Recommended Papers

Masked-attention Mask Transformer for Universal Image Segmentation

Introduction

Image segmentation groups pixels with different semantics, where each choice of semantics defines a task. While only the semantics of each task differ, current research focuses on designing specialised architectures for each task.

We present ***Masked-attention Mask Transformer****, a new architecture capable of addressing any image segmentation task, like panoptic, instance or semantic. Its key components include* masked attention*, which extracts localised features by constraining cross-attention within predicted mask regions. In addition to reducing the research effort by at least three times, it outperforms the best specialised architectures by a significant margin on four popular datasets.*

Challenge

- Per-pixel classification architectures based on Fully Convolutional Networks (FCNs) are used for semantic segmentation.
- Mask classification architectures that predict a set of binary masks each associated with a single category, dominate instance-level segmentation.

Although such *specialised* architectures have advanced each individual task, they lack the flexibility to generalise to the other tasks. Thus, duplicate research and hardware optimisation effort is spent on each specialised architecture for every task.

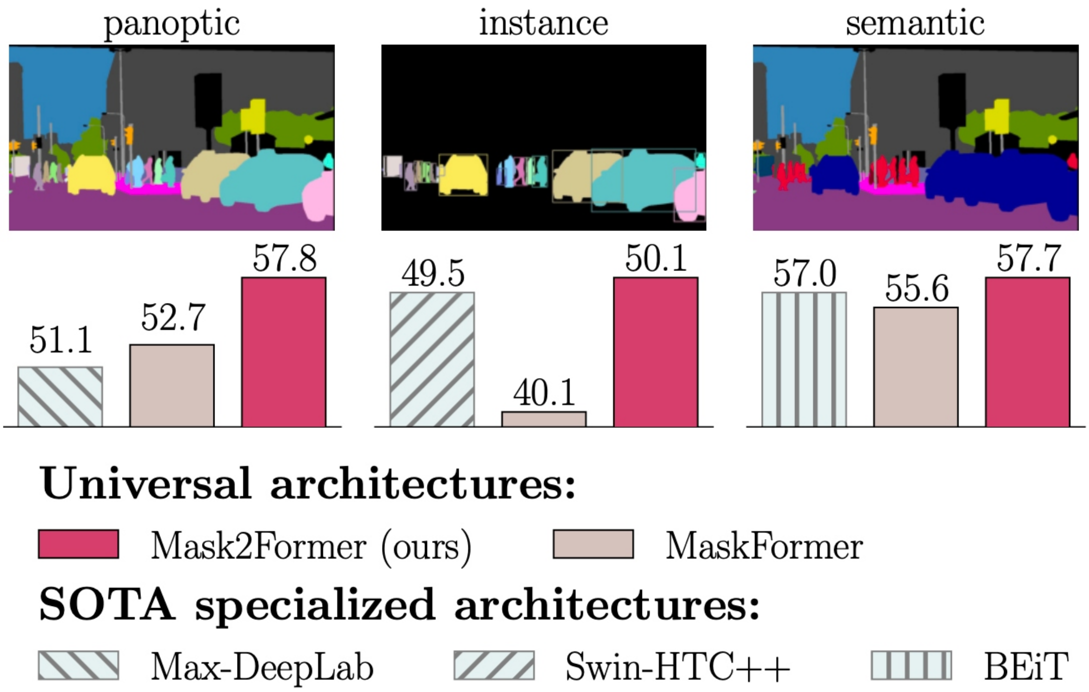
To address this fragmentation, recent work has attempted to design *universal architectures*, that are capable of addressing all segmentation tasks with the same architecture. These architectures are typically based on an end-to-end set prediction objective, and successfully tackle multiple tasks without modifying the architecture, loss, or the training procedure.

Challenge

Note, that universal architectures are still trained separately for different tasks and datasets, albeit having the same architecture. In addition to being flexible, universal architectures have recently shown state-of-the-art results on semantic and panoptic segmentation.

However, recent work still focuses on advancing specialised architectures , which raises the question: why haven't universal architectures replaced specialised ones?

Although existing universal architectures are flexible enough to tackle any segmentation task, as shown below, in practice their performance lags behind the best specialised architectures.



Challenge

Beyond the inferior performance, universal architectures are also harder to train. They typically require more advanced hardware and a much longer training schedule. For example, training Mask- Former takes 300 epochs to reach 40.1 AP and it can only fit a single image in a GPU with 32G memory. In contrast, the specialised Swin-HTC++ obtains better performance in only 72 epochs. Both the performance and training efficiency issues hamper the deployment of universal architectures.

