Multi Modal Learning using AI

A Capstone Project Report submitted in partial fulfillment of the requirements for the degree of B.Tech in Electrical Engineering

by

Hiya Kwatra
(Entry Number - 2020EEB1173)

Under the guidance of
Dr. Santosh Kumar Vipparthi

DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY ROPAR

2024

Declaration

I hereby dec1are that this written submission represent ts my ideas in my own of words and where others' words or ideas have been inc1uded, I have adequate1y cited and referenced the origina1 sources. I also dec1are that I have adhered to a11 the princip1es of academic-honesty and integrity, and also haven't mis-represented, fabricated or fa1sified any of the ideas/data/facts/sources in my submission. I also understand that any vio1ation of the above stated will be the cause for discip1inary action by the 1nstitute and can evoke pena1 action from sources which haven't been adequately cited or from whom proper permission has not been taken when needed.

Name : Hiya Kwatra
Entry No : 2020EEB1173

Date: 06/05/2024

Abstract

We here have presented our study on several models utilized for multi-modal AI studies in this report. We will deep dive into their designs, methods, and outcomes. This research aimed to develop a unified framework that can efficiently perform semantic segmentation on any given input, be it natural images, medical scans, satellite imagery, or other data types. Traditionally, segmentation models are designed and trained for a specific domain or data modality only, so the researchers sought to build a model that could segment unseen data types without any domain-specific tuning or fine-tuning.

Acknowledgement

I would take this opportunity to thank our course coordinator, Dr. Mahendra Sakare, and our supervisor in this course, Dr. Santosh Kumar Vipartthi, for their invaluable support and guidance throughout this project.

Table of Contents

List of Figures

1. Introduction

1.1 Referred object Video/Image Segmentation

Traditionally in video object segmentation (VOS) tasks, the process involved a user manually providing pixel-level annotations for an object of interest in a reference frame of a video. The segmentation system was then responsible for generating binary masks that accurately delineated that same object across all other frames where it appeared, based on the user-provided annotations in the initial frame. The idea is to improve computer-human inter action by allowing the 1inguistic expressions as initial ization cues instead of inter‘active segmentations in a detai1ed binary mask, bounding box, point or scribble. The particularities of Refer red Expressions for videos were initia11y addressed by Khorev‘a et a1 , building a dataset of RefExps dividing into two classes: REs for the first frame of a video ; REs for the fu1l clip.

Video object segmentation is achieved using annotations manually added to a frame of video, which are procured by the pixel wise binary mas|< for all of the video's frames where it has visiblity, which is generated by the segmentation system. A class of multimodal machine learning problems known as "Referred object segmentation" attempts to combine two modalities—text and image/video (series of images)—in order to give additional context for the objects or entities that should be identified in the given picture frame. Prior works have looked at using referring expressions (REs) as input for tasks involving language-guided video object segmentation (LVOS). However, referring expressions from existing LVOS benchmark datasets can sometimes be challenging to interpret and analyze directly. To address this, a recent study proposed categorizing referring expressions into seven semantic categories. This semantic categorization scheme, referred to as RefVOS, provided a framework to better understand the nature and properties of referring expressions typically used as inputs for LVOS tasks.

1.1.1 Previous Work - RefVOS

RefVOS is a competitive novel model for a 1anguage-guided system of image segment ation and state-of-the-art VOS which is 1anguage-guided . A2D is extended with new Referred Expressions with diverse seman+ic categories for non-trivia1-cases, revealing that it struggles to exploit static and motion even+s and mainly benefiting from REs based on appearance and 1ocation from the tests.

The given model aims to individually produce embeddings for the image (visual features) and text separately. It is achieved using the following:-

1. **Visual Encoder**
   The images are encoded using DeepLabv3, a semantic segmentation network that relies solely on atrous convolutions. DeepLabv3 uses ResNet101 as its backb0ne and has an output s+ride of 8 pixels. It employs Atrous Spatia1 Pyramid Pooling (ASPP), which consists of parallel atrous convolutional layers with different sampling rates of 12, 24 and 36. ASPP utilizes atrous convolutions to capture multi-scale contextual information without increasing the number of parameters or the computational cost.

2. **Language Encoder**
   BERT, which is a bi-directional transformer model, is used as a language encoder to obtain embeddings for input linguistic phrases. The BERT model is used which is fine-tuned with/to the REs (Referred expressions) of RefCOCO(dataset with annotations) with the MLM i.e. masked language modelling with one epoch for a loss, which is tokenized to the L-guided image-segmentation act, adding the [CLS] a nd [SEP] tokens at the end and begin of the linguistic phrase , henceforth to produce an embedding of 768-dimensions.

These features are then combined to get a multimodal feature space. Its performance is increased when i+ is initially trained on RefCOCO and later finetuned on A2D in b0th of its branches, visua1 and 1anguage.

The encoded language phrase is firstly transformed via a linear projection into a 256-dimensional embedding, and it is then multiplied element-wise by the visual features that were collected from the ASPP from DeepLabv3 to extract Multi-modal Embedding. Standard cross entropy 1oss is used for the segmentation task. The GloVe embeddings are averaged for each token and concatenated with the mean forward and backward pass embeddings. The two configurations that use BERT are compared to this baseline. The second step consists as to first train the BERT with the masked 1anguage mode11ing loss with the REs from RefCOCO, and later fine-tune it on the 1anguage-guided image segmentati0n task, proving the configuration with an an additional gain.

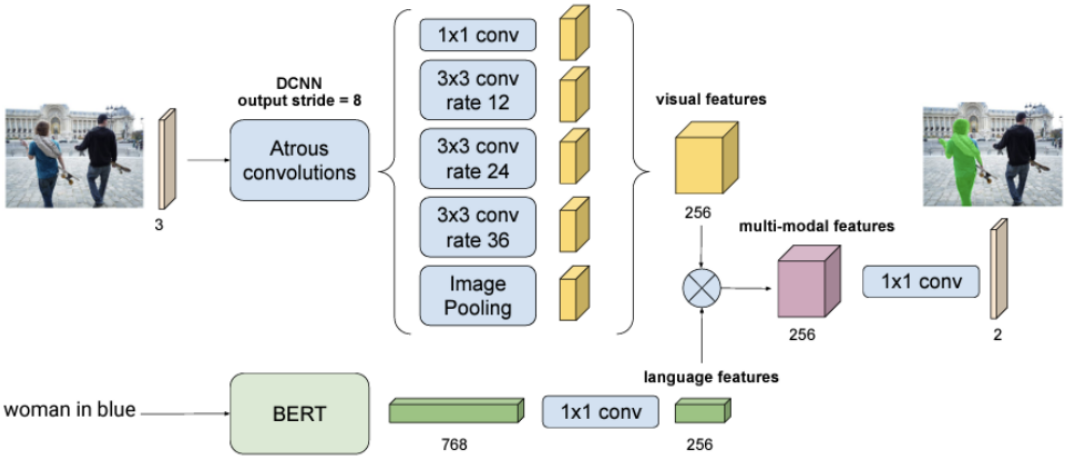


Figure 1.1: Architecture of RefVOS

1.1.2 Datasets

The RefCOCO, RefCOCO+, and RefCOCOg datasets are commonly used in computer vision for referred object segmentation. In the three papers discussed, these are the common and standard datasets used for training and evaluation purposes. A new method of GRES (Generalised Referring Expression Segmentation) is also presented. It introduced an addition to the previous benchmarks by allowing the referring to arbitrary (even 0) objects in number.

