# Per-Pixel Classification is Not All You Need for Semantic Segmentation

sgc

April 28, 2022

## 1 Introduction

The authors think that a single mask classification can be used in both semantic and instance segmentation. So they proposed MaskFormer approach that seamlessly converts any existing per-pixel classification model into a mask classification. In MaskFormer, a transformer decoder is employed to compute a set of pairs, each consisting of a class prediction and a mask embedding vector. The mask embedding vector is used to get the binary mask prediction via a dot product with the per-pixel embedding obtained from an underlying fully-convolutional network.
Authors evaluated their model in couple of datasets and achieved great performance.

## 2 Method

First the author introduced per-pixel and mask strategy briefly. The mask formulation: First partition N regions (represented by masks) and then label each region.

### 2.1 MaskFormer

The MaskFormer consists of 3 parts: 1) a pixel-level model that generates binary mask predictions; 2) a stack of transformer decoders that generates N per-segment embeddings; 3) a segmentation module generate predictions.

**Pixel-level module** First a backbone generates a low-resolution feature map(here one of the dimensions represents the channel), and then a decoder upsample the feature to full image resolution (One dimension represents the embedding size)

**Transformer module** A transformer compute its output from the feature and its positional embedding.

**Segmentation module** Segmentation module applies a linear classifier, followed by a softmax activation, on top of the per-segment embeddings Q to yield class probability predictions $p_i \in \Delta_{K+1}{}_{i=1}^{N}$ for each segment.

## 2.2 Mask-classification inference

**General inference** General inference partitions an image into segments by assigning each pixel $[h, w]$ to one of the N predicted probability-mask pairs via $argmax_{i:c_i \neq \emptyset} p_i(c_i) m_i[h, w]$. Here $c_i$ is the most likely class label $c_i = argmax_{c \in 1,\ldots,K,\emptyset} p_i(c)$ for each probability-mask pair i. Intuitively, this procedure assigns a pixel at location [h, w] to probability-mask pair i only if both the most likely class probability $p_i(c_i)$ and the mask prediction probability $m_i[h, w]$ are high.

**Semantic interference** Just simply multiply works pretty well. $argmax_{i:c_i \in 1,2,3,\ldots,K} p_i(c_i) m_i[h, w]$ notably "no object" label is not included as each output pixel requires a label for semantic segmentation tasks.

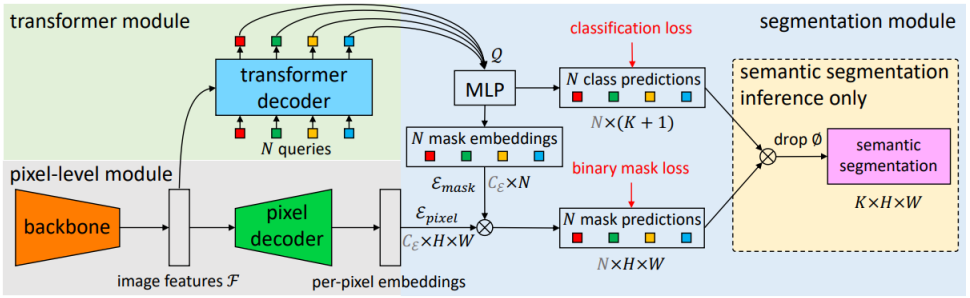


Figure 2: **MaskFormer overview.** We use a backbone to extract image features $\mathcal{F}$. A pixel decoder gradually upsamples image features to extract per-pixel embeddings $\mathcal{E}_{\text{pixel}}$. A transformer decoder attends to image features and produces $N$ per-segment embeddings $\mathcal{Q}$. The embeddings independently generate $N$ class predictions with $N$ corresponding mask embeddings $\mathcal{E}_{\text{mask}}$. Then, the model predicts $N$ possibly overlapping binary mask predictions via a dot product between pixel embeddings $\mathcal{E}_{\text{pixel}}$ and mask embeddings $\mathcal{E}_{\text{mask}}$ followed by a sigmoid activation. For semantic segmentation task we can get the final prediction by combining $N$ binary masks with their class predictions using a simple matrix multiplication (see Section 3.4). Note, the dimensions for multiplication $\otimes$ are shown in gray.

# 3 Results

This method is compatible for any backbones.
In semantic segmentation, MaskFormer outperforms the best per-pixel classification-based models while having fewer parameters and faster inference time. This result suggests that the mask classification formulation has significant potential for semantic segmentation.
In panoptic segmentation, MaskFormer actually outperformed some recent(2021) models.

# 4 Ablation study

1, number of queries, 100 performs the best and 20 already outperforms the baseline.
2, number of transformer decoder layers. Interestingly, a single layer of decoder already outperforms a 6-layer-decoder PerPixelBaseline+. For panoptic segmentation, however, multiple decoder layers are required to achieve competitive performance.
3, Per-pixel vs. mask classification. the shift from per-pixel to mask classification is the key of the improvement. Bipartite matching is not only more flexible (can make less prediction than total class count) but also gives better results.

# 5 Summary

MaskFormer is the first model that unifies semantic- and instance-level segmentation with the exact same model, loss, and training pipeline.

# References