In order to solve all the challenges above and achieve a great performance in all kinds of segmentations we created the ***Masked-attention Mask Transformer.***

- First, we use *masked attention* in the Transformer decoder which restricts the attention to localised features centered around predicted segments, which can be either objects or regions depending on the specific semantic for grouping. Compared to the cross-attention used in a standard Transformer decoder which attends to all locations in an image, our masked attention leads to faster convergence and improved performance.
- Second, we use *multi-scale high-resolution features* which help the model to segment small objects/regions.
- Third, we propose *optimisation improvements* such as switching the order of self and cross-attention, making query features learnable, and removing dropout; all of which improve performance without additional compute.
- Finally, we save 3×training memory without affecting the performance by *calculating mask loss on few randomly sampled points*.

These improvements not only boost the model performance, but also make training significantly easier, making universal architectures more accessible to users with limited compute.

We evaluate Mask2Former on three image segmentation tasks of Panoptic, Instance, Semantic Segmentation using four popular datasets: COCO, Cityscapes, ADE20K and Mapillary Vistas

Types Of Segmentation Architecture

Specialised semantic segmentation architectures typically treat the task as a per-pixel classification problem. FCN-based architectures independently predict a category label for every pixel. Follow-up methods find context to play an important role for precise per-pixel classification and focus on designing customised context modules or self-attention variants.

Specialised instance segmentation architectures are typically based upon mask classification. They predict a set of binary masks each associated with a single class label. The pioneering work, Mask R-CNN, generates masks from detected bounding boxes. Although the performance has been advanced in each task, these specialised innovations lack the flexibility to generalise from one to the other, leading to duplicated research effort.

Panoptic segmentation has been proposed to unify both semantic and instance segmentation tasks. We find panoptic architectures usually only report performance on a single panoptic segmentation task, which does not guarantee good performance on other tasks, as we can see on the figure above. For example, panoptic segmentation does not measure architectures abilities to rank predictions as instance segmentations. Thus, we refrain from referring to architectures that are only evaluated for panoptic segmentation as universal architectures.

Universal architectures have emerged with DETR (Detection With Transformers) and show that mask classification architectures with an end-to-end set prediction objective are general enough for any image segmentation task. MaskFormer shows that mask classification based on DETR not only performs well on panoptic segmentation but also achieves state-of-the-art on semantic segmentation. Unfortunately, these architectures fail to replace specialised models as their performance on tasks or datasets is still worse than the best specialised architecture with the main challenge being instance segmentation.

Masked-attention Mask Transformer

Mask classification architectures group pixels into N segments by predicting N binary masks, along with N corresponding category labels. Mask classification is sufficiently general to address any segmentation task by assigning different semantics.

Challenge

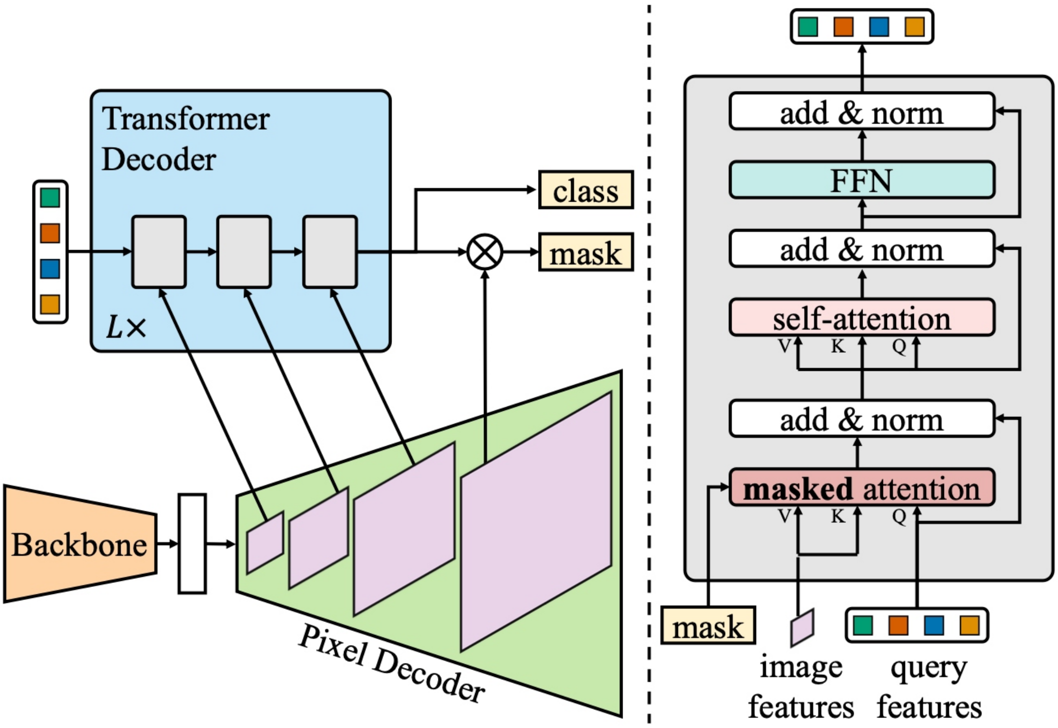
However, the challenge is to find good representations for each segment. For example, Mask R-CNN uses bounding boxes as the representation which limits its application to semantic segmentation.