1. **RefCOCO:**

   - **Description:**
     Images from the Microsoft COCO dataset are combined with natural language expressions that refer to the specific objects in the images to create the RefCOCO (Referential COCO) dataset. Work related includes creating pixel-wise object masks for the referred items based on the textual descriptions
   - **Statistics:**
     - Number of Images: 19,994 images
     - Number of Objects: Approximately 142,000 instances
     - Average Objects per Image: Around 7 objects per image
     - Annotations: Each annotation includes a referring expression (a textual description) and a bounding box encircling the referred object.

2. **RefCOCO+:**

   - **Description:**
     RefCOCO+ is an extension of the RefCOCO dataset. The problem of object referential ambiguity—the possibility that a referring expression may point to more than one object with the same attributes—is addressed by RefCOCO+. Annotations are included in this dataset to help identify the correct object among the imprecise referents.
   - **Statistics:**
     - Number of Images: 19,992 images
     - Number of Objects: Approximately 141,000 instances
     - Average Objects per Image: Around seven objects per image
     - Annotations: Each of the annotations includes a referring expression,also a bounding box encircling the 'referred' object, and some additional information for the disambiguation of multiple objects with the same attributes.

3. **RefCOCOg:**

   - **Description:**
     RefCOCOg (Referential COCO with a gaze prior) is an extended version of the RefCOCO dataset that includes gaze information. It also incorporates gaze data obtained from human subjects, which gives insights into the typical gaze locations of individuals when deciphering referencing expressions. The referred object segmentation task can be made effective using this gaze data.

- **Statistics:**
  - Number of Images: 25,010 images
  - Number of Objects: Approximately 178,000 instances
  - Average Objects per Image: Around 7 objects per image
  - Annotations: Each annotation includes a referring expression, a bounding box surrounding the referred object, and gaze information.

2. Literature survey

2.1 Referring Expression Segmentation

Referring Expression Comprehension (REC) predicts, for each image that corresponds to a referring expression, a bounding box that tight1y enc1oses the Target 0bject. One-stage methods that predict the target bounding box directly and two-stage methods based on region proposal ranking are examples of existing works. REC is the source of Referring Expression Segmentation (RES), as defined by Hu et al. ReferIt is the original dataset for RES and REC, where a single expression refers to a single instance. RefCOCO is later proposed by Yu et al. for RES and REC. It does, however, only have single-target expressions, just like ReferIt. This is also inherited by RefCOCOg, another well-known dataset. "One expression, one instance" has evolved into the "de-facto" rule for RES tasks, despite the fact that the origina1 definition of RES doesn‘t place a 1imit on the number of target instances. RES is not far from image caption datasets, but they are not able to guarantee un‘ambiguity of expression→objec+(s). They are, therefore, unfit for tasks involving referrals. Certain referring datasets, such as Scanrefer, which focuses on 3D 0bjects, and Clevrtex, which focuses on unsupervised learning, use different data modalities or learning schemes. Furthermore, there is no-target expression in any of the aforementioned datasets. Prior research has mostly concentrated on two areas: (1) multi-modal feature fusion and (2) vision and language feature extraction. There has been a wealth of research on feature extraction, using transformer models, recurrent neural networks, CNNs, and more. Multi-modal transformers, attention mechanisms, and feature concatenation have all been investigated in feature fusion efforts.

We will discuss the difficulties in localizing objects in images using referring expressions in Referring Image Segmentation (RIS). Prior research has concentrated on multi-modal feature fusion and the extraction of features from language and vision. However, the methods currently in use limit the ability to create precise segmentation masks for objects with intricate shapes and occlusion.

2.1.1 GRES

Referring expression tasks have seen rapid progress in recent years. Early works mainly focused on Referring Expression Comprehension (REC) which locates target objects based on text queries. Datasets like ReferIt and RefCOCO established the tasks and attracted researchers to explore various approaches. Learning-based methods largely dominated this field and achieved superior performance over traditional NLP and CV pipelines.

Referring Expression Segmentation (RES) was later proposed as a more challenging and practical variant of REC, requiring models to output pixel-level target masks instead of bounding boxes. Representative RES datasets including ReferIt and RefCOCO defined clear settings where each expression refers to a single target. Pioneering works on RES investigated one-stage end-to-end networks and two-stage proposal-based pipelines. Transformer-based approaches more recently lead to a new leap in RES performance.

While existing RES datasets and methods made encouraging progress, they were subject to strong constraints of only supporting single-target expressions. This limited the applicability of RES systems in real applications. Some follow-up datasets addressed this to some extent. PhraseCut introduced multi-target samples but treated them as exceptions, and expressions were restricted to templates. Image caption datasets enabled referring in context but lacked unambiguous expression-object mappings required by RES.

To overcome these limitations, the novel GRES benchmark relaxes constraints to allow counting and describing arbitrary numbers of targets and even no targets. It also constructs the first dataset gRefCOCO containing diverse expression types for the GRES task. This significantly broadens the application scope of referring models toward more realistic scenarios.

2.1.2 PolyFormer

The suggested method, named PolyFormer, carries out multitask learning of Referring Expression Comprehension (REC) and RIS using a sequence-to-sequence (seq2seq) framework. When compared to earlier approaches that called for task-specific heads, this framework performs better. Geometric localization is handled by PolyFormer as a regression task, with continuous coordinate prediction for improved accuracy. To highlight the shortcomings of current methods for handling fragmented objects by discussion related works in contour-based instance segmentation and NLP, where seq2seq models have shown success. All things considered, PolyFormer presents a viable option for precise and comprehensive image segmentation.

2.1.3 Segment Anything

The Segment Anything Model (SAM) is a groundbreaking computer vision model that can segment any item in an image using customizable prompts such as text descriptions, clicks, or bounding boxes. Meta AI researchers presented it in their work "Segment Anything".

- It is trained using a large dataset known as Segment Anything 1B (SA-1B), which consists of 11 million licensed, privacy-respecting images from multiple sources and over 1 billion segmentation masks. Strong generalization is made achievable by the abundance and diversity of training data.
- Compared to traditional models that require significant task-specific architecture, SAM avoids specialization by acting as a basic promptable segmentation model that can handle a variety of input prompts.
- By functioning as a foundational model that can be prompted via multiple inputs,

such as clicks, boxes, or text, it may be used by a wider range of users and applications, with the main goal being to simplify the segmentation process.

- SAM simplifies segmentation by breaking down it into two parts: mask prediction, which it solves, and label prediction. Several studies are looking into expanding SAM's capabilities by mixing it with other models such as DINO, Stable Diffusion, and language models.
- It achieves cutting-edge performance on 23 datasets, exceeding previous approaches, particularly with only a few guiding points, such as clicks.(add pic for same from paper RITM)
- SAM's mask quality is high, comparable to ground truth annotations.
- Its predictive ability improves with larger network sizes, training data, and aggressive cues during inference.

The performance of SAM shows how strong vision foundation models trained on large datasets can be constructed, with flexible prompting enabling generalization to open-vocabulary tasks. SAM represents a substantial step toward general visual intelligence by proposing a unified model capable of segmenting arbitrary objects using just prompts, hence reducing specialization.

