УДК 004.932.2

OVERVIEW OF THE ONEFORMER NEURAL NETWORK MODEL APPLICABLE TO THE PROBLEM OF THE INSTANCE SEGMENTATION IN AN IMAGE

БОБРОВА НАТАЛЬЯ ЛЕОНИДОВНА,
к.т.н., доцент
ХАРКЕВИЧ АНТОН ПАВЛОВИЧ,
СТЕЦКО ВАДИМ ЮРЬЕВИЧ
магистранты
УО «Белорусский государственный университет информатики и радиоэлектроники»

Научный руководитель: Шевалдышева Елена Зигфридовна
к.ф.н., доцент
УО «Белорусский государственный университет информатики и радиоэлектроники»

Аннотация. В данной работе рассмотрена нейросетевая модель OneFormer, подходящая для осуществления семантической сегментации, сегментации экземпляров и паноптической сегментации изображения, рассмотрена архитектура модели OneFormer, ключевые компоненты архитектуры, обеспечивающие универсальность модели, указаны недостатки модели.
Ключевые слова. OneFormer, сегментация экземпляров, семантическая сегментация, паноптическая сегментация, Qtext, Qtask.

ОБЗОР НЕЙРОСЕТЕВОЙ МОДЕЛИ ONEFORMER, ПРИМЕНИМОЙ К ЗАДАЧЕ СЕГМЕНТАЦИИ ЭКЗЕМПЛЯРОВ НА ИЗОБРАЖЕНИИ

Bobrova Natalya Leonidovna,
Kharkevich Anton Pavlovich,
Stetsko Vadim Yurievich

Scientific adviser: Shevaldysheva Elena Zigfrididovna

Abstract. This paper reviews the neural network model OneFormer suitable for semantic segmentation, instance segmentation and panoptic segmentation in an image, reviews the architecture of OneFormer model, its key components that ensure the versatility of the model, points out the limitations of the model.
Keywords. OneFormer, instance segmentation, semantic segmentation, panoptic segmentation, Qtext, Qtask.

This paper reviews the OneFormer model. This model was released in late 2022. It is one of the most popular models applicable to the task of instance segmentation in an image. The key feature of the model is its versatility in performing semantic segmentation, instance segmentation and panoptic segmentation tasks. The OneFormer model uses a single unified framework that covers all aspects of image segmentation. It is trained

for all three tasks only once, significantly reducing the time and complexity to achieve results.

The general idea of OneFormer architecture, its versatility and compactness compared to other architectures are shown in the figure below:

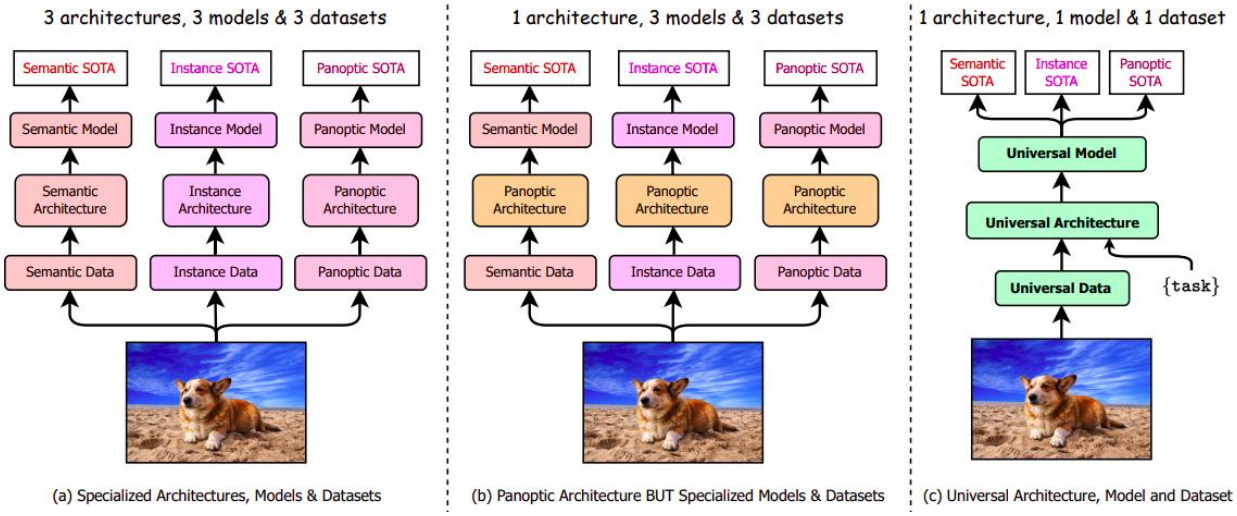

Fig. 1. General idea of OneFormer model architecture

During the training process, the OneFormer architecture dynamically adapts the model's behavior based on the current segmentation task. For this purpose, information in the format «task - {task}» is included in the input data.

The input task is randomly selected from a set of available tasks that includes panoptic segmentation, semantic segmentation, and instance segmentation for each training image. This random selection allows the model to learn and adapt to different types of segmentation tasks during training.

A one-dimensional task token is generated from the input data. OneFormer thus ensures that the model has knowledge of the task it needs to perform and can adjust its behaviour accordingly.

In addition, input data in the format «task - {task}» influences the creation of a text list that represents the number of binary masks for each class in the source label. This text list is mapped to text query representations, providing task-specific information that helps improve model predictions and segmentation.

OneFormer provides task-specific training and enables the development of a unified segmentation system because the model depends on «task - {task}» inputs and takes into account specific information.

OneFormer is also designed to accept user query sets as inputs. Query set representations facilitate communication and interaction between different components in the model architecture. These representations are used in the transducer decoder, where they are responsible for collecting and integrating information from both the input image and the task-specific context.

During training, the OneFormer model uses two sets of queries: text queries (Qtext) and object queries (Q). Qtext is a textual representation of segments in an image, while Q is an image-based representation.

To obtain Qtext, text entries are tokenised and passed through a text encoder, which consists of a 6-layer transducer. This encoding process generates a list (Ntext) of text attachments that capture information about the number of binary masks and their corresponding classes present in the input image.

The set of trained textual contextual embeddings (Qctx) is then combined with the encoded textual embeddings, resulting in N text queries (Qtext).

The architecture of the text transducer is shown in the figure below.

Object queries (Q') are initialised as a repetition of the token task (Qtask) N - 1 times to retrieve an object Q. The Q' are then updated using instructions from features in a two-layer transformer. The updated object queries are merged with Qtask, resulting in a task-conditioned representation of N queries, denoted as Q.

This initialisation and concatenation step, as opposed to random initialisation, is crucial to effectively train the model for multiple segmentation tasks.