Inspired by Detection with Transformers, each segment in an image can be represented as a **C-dimensional feature vector / object query** and can be processed by a Transformer decoder, trained with a set prediction objective.

A simple meta architecture would consist of three components.

1. A *backbone* that extracts low-resolution features from an image.
2. A *pixel decoder* that gradually upsamples low-resolution features from the output of the backbone to generate high-resolution per-pixel embeddings.
3. A *Transformer Decoder* that operates on image features to process object queries. The final binary mask predictions are decoded from per-pixel embeddings with object queries.

Mask2Former adopts the aforementioned meta architecture, with our proposed Transformer decoder replacing the standard one. The key components of our Transformer decoder include a *masked attention* operator, which extracts localised 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 utilise 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 optimisation improvements that boost model performance without introducing additional computation.

Masked attention

Challenge

Context features have been shown to be important for image segmentation. However, recent studies suggest that the slow convergence of Transformer-based models is due to global context in the cross-attention layer, as it takes many training epochs for cross-attention to learn to attend to localised object regions.

We hypothesise that local features are enough to update query features and context information can be gathered through self-attention. 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.

High-resolution features

Challenge

High-resolution features improve model performance, especially for small objects. However, this is computationally demanding.

We propose an efficient multi-scale strategy to introduce high-resolution features while controlling the increase in computation. Instead of always using the high-resolution feature map, we utilise a feature pyramid which consists of both low- and high-resolution features and feed one resolution of the multi-scale feature to one Transformer decoder layer at a time.

Specifically, we use the feature pyramid produced by the *pixel decoder* with resolution 1/32, 1/16 and 1/8 of the original image. We repeat this 3-layer Transformer decoder L times. Our final Transformer decoder hence has 3L layers.

More specifically, the first three layers receive a feature map of resolution:

1. height = [H1 = H/32, H2 = H/16, H3 = H/8]
2. width = [W1 = W/32, W2 = W/16, W3 = W/8]
3.

Optimisation improvements

A standard Transformer decoder layer consists of three modules to process query features in the following order:

1. self-attention module
2. cross-attention module
3. feed-forward network (FFN).

Moreover, query features are *zero initialised* before being fed into the Transformer decoder and are associated with *learnable* positional embeddings. Furthermore, dropout is applied to both residual connections and attention maps.

To optimise the Transformer decoder design, we make the following three improvements.

1. First, we switch the order of self- and cross-attention to our new masked attention to make computation more effective: query features to the first self-attention layer are image-independent and do not have signals from the image, thus applying self-attention is unlikely to enrich information.
2. Second, we make query features Xo learnable as well and learnable query features are directly supervised before being used in the Transformer decoder to predict masks. We find these learnable query features function like a region proposal network and have the ability to generate mask proposals.
3. Finally, we find dropout is not necessary and usually decreases performance. We thus completely remove dropout in our decoder.

Semantic Representations with Attention Networks for Boosting Image Captioning

Introduction

A semantic attention network is proposed to incorporate general-purpose knowledge into a transformer attention block model. This design combines visual and semantic properties of internal image knowledge in one place for fusion, serving as a reference point to aid in the learning of alignments between vision and language and to improve visual attention and semantic association. The proposed framework is validated on the COCO dataset, and experimental results demonstrate competitive performance against the current state of the art.

Image captioning techniques, which automatically create a natural language description from an image, are an important aspect of multimedia content analysis. They have attracted much attention since they offer insight into the relationships between the multi-modal mapping of vision and natural language tasks. Image captioning, which aims to describe the image in continuous natural language, also has a variety of practical applications.