3. Methodology

3.1 GRES: Generalized Referring Expression Segmentation -Liu et al.

Acknowledging the limitations of RefCOCO and RefCOCO+, a new dataset benchmark GRES is introduced. They provide a new baseline model for it, named ReLA, along with the dataset. ReLA, dividing the image/series of images into regions adaptively using substance clues and inherently, for sole modelling the dependencies between regions and languages, is used.

3.1.1 New benchmark and Dataset

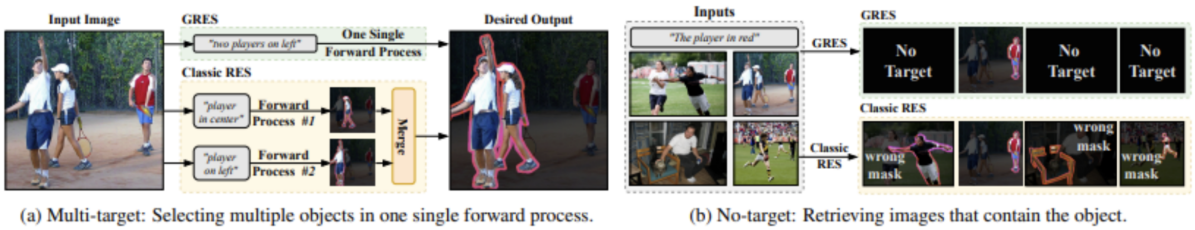
Genera1ized Referring Expression Segmentation (GRES) is a new benchmark put forth that permits expressions that indicate any number of targe+ 0bjects. Like classic RES, GRES accepts an image & a refer‘ring expression as input. Additionally, GRES supports multi-target expressions, which allow the in a sing1e expression a specification of severa1 target objects . In practice, referring expression segmentation becomes more robust and useful as a result of the increased flexibility this gives the input expression. However, no target expressions are not supported by the earlier datasets. In order to support studies on rea1istic referring segmentation, a new data‘set gRefCOCO, is constructed for GRES . Provideing two types of samples to supplement RefCOCO: no-target samples (None of the objects in the image matches the expression) and multi-target samples (where two or more target instant ces in the image are pointed towards by the expression), it is a new breakthrough.

3.1.2 GRES : Generalized Referring Expression Segmentation

Each data sample in the GRES dataset consists of four items: a language expression (T),a n image (I), a binary no-target label (EGT) indicating whether T refers to any targets, and a ground-truth segmentation mask (MGT) covering all pixels of targets referred to in T. Expression T can refer to multiple targets without any limit on the number. The GRES models are tasked with forecasting a mask M using inputs I and T. For expressions with no target, the predicted M should contain no pixels. By including mu1ti-target and no-target exps‘, GRES introduces more sophisticated and realistic scenarios for referring image segmentation, such as when a user wants to find and extract something within

a collection of images using a free-form description. This makes the task more flexible and specific than conventional image retrieval. Additionally, allowing both no-target and multi-target expressions strengthens the model's resilience and reliability for real-world use cases where any kind of expression can occur spontaneously, intentionally or unintentionally.

Share



(a) Multi-target: Selecting multiple objects in one single forward process. (b) No-target: Retrieving images that contain the object.

Figure 3.1: Application of GRES

3.1.3 Proposed ReLa model (Uses region based cross attention)

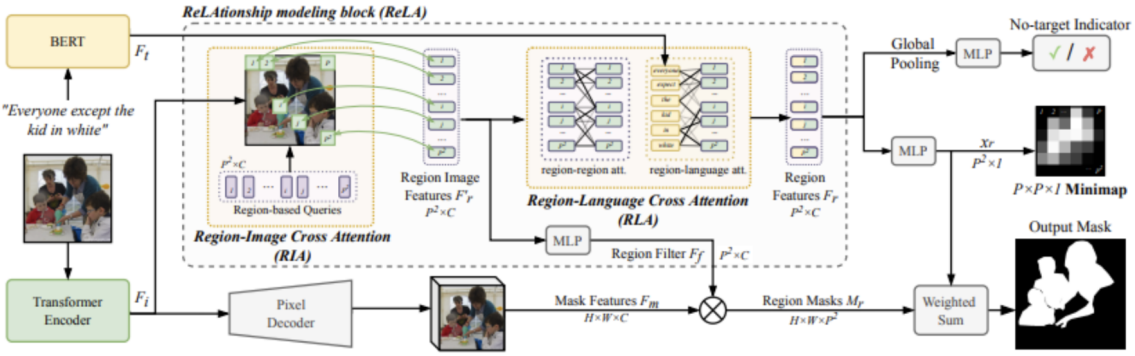


Figure 3.2: Architecture and use of ReLA in GRES

1. The text data goes to BERT, and the image data to a separate transformer encoder.

2. Both the encodings are then sent to the ReLationship modelling block (ReLA), where first the image features are divided into regions (P2 regions), i.e., p*p squares for each 2d channel. Thus a vector of shape p2 * C is generated in the RIA (Region Image cross attention) block within the ReLa block. In this manner, each region's feature is dynamically gathered from its pertinent locations. This method offers greater flexibility than hard-splitting the image into patches. Multiple regions can represent an instance in the minimap, with regions representing finer-grained attributes at the sub-instance level, such as a person's head and upper body. Sub-instance representations of this kind are needed to handle the intricate relationship and attribute descriptions found in GRES.

3. These are then further used in RLA (Region Language Cross Attention) where,
   I) self Attention is calculated of each of these p2 block with themselves.
   II) Cross attention with the Bert embeddings from the text.

The relationships between dependent regions are modeled by the self-attention. It produces the relationship-aware region feature Fr1 after computing the attention ma1rix by interacting one region's features with every other region. Concurrently, the cross attention receives region 1mage Feature F r as query input and 1anguage feature Ft as Value and Key input. The re1ationship between each word and each $^{e}gionisfirstmodeledinthisway : whereAlRP2Nt.Next, itusesthederivedword*regionattenti0n, I AlFt, toformthelanguage-awa+eregionfeatures.Lastly, theregionimagefeaturesFr, languageawareregionfeatureFr2, andinteraction - awareregionfeatureFr1arecombined.$

