked attention	No		
Mask DINO (ours)	✓	Panoptic, instance, and semantic	✓	Deformable attention	DN for both mask and box	✓	✓

Mask DINO: Towards A Unified Transformer-based Framework for Object Detection and Segmentation

Here is the latest performance comparison between the top-ranking models.

Method	Params	Backbone	Backbone Pre-training Dataset	Detection Pre-training Dataset	val w/o TTA	val w/ TTA
Instance segmentation on COCO					AP	
Mask2Former [4]	216M	SwinL	IN-22K-14M	–	50.1	–
Soft Teacher [36]	284M	SwinL	IN-22K-14M	O365	51.9	52.5
SwinV2-G-HTC++ [23]	3.0B	SwinV2-G	IN-22K-ext-70M [23]	O365	53.4	53.7
MasK DINO(Ours)	**223**M	SwinL	IN-22K-14M	O365	**54.5**(+1.1)	–
Panoptic segmentation on COCO					PQ	
Panoptic SegFormer [19]	–M	SwinL	IN-22K-14M	–	55.8	–
Mask2Former [4]	216M	SwinL	IN-22K-14M	–	57.8	–
MasK DINO (ours)	223M	SwinL	IN-22K-14M	O365	**59.4**(+1.6)	–
Semantic segmentation on ADE20K					mIoU	
Mask2Former [4]	215M	SwinL	IN-22K-14M	–	56.1	57.3
Mask2Former [4]	217M	SwinL-FaPN	IN-22K-14M	–	56.4	57.7
SeMask-L MSFaPN-Mask2Former [14]	–M	SwinL-FaPN	IN-22K-14M	–	–	58.2
SwinV2-G-UperNet [23]	3.0B	SwinV2-G	IN-22K-ext-70M [23]	–	59.3	59.9
MasK DINO (ours)	**223M**	SwinL	IN-22K-14M	O365	**59.5**	**60.8**(+0.9)

Mask DINO: Towards A Unified Transformer-based Framework for Object Detection and on

Chosen Model

We choose the DETR (DE:TR: End-to-End Object Detection with Transformers) which is a Facebook project that was used by the SegFormer project as described above. The model uses CNNs as the fundamental deep neural network, which will be a great opportunity to apply one main model covered in the course. The DETR model has a very well-documented git repository, with available zoo models and weights. Moreover, the repo provides a set of google collab notebooks for implementing different segmentation projects, which are going to be very useful as a base model. Specifically, it provides a Panoptic segmentation notebook referenced below.

Model Source:

Github: GitHub - facebookresearch/detr: End-to-End Object Detection with Transformers

Google collab : Panoptic segmentation using DETR

Previous projects in machine learning

- Loan default prediction project in a machine learning course.

Proposed Changes:

- Try different hyperparameters (batch sizes, learning rate, loss functions, etc.)
- Generate image crops from the datasets and train on these image crops, and predict on full images.
- Create an ensemble model that utilizes the different models discussed above.

Per instructor's feedback:

- Try different CNN architecture models.
    - We can do the ensemble model here by combining multiple CNNs.
- We can also try to build a hierarchical model that splits the larger tasks of panoptic segmentation into two smaller models (Semantic and instance), then combine their outputs
    - The dataset provides ground truths for both semantic and instance segmentation.
    - The Segformer model took a simpler approach by what they call coupling the queries, maybe we can look further into it.
- We can also do data augmentation by creating images with different images with different saturation, contrast, etc., however we already have a relatively large dataset, so this might be our last option.

Expected Benefits:

By doing these changes or a subset of them we may get better results for the PQ index from the SegFormer model. However, if we got worse results, we will be able to document our approach; stating why our modifications did no inhancement . This can be very beneficial for further researches to try other promising paths.

Contribution:

Youssef Ashraf Kandil:

- Researching for the idea,
- Introduction
- Evaluation metrics
- Literature review
- Papers comparison
- Choosing model
- Proposed changes
- Updating Proposal

Andrew Sinout Shenouda:

- Researching for the idea,
- problem statement
- goals of Panoptic Segmentation
- getting data sets
- Literature review
- Choosing model
- Proposed changes
- Updating Proposal

References:

Panoptic Segmentation | Papers With Code

Semantic vs. Instance vs. Panoptic Segmentation - PyImageSearch

Metrics to Evaluate your Semantic Segmentation Model | by Ekin Tiu | Towards Data Science

https://arxiv.org/pdf/2206.02777v1.pdf

COCO Dataset | Papers With Code

Cityscapes Dataset | Papers With Code

KITTI Dataset | Papers With Code

https://arxiv.org/pdf/2109.03814v4.pdf

https://arxiv.org/pdf/2112.01527v3.pdf

GitHub - facebookresearch/detr: End-to-End Object Detection with Transformers

Panoptic segmentation using DETR

Image Segmentation: Tips and Tricks from 39 Kaggle Competitions | Neptune Blog