Existing approaches to image captioning have evolved an encoder-decoder structure, based on the sequence-to-sequence paradigm for machine translation. Typically, Convolutional Neural Networks are utilised as encoders, converting image data into usable visual features. Alongside this, Recurrent Neural Networks are typically utilised as decoders for generating a language description. The standard encoding and decoding structure works on the input image to provide a related description of the scene, objects, and their relationships. The majority of current methods investigate mapping relationships between words in a sentence and specific regions of an image.

Challenge

- The main challenge faced by researchers in the application of vision-and-language models is **data-related**, since the majority of image captioning models are trained on a large amount of paired image and caption data, but these datasets typically have only a few ground truth captions per image, which are insufficient to provide a clear description of the contents of each image. It’s common knowledge in this area of research that not all captions hold equivalent importance in describing the contents of a given image.
- Another limitation is that many image captioning models use just the visual characteristics of an image to direct the encoder, while the decoder typically relies on the textual information from the training set – this can result in difficulties accurately identifying objects in a given image.
- When several objects are present in an image, the described structure may not be able to identify all objects present and may especially struggle to identify any relationships between objects.

These weaknesses can result in the model missing tiny objects, providing incorrect object relationships, or producing incorrect text representations, which go on to affect the quality of the resulting caption. This shows that it may be beneficial to introduce more knowledge sources to the network during training in an aims to increase contextual understanding and further caption generation accuracy.

In terms of improving visual comprehension, an **external knowledge network** is introduced, which aids in the generation of more flexible description sentences by utilising information other than the basic content of the image provided. In this manner, the proposed model can gather data from external sources other than the image’s basic content to use in creating more flexible description sentences. We believe that **embedding** this type of knowledge in a model is a necessary step to enable progress on complex multi-modal image captioning problems.

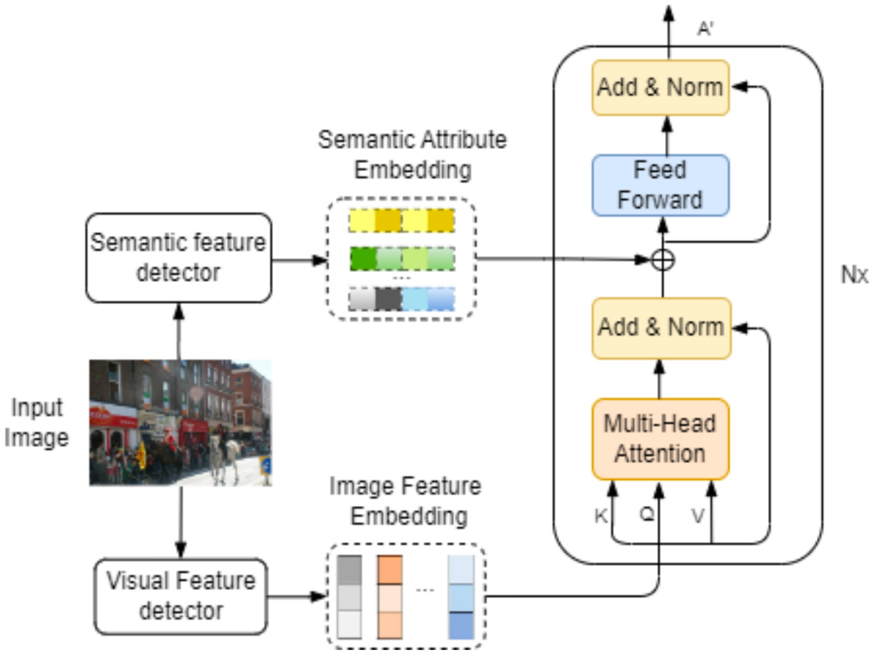
Our contributions are summarised as follows:

- Semantic-guided attention network based on the Transformer model.

- Use of semantic features of the image’s main elements to link and guide visual features, such as spatial relationships between objects, so that the information in the image is more highly integrated.
- Use of auxiliary knowledge base source to increase our model’s reasoning capabilities. This involves gathering information from outside of an image’s basic content and allows for the creation of more appropriate image descriptions.

DESIGN OF THE PROPOSED FRAMEWORK

Semantic guided-attention networks can adaptively perform the image encoder procedure to describe a given image. The figure below depicts the proposed framework for image captioning. First, an object detection model is used, i.e., Faster R-CNN, to extract the feature of the original image. Then, an attention module is needed to encode the visual features and output an attentive feature. Following this, common sense embedding features are extracted from external KBs ConceptNet and share the common feature space with other input image information in the encoder transformer to depict relationships between different objects and scenes in image. Finally, a language decoder is applied to generate language descriptions.