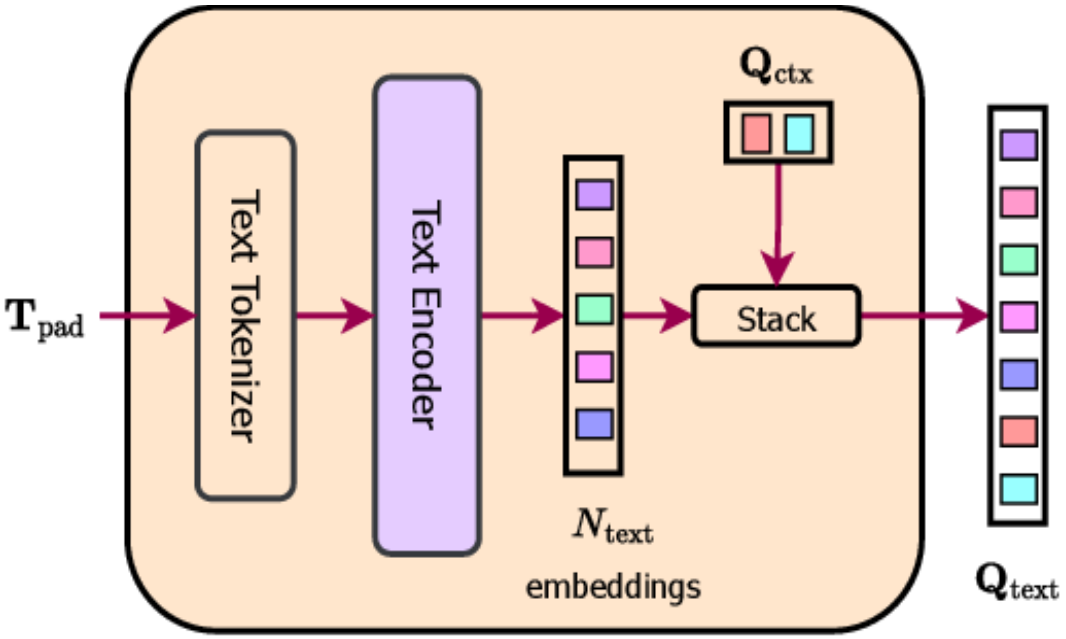

Fig. 2. The architecture of the text transducer

The overview architecture of the OneFormer model is shown in the figure below:

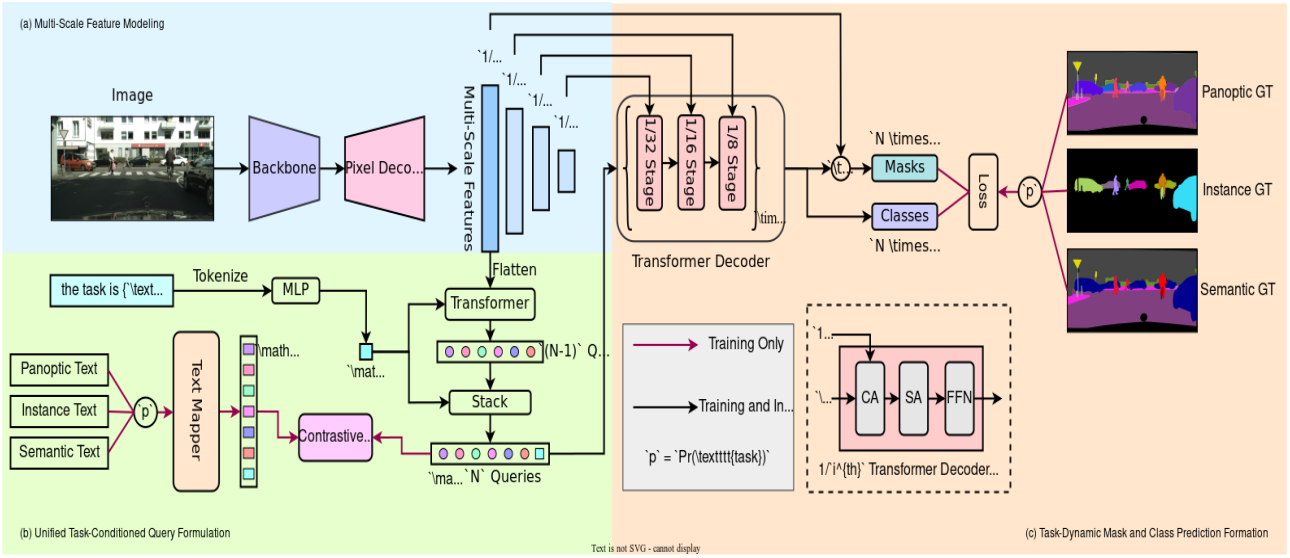

Fig. 3. The architecture of the OneFormer model

The OneFormer model shows high accuracy which has been tested on three datasets (CityScapes Dataset, ADE20K Dataset, COCO Dataset).

An example of image segmentation using the OneFormer model showing higher accuracy compared to the Mask2Former model is shown in the figure below:

Despite its versatility, the OneFormer has some disadvantages. These include:

1) The model has high computational requirements and may need a lot of resources for training and deployment.