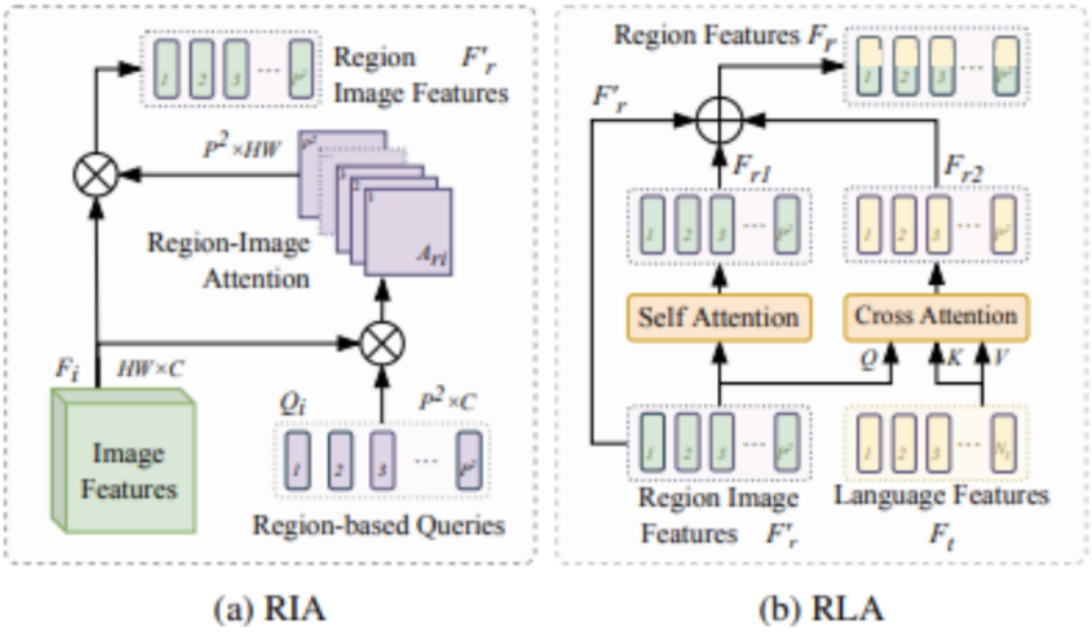


Figure 3.3: RIA vs RLA

4. All these are combined with the visual embeddings created in step 1 using MLP to get-
    - Output map (To segment the original image)
    - No target indicator (Binary output that tells if there is no target in the image that satisfies the criteria laid out in the text).
    - Minimap (A downsampled sort of version of output feature map to map each pixel to its corresponding patch created in step 2.)

3.1.4 Results

The outcomes of the same mode! +rained on RefCOCO and gRefCOCO are compared in order to demonstrate the value and necessity of gRefCOCO on the GRES task. Image (a) in Fig. 3.6 illustrates a sample of multi-target that locates "two guys" by using a shared attribu1e ("in black jacket"). The expression clearly indicates that there are two target objects, but the model trained on RefC0C0 on1y finds one. The no-target expression in Image (b) causes the RefCOCO-trained model to produce an insignificant mask. Findings show that models cannot be effectively generalized to GRES if they are solely trained on single-target referring expression datasets, such as RefCOCO. On the other hand, the

Methods	val	val	testA	testA	testB	testB
	clou	gloU	clou	gloU	clou	gloU
MattNet	47.51	48.24	58.66	59.30	45.33	46.14
LTS	52.30	52.70	61.87	62.64	49.96	50.42
VLT	52.51	52.00	62.19	63.20	50.52	50.88
CRIS	55.34	56.27	63.82	63.42	51.04	51.79
LAVT	57.64	58.40	65.32	65.90	55.04	55.83
VLT + ReLA	58.65	59.43	66.60	65.35	56.22	57.36
LAVT + RELA	61.23	61.32	67.54	66.40	58.24	59.83
RELA (ours)	62.42	63.60	69.26	70.03	59.88	61.02

Figure 3.4: Comparison on gRefCOCO dataset

Methods	val N - acc .	val T - acc .	testA N - acc .	testA T - acc .	testB N - acc .	testB T - acc .
MattNet	41.15	96.13	44.04	97.56	41.32	95.32
VLT	47.17	95.72	48.74	95.86	47.82	94.66
LAVT	49.32	96.18	49.25	95.08	48.46	95.34
RELA - 50pix	49.96	96.28	51.36	96.35	49.24	95.02
RELA	56.37	96.32	59.02	97.68	58.40	95.44

Figure 3.5: No-target results comparison on gRefCOCO dataset

model can now handle an arbitrary number of objects indicated by an expression, thanks to the recently constructed gRefCOCO.

Figure 3.6: Visual Representation of results of ReLA trained on RefCOCO & gRefCOCO

3.2 Referring Image Segmentation as Sequentia1 Po1ygon Generation -Liu et al.

Referring image segmentation (RIS) is a concept that localizes the segmenta1ion mask of a objec-tbased on a natura1 1anguage query by combining vision-language understanding and instance segmentation. The traditional RIS pipeline consists of taking text and image inputs, extracting features, and combining them to predict the mask. Nevertheless, this method ignores the organization of the output forecasts. PolyFormer is a sequence-2-sequence (seq2seq) framework that is suggested as a solution. It receives as input text query tokens and image patches and outputs a seq. of po1ygon vertices. Structured output predictions are made possible by PolyFormer, which conditions each vertex prediction on vertices that have already been predicted. We go over how localization can be formulated as a regression task and show how well PolyFormer predicts bounding box corner points and polygon vertices. Three referring image segmentation benchmarks are used to assess PolyFormer's performance, and it outperforms previous methods in each case.

3.2.1 Polyformer

While PolyFormer predicts the vertices of polygons bounding the object and the corner points of bounding boxes, traditional image segmentation methods predict dense segmentation masks. The framework comprises of two encoders: one for text and the visual, which extract features from the text and images respectively. Then, a multi-modal transformer encoder receives these features projected in1o a shared embedd1ng space. The bounding box and polygon vertices' continuous floating-point coordinates are produced in an autoregressive fashion by a regression-based transformer decoder. The polygons

serve as the basis for creating the segmentation mask. Cutting-edge outcomes are obtained by PolyFormer on a number of image segmentation benchmarks, such as COCO, LVIS, and Cityscapes. The suggested technique shows superior geometric localization capabilities and performs better than current image segmentation techniques.

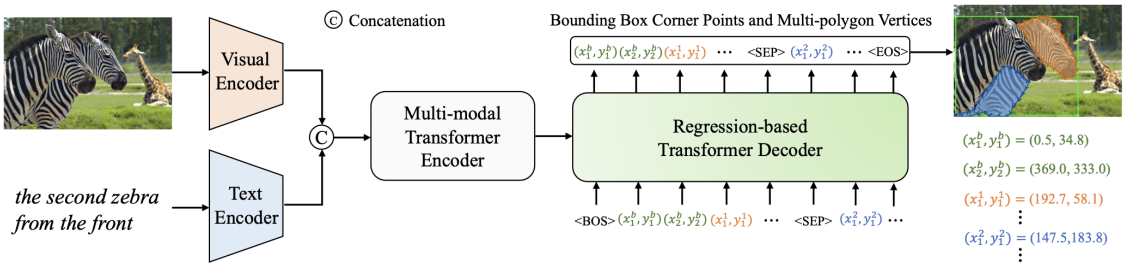


Figure 3.7: Architecture of Polyformer

3.2.2 Regression-based Transformer Decoder

Previous visual sequence-2-sequence methods discretized continuous coordinates into bins, resulting in unavoidable quantization errors between the original coordinate value x and its binned approximation [x]. To predict one of these bins, they formulated coordinate 1oca1ization as a c1assification prob1em. However, classification is suboptimal for geometric localization where precise coordinate prediction is required. To address this limitation, we propose a regressio-nbased decoder that directly forecasts the actual continuous coordinate values. In contrast to classification-based approaches, this avoids discretizing and quantizing the coordinates during both training and inference.

3.2.3 2D Coordinate Embedding

A 2D coordinate embedding approach is proposed/introduced to improve the accuracy of coordinate representation in image segmentation produced as output allowing precise coordinates of embedding for any of the floating-point coordinates (x, y) in an image. Four discrete bins are created by performing ceiling and floor operations on (x, y), and the matching embeddings are indexed from the 2D codebook. Accurate coordinate embeddings are then obtained by means of bilinear interpolation. Decoder layers consisting of a feed-forward-network, a multi-head-cross-attention layer, and a multi-head-self-attention layer, are used for capturing the relations between both, the multi-modal features and output 2D coordinate embeddings.

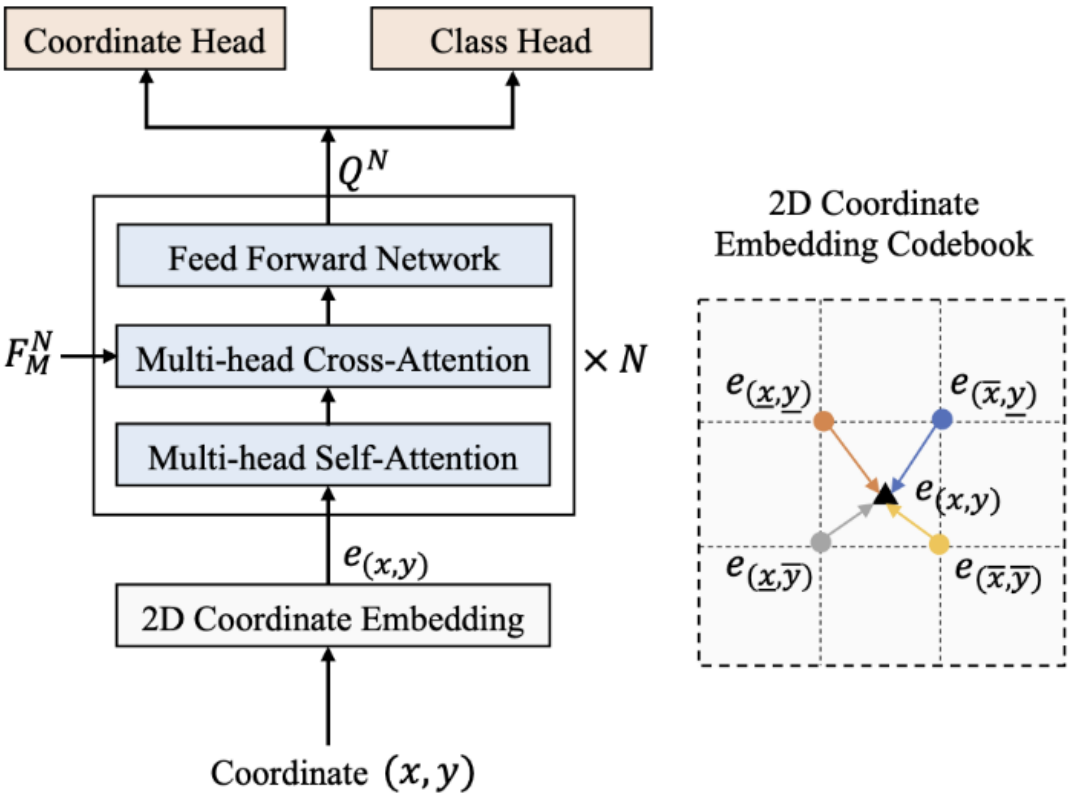


Figure 3.8: 2D coordinates

3.2.4 Proposed Model

The groun-dtruth bounding box and polygon seq2encs are represented by levying a target sequence construction.

- Polygon Representation:One or more polygons re[resenting a segmentation mask that encircle the object being referred to. The vertex nearest 2 the upper 1eft corner of the image is selected as the sequence's beginning p0int.
- Vertex and Special Token: Without using any quantization, we preserve the continuous f1oating point value of the origina1 vertex coordinate, which is either the x or y coordinate. For representing multiple polygons, the introduction of a seperator token <SEP> is levied between two po1ygons. The order of polygons belonging to the same object is determined by the separation between their origin and starting points. Lastly, we designate the start and finish of the sequence using the <BOS> and <EOS> tokens.
- Unified Sequence with Boundin Box: Top-1eft and bottom-right are two corner points that make up a bounding box. The coordinate tokens <COO> are defined as the bounding box corner points and po1ygon vertices.

The input of the framework consists of a referring expression T and an image I.

- Image Encoder: For the 1nput 1mage is encoded using a Swin trans former to extract features from the 4th stage as visual representation.

- Text Encoder: The 1anguage embedding mode1, BERT is used to extract word features for a given language description.
- Multi-Model Transformer Encoder: The mu1ti-moda1 encoder contains N transformer layers for fusing image and text features. Each layer has a mu1ti-head se1f-attention mechanism, followed by layer normalization and a feed-forward network. Abso1ute positiona1 encodings are added to the image + text featur‘es to preserve positional information. Additionally, 1D relative position biases are added to the 1mage features to capture relationships between image patches, while 2D relative biases are used for the text features to model relationships between tokens. This fusion architecture leverages self-attention and relative positioning to process image and language representations jointly.

It is later fed to the Regression-based Transformer Decoder for outputting the corresponding predictions of the 2D c00rdinates of the Referred-0bject bounding box corner points and po1ygon vertices.

3.2.5 Training

It is trained using the technique of polygon augmentation to increase diversity in sparse object contours, achieved by the interpolation of the dense contour later applying uniform down-sampling with a randomly sampled interval. The objective of the model (trained using a combination of L1 regression loss and 1abe1-smoothed crosse-ntropy 1oss)is the prediction of the sequential token/tokens and it s type based on a referring expression, image, and preceding tokens. Starting with the token, the model progressively generates the output sequence during inference. Up until the token is output, it predicts the token type, coordinates, and separators. The anticipated polygons are used to create the final segmentation mask.

3.2.6 Results

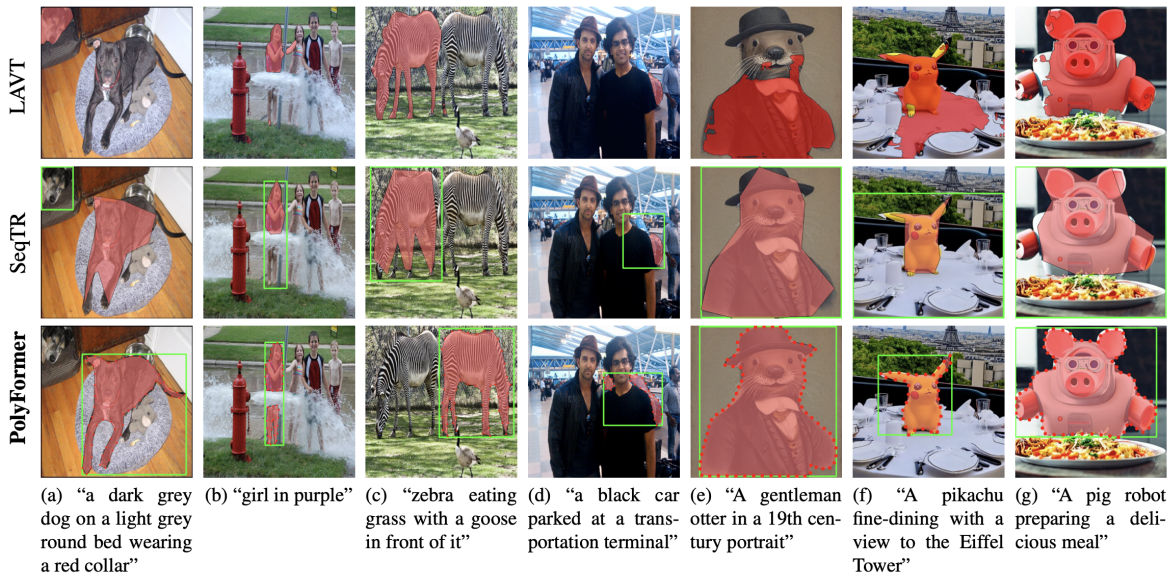
Experimentations are conducted on the following four benchmarks: RefCOCO, RefCOCO+, RefCOCOg, and ReferIt. Mean Intersection-over-Union i.e. mIoU is used for the Referring Image Segmentation (RIS) evaluation metric, and Precision@0.5 is used for the Referring Expression Comprehension (REC). PolyFormer-B, with Swin-B as visual encode and BERTBASE as text-encoder, achieves superior results in comparison to previous method/methods on the mentioned three datasets under all metrics.

	Method	Visual	Text	RefCOCO			RefCOCO +			RefCOCOg		ReferIt
		Backbone	Encoder	val	test A	test B	val	test A	test B	val	test	test
oIoU	STEP	RN101	Bi - LSTM	60.04	63.46	57.97	48.19	52.33	40.41	-	-	64.13
	BRINet	RN101	LSTM	60.98	62.99	59.21	48.17	52.32	42.11	-	-	63.11
	CMPC	RN101	LSTM	61.36	64.53	59.64	49.56	53.44	43.23	-	-	65.53
	LSCM	RN101	LSTM	61.47	64.99	59.55	49.34	53.12	43.50	"	-	66.57
	CMPC +	RN101	LSTM	62.47	65.08	60.82	50.25	54.04	43.47	-	-	65.58
	MCN	DN53	Bi - GRU	62.44	64.20	59.71	50.62	54.99	44.69	49.22	49.40	-
	EFN	WRN101	BI - GRU	62.76	65.69	59.67	51.50	55.24	43.01	I	-	66.70
	BUSNet	RN101	Self - Att	63.27	66.41	61.39	51.76	56.87	44.13	I	I	-
	CGAN	DN53	Bi - GRU	64.86	68.04	62.07	51.03	55.51	44.06	51.01	51.69	-
	LTS	DN53	BI - GRU	65.43	67.76	63.08	54.21	58.32	48.02	54.40	54.25	-
	RESTR	VIT - B	Transformer	67.22	69.30	64.45	55.78	60.44	48.27	-	-	70.18
	PolyFormer - B	Swin - B	BERT - base	74.82	76.64	71.06	67.64	72.89	59.33	67.76	69.05	71.91
	PolyFormer - L	Swin - L	BERT - base	75.96	78.29	73.25	69.33	74.56	61.87	69.20	70.19	72.60
mIoU	VLT	DN53	BI - GRU	65.65	68.29	62.73	55.50	59.20	49.36	52.99	56.65	-
	CRIS	RN101	GPT - 2	70.47	73.18	66.10	62.27	68.06	53.68	59.87	60.36	-
	SeqTR	DN53	Bi - GRU	71.70	73.31	69.82	63.04	66.73	58.97	64.69	65.74	-
	RefTr	RN101	BERT - base	74.34	76.77	70.87	66.75	70.58	59.40	66.63	67.39	-
	LAVT	Swin - B	BERT - base	74.46	76.89	70.94	65.81	70.97	59.23	63.34	63.62	-
	PolyFormer - B	Swin - B	BERT - base	75.96	77.09	73.22	70.65	74.51	64.64	69.36	69.88	65.98
	PolyFormer - L	Swin - L	BERT - base	76.94	78.49	74.83	72.15	75.71	66.73	71.15	71.17	67.22

Figure 3.9: Comparison of the model on 3 refer‘ring 1mage Segmentation benchmarks with stat-eof-art methods

Method	Visual	Text	RefCOCO			RefCOCO +			RefCOCOg	RefCOCOg	ReferIt
	Backbone	Encoder	val	test A	test B	val	test A	test B	val	test	test
UNTIER - L	RN101	BERT	81.41	87.04	74.17	75.90	81.45	66.70	74.86	75.77	-
VILLA - L	RN101	BERT	82.39	87.48	74.84	76.17	81.54	66.84	76.18	76.71	-
RefTr	RN101	BERT - base	85.65	88.73	81.16	77.55	82.26	68.99	79.25	80.01	76.18
Seq TR	DN53	Bi - GRU	87.00	90.15	83.59	78.69	84.51	71.87	82.69	83.37	69.66
MDETR	ENB3	ROBERTa - base	87.51	90.40	82.67	81.13	85.52	72.96	83.35	83.31	-
OFA - B	RN101	Embedding layer	88.48	90.67	83.30	81.39	87.15	74.29	82.29	82.31	-
UniTAB	RN101	ROBERT - base	88.59	91.06	83.75	80.97	85.36	71.55	84.58	84.70	-
OFA - L	RN152	Embedding layer	90.05	92.93	85.26	85.80	89.87	79.22	85.89	86.55	-
PolyFormer - B	Swin - B	BERT - base	89.73	91.73	86.03	83.73	88.60	76.38	84.46	84.96	80.90
PolyFormer - L	Swin - L	BERT - base	90.38	92.89	87.16	84.98	89.77	77.97	85.83	85.91	81.50

Figure 3.10: Comparison on 3 refer-ring expression c0mprehension benchmarks with the stateo-f-the-ar1 methods



(a) "a dark grey dog on a light grey round bed wearing a red collar" (b) "girl in purple" (c) "zebra eating grass with a goose in front of it" (d) "a black car parked at a transportation terminal" (e) "A gentleman otter in a 19th century portrait" (f) "A pikachu fine-dining with a view to the Eiffel Tower" (g) "A pig robot preparing a delicious meal"

Figure 3.11: Visual Result Comparison

Fig. 3.11(a)–(d) displays the PolyFormer visualization results on the RefC0C0g test set. It is evident that Poly'Former can segment the referred 0bject in difficult situations, such as those involving complex shapes and occlusion, partially displayed objects, or situations requiring complex language comprehension. In Fig. 3.11(e)–(g), we also display the outcomes for the images produced by stable diffusion. With synthe1ic images and text descri1tions that it has never seen before, PolyFormer shows good generalization ability. On the other hand, the most advanced LAVT and SeqTR algorithms are unable to produce satisfactory outcomes.

3.3 Segment Anything Model by Meta AI

The Architecture of SAM comprises three components working together to return a valid segmentation mask:

- Image encoder: one-time image embedding generation.
- Pompt encoder: embedding the prompts.
- Lightweight mask decoder: for combining the output of the embedding from the prompt and image encoders.

3.3.1 Segment Anything Data Engine

Here's a summary to add to the methodology section regarding the Segment Anything Data Engine:

- Developed to create the large-scale SA-1B dataset with 1.1 billion+ segmentation masks.

- Employs a model-in-the-loop data annotation approach with three stages:
  1. **Model-Assisted Manual Annotation**
     - Professional annotators use an interactive SAM-powered tool.
     - Label masks with foreground/background point prompts.
     - Annotation efficiency increases as the model improves.
  2. **Semi-Automatic Annotation**
     - SAM by prompting with likely locations generates masks for a subse1 of objects.
     - Annottators focus on labelling the ramaining of objects.
     - Increases maskd-iversity in the dataset.
  3. **Fully Automatic Annotation**
     - Annotators promp+ SAM with a regular grid of for-eground points.
     - Generates around 100 high-qua1ity masks per image on average.
     - Uses ambiguity-aware predictions and non-maximal suppression.

This iterative data engine approach enabled the creation of the massive SA-1B dataset:

- 1.1 billion high-quality masks from 11 million licensed, privacy-preserving images.
- Facilitated advanced research in computer vision and segmentation.

The data engine leverages SAM's capabilities in an iterative process, transitioning from manual to semi-automatic to fully automatic stages, allowing the efficient creation of a diverse, large-scale segmentation dataset that powers the training of SAM.

3.3.2 Architecture

The Segment Anything Model (SAM) network architecture contains three crucial components: the 1mage Encoder, the Promp1 Encoder, and the Mask Decoder.

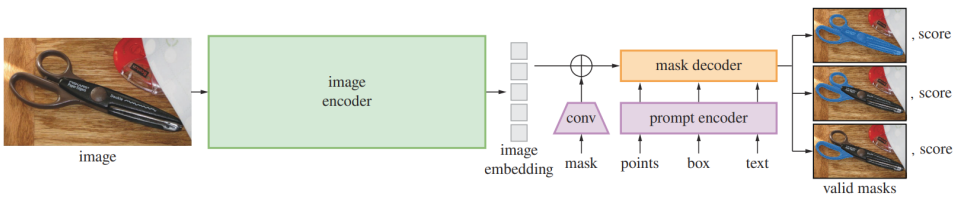


Figure 3.12: Architecture

- Image Encoder:
  The Masked Autoencoder (MAE) pre-trained Vision Transformer (ViT) backbone serves as the foundation for SAM's Image Encoder. It employs the ViT-H/16 (Vision Transformer - Huge with 16x16 patch size) design, a robust and scalable vision transformer framework [**?**]. The Image Encoder processes a superior resolution input

image of 1024x1024x3 (common for many applications) using the ViT-H/16 backbone. This encoder produces a dense feature embedding with dimensions 64x64x256, resulting in a 16x downscaled replica of the original image.
This downscaling step is critical for effective processing while preserving important image information. The 64x64x256 feature embedding provides a small but useful depiction of the input image, collecting local as well as global visual information.
The Image Encoder runs once each input image, creating these image-embedded data before triggering the model. This approach enables the Image Encoder to be seamlessly integrated into the overall segmentation process, because the image embeddings may be efficiently utilized for many prompts without demanding unnecessary computations.

- Prompt Encoder:

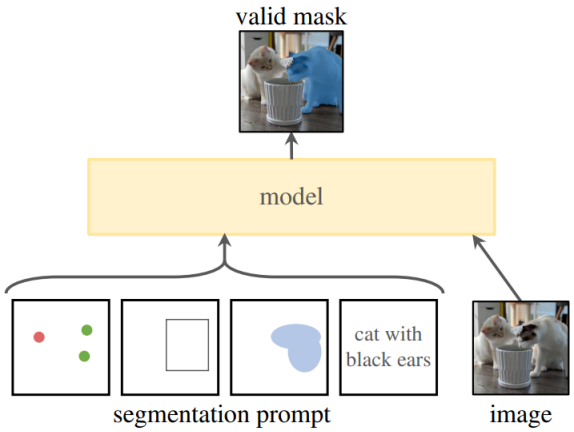


Figure 3.13: Prompt Encoder

  - Encodes several forms of prompts into embedded vectors in real time.
  - Handles two kinds of prompts:
    - Sparse prompts (boxes, points, text):
      - Points: Points are represented by positional encoding plus learnt embedding (foreground/background).
      - Boxes: Positiona1 encodings of top-1eft and bottom-right borders plus learnt embeddings.
      - Text: Encoded with an off-the-shelf CLIP text encoding.
    - Dense Prompts (Masks):
      - Downsized by a factor of four.
      - Convolution layers (2x2 stride-2 + 1x1) were used to generate 256 channels.
      - Mask embedding added element-wise to image embedding.
      - For no mask, the learnt 'no mask' embedding was applied.
  - Designed to interpret multiple prompt types.
  - Output prompt embedded data for fusion with picture embeddings.

The prompt encoder effectively handles multimodal prompts, employing positional/learned embedded data for sparse prompts such as points, boxes, and text, and convolution encoding for dense mask prompts.

- Mask Decoder:

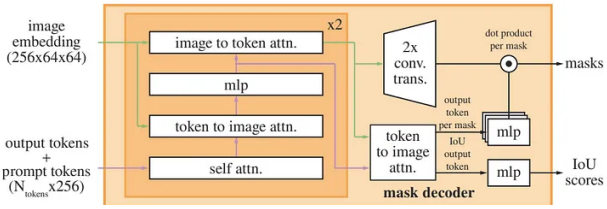


Figure 3.14: Mask Decoder

  - A lightweight decoder that predicts final segmentation masks.
  - Accepts image-embedded data generated by the Image Encoder and prompt-embedded data from the Prompt Encoder as input.
  - Based on the modified Transformer decoder architecture.
  - Incorporates learned output token embedding with prompt embeddings.
    1. Plays an important role, analogous to class tokens in visual transformers.
    2. Guides for decoding to ensure appropriate image segmentation.
  - Each decoder layer performs four steps:
    1. Se1f-attention to tokens (prompt embedded data + output token).
    2. Cross-attention between tokens and picture embeddings.
    3. Point-by-point MLP updates each token.
    4. Cross-attention between picture embeddings and tokens (updates image using prompt information).
  - Utilizes prompt se1f-attention and cross-attention techniques in both direct1ons.
  - Incorporates dynamic mask prediction head.
  - Optimized for real-time performance.
  - Predicted masks can be annotated to continuously update model weights.
  - Enables efficient mapping of images and rapid embeddings to segmentation masks.
  - Enables responsive prompting and continuous learning.

The Mask Decoder combines picture and prompt information using a modified Transformer decoder architecture that employs cross-attention in each direction. The learned output token embedding guides the segmentation procedure. The total architecture supports performance in real-time, interactive prompting, and continuous learning from annotated data.

3.3.3 Key Findings

- SA is evaluated on 23 different semantic segmentation tasks spanning domains like natural images, biomedical images, satellite imagery, etc.
- SA achieved strong zero-shot and few-shot segmentation performance directly on test sets of unlabeled data from new domains, outperforming previous approaches that require domain-specific tuning.
- When fine-tuned on small labelled datasets, SA achieved SOTA results on 16 out of the 23 segmentation tasks, often surpassing models specially designed for each domain. This demonstrates its ability to leverage its broad pre-trained knowledge.
- SA also showed good generalization within domains by performing well on the tasks with different object categories or image modalities from fine-tuning datasets.
- Notably, SA was able to segment objects in satellite and medical image modalities that are visually very different from natural scenes, through cross-domain knowledge transfer.
- After an analysis of the model's representations, it is found that SA learns domain-general concepts like edges, regions and textures during pre-training that help with the segmentation. Its Transformer also captures the global context relationships.
- The Segment Anything model demonstrates that a single CNN-Transformer can perform semantic segmentation competently across various domains and modalities without any domain-specific design or tuning. Its self-supervised pre-training approach effectively teaches the model structural image properties, which are relevant for segmentation in diverse contexts, simplifying the process of applying deep learning tools to new domains and tasks.

3.3.4 Results

Zero-Shot Sing1e Point Valid Mask Eva1uation

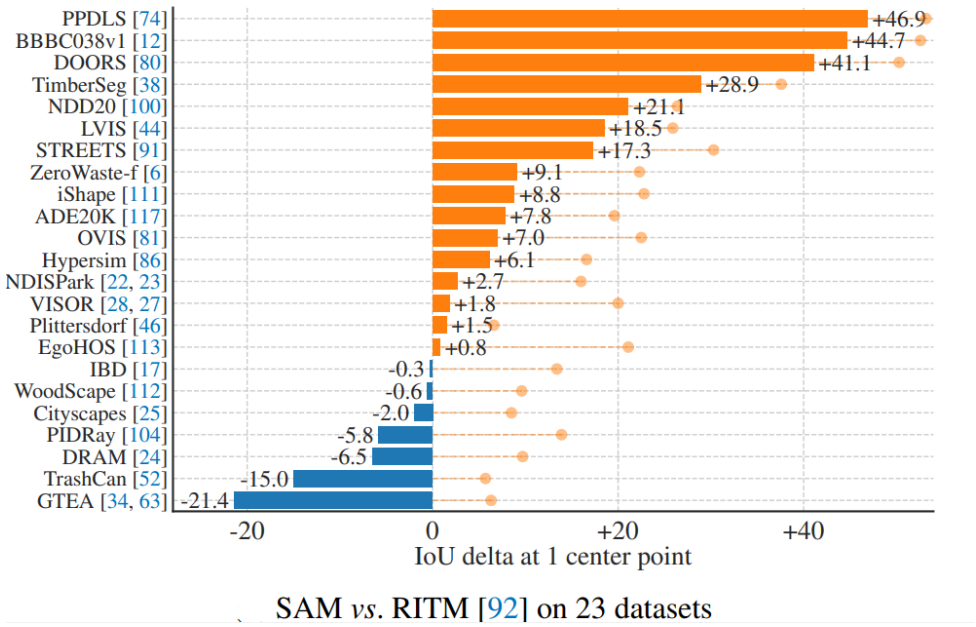


Figure 3.15: SAM vs. RITM on 23 datasets

The average IoU of SAM and the greatest sing1e-point segmenter of R1TM. Due to uncertainty, a single mask could differ from the ground truth. Circles represent the "oracle" outcomes which are the most relevant of SAM's three forecasts.

Zero-shot transfer to edge detection

Qualitatively, we see that, despite not being trained for edge identification, SAM provides reasonable edge maps. SAM predicts more edges than the ground truth, including sensible ones not indicated in BSDS500.

Table 3.1: Edge Detection Performance Comparison

Method	Year	ODS	OIS	AP
HED	2015	0.788	0.808	0.84
EDETR	2022	0.84	0.858	0.896
Sobel filter	1968	0.539	-	-
Canny	1986	0.6	0.64	0.58
Felz-Hutt	2004	0.61	0.64	0.56
SAM	2023	0.768	0.786	0.794

Zero-Shot Object Proposals

SAM outperforms ViTDet-H in the detection of large and medium-sized objects, both frequent and rare. However, SAM falls short of ViTDet-H in detecting small and common objects.

Table 3.2: Object Proposal Performance Comparison

Method	**All**	**Small**	**Med.**	**Large**	**Freq.**	**Com.**	**Rare**
ViTDet-H	63	51.7	80.8	87	63.1	63.3	58.3
SAM - sing1e out.	54.9	42.8	76.7	74.4	54.7	59.8	62
SAM	59.3	45.5	81.6	86.9	59.1	63.9	65.8

Zero-Shot Instance Segmentation

When comparing the masks generated by SAM and V1TDet on the two different C0C0 and LVIS datasets, we notice variances in mask AP metrics, with SAM behind ViTDet but being somewhat close. Yet, upon visual inspection, SAM masks frequently demonstrate superior qualitative characteristics, particularly finer boundaries, as compared to ViTDet.

Table 3.3: Instance Segmentation Performance Comparison

COCO				
Method	**AP**	**APs**	**APm**	**APl**
ViTDet-H	51	32	54.3	68.9
SAM	46.5	30.8	51	61.7
LVIS v1				
ViTDet-H	46.6	35	58	66.3
SAM	44.7	32.5	57.6	65.5

Testing SAM

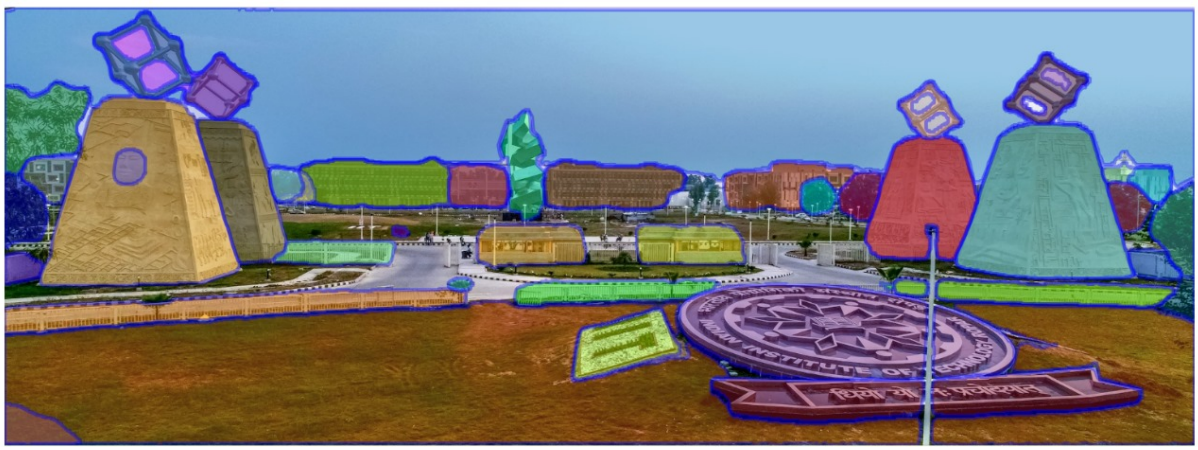


Figure 3.16: Multiple Object Proposals by Segment Anything Model on IIT Ropar Image

4. Conclusion and Future Work

4.1 Conclusion

4.1.1 Segment Anything

Finally, this study describes the Segment Anything (SA) project, which consists of three interrelated components: a unique promotable segmentation job, a successful Segment Anything Model (SAM), & the construction of a gigantic dataset known as SA-1B, which contains over 1 billion segmentation masks. The promptable segmentation challenge tries to generate correct segmentation masks using customizable prompts such as points, boxes, or text descriptions. SAM is built with an effective architecture to allow for real- time interactive prompting, using separate encoders for the picture and prompts that are integrated with a lightweight mask decoder.

To train SAM, a "data engine" was created to repeatedly collect a large-scale dataset by having the model generate masks while human annotators provided further supervision. The resulting SA-1B collection has 1.1 billion high-quality masks on 11 million licensed photos, which is far larger than any previous segmentation dataset.

Extensive tests show that SAM has outstanding zeros-hot capabilities on