A. OBJECT DETECTION

Following, a Faster R-CNN in conjunction with ResNet-101 is adopted, which has a CNN base, for object detection and feature extraction. The Faster R-CNN model is pre-trained on the Visual Genome dataset and outputs object classes. Its first stage is a Region Proposal Network (RPN) that uses intermediate feature maps from ResNet-101 as inputs and generates bounding boxes for proposed objects. Intersection-Over-Union (IOU) is metric that measures the overlap between two bounding boxes. The reference boxes n that have an IoU more than 0.7 are selected. In the second stage, Region-Of-Interest (ROI) pooling layer is used to convert all proposal bounding boxes to the same spatial size feature map (e.g., 14 x 14 x 2048). For
simplicity, the top 36 ROIs are only used. These are followed by a softmax distribution to predict the bounding box object classes and refinements for each box proposal.

B. IMAGE ENCODER

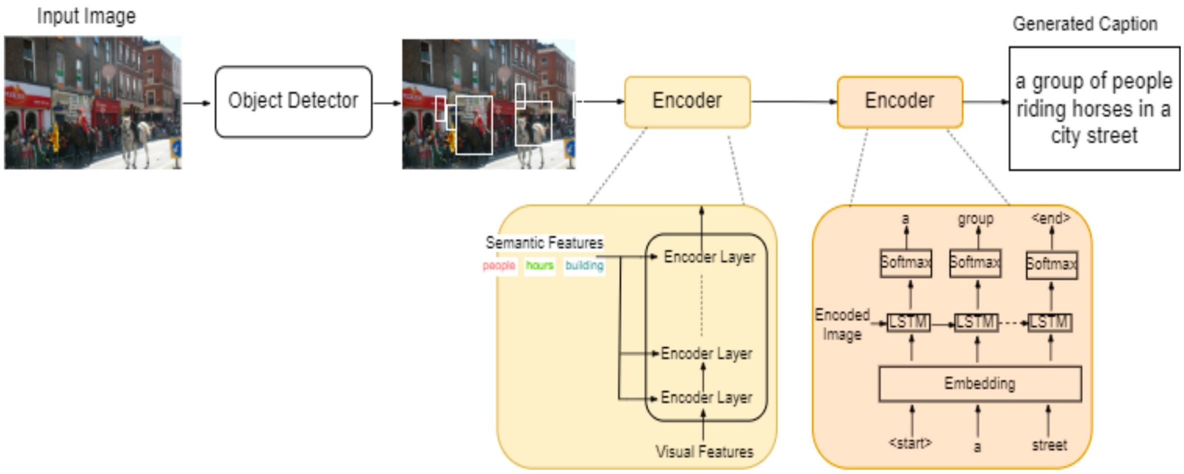
To improve the image understanding capability of the image encoder, we construct the merged box as a semantic relationship guide to direct attention and enhance visual feature representation. Based on the detection results of Faster R-CNN, self-attention layers contain two kinds of inputs:

1. Object’s visual features Vi of previously detected objects
2. Those objects semantic classes Si.

First, the visual object’s attention is founded. This is done by employing layers of MHA.
Secondly, the ConceptNet knowledge base is used to generate objects with semantic concepts embedded Ese(S). Thus, semantic concepts S = {s1, s2, ..., sn} are embedded to semantic concepts features O = {o1, o2, ..., on}, where oi ∈ Rd2, and d2 = 300. Notably, towards a specific object visual features may vary from object to object while semantic features always remain unchanged. The merged box combines the visual attention features Attention(Q,K,V) and semantic vectors to get visual semantic representation for input image.

C. LANGUAGE DECODER

In this paper, the LSTM units are selected as the decoder. The decoding component is used to decode visual features to iteratively generate descriptive text sequences. The improvement of the decoder mainly focuses on enriching the information in both the visual and text. As shown in Figure 2, we jointly integrate both the attention features from encoder output A′ and the word embedding vectors Wt in one fusion representation into the LSTM unit. The LSTM network has a state cell and several gates, such as a forget gate and an input gate. These gates ensure the effective memory and updating of information at each time step t ∈ [1,T], and generate output word yt.


Input Image
Object Detector
Encoder
Encoder
Generated Caption
a group of people riding horses in a city street
Semantic Features
people
hours
building
Encoder Layer
Encoder Layer
Encoder Layer
Visual Features
a
group
<end>
Softmax
Softmax
Softmax
Encoded Image
LSTM
LSTM
LSTM
Embedding
<start>
a
street