2) The model is sensitive to the dataset used as its performance depends on it. Additional tuning is also needed for specific datasets.

3) Training OneFormer can be time consuming due to the complexity of the model architecture.

4) The large size of the OneFormer model due to the high number of multitasking components can affect memory usage and limit the ability to deploy in resource-constrained environments.

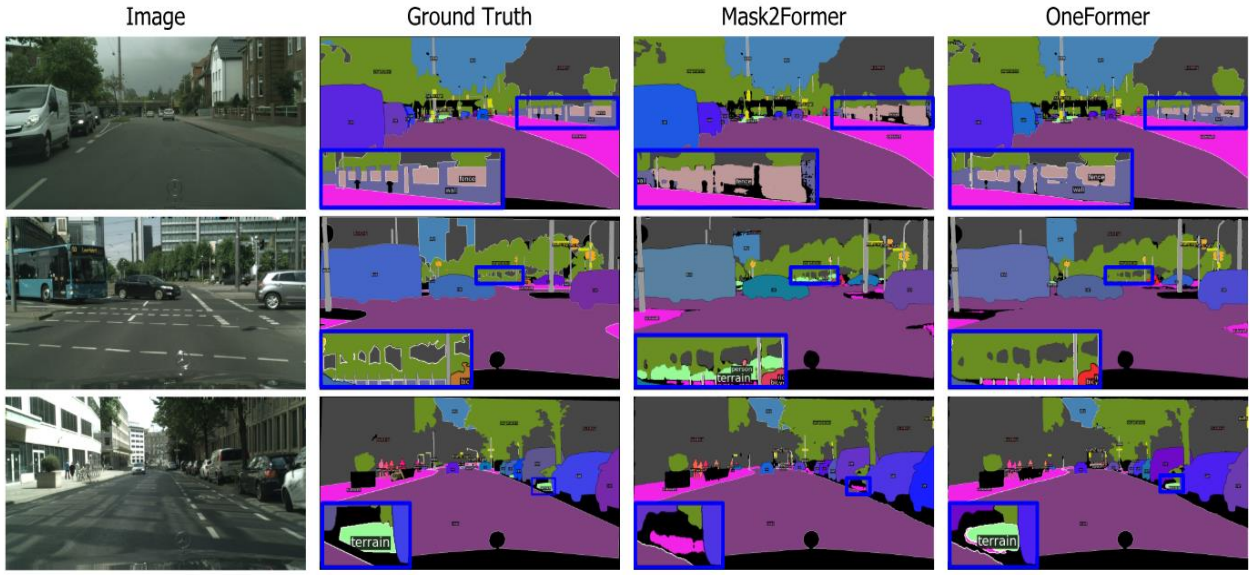

Fig. 4. The example of image segmentation using Mask2Former and OneFormer models

Thus, the OneFormer model is a universal model that uses a unified structure for all image segmentation tasks. The model achieves high performance for image segmentation tasks in various domains.

OneFormer's compact architectural design provides robustness and adaptability. At the same time, the ability to include task-specific information through user query views improves model understanding.

The model shows high accuracy and efficiency on benchmark datasets.

Although the OneFormer model has limitations, it represents a significant step forward in the field of image segmentation.

References

1. Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, Humphrey Shi “OneFormer: One Transformer To Rule Universal Image Segmentation”//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2989-2998

2. Deny Wiria Nugraha, Amil Ahmad Ilham, Andani Achmad, Ardiaty Arief “Transformers for aerial images semantic segmentation of natural disaster-impacted areas in natural disaster assessment”//Bulletin of Electrical Engineering and Informatics, Vol. 14, No. 2, April2025, pp. 1391-1406

3. Maxim Kolodiazhnyi, Anna Vorontsova, Anton Konushin, Danila Rukhovich “OneFormer3D: One Transformer for Unified Point Cloud Segmentation”//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 20943-20953

4. Qu, Y.; Kim, J. “Enhancing Query Formulation for Universal Image Segmentation”//Sensors 2024, 24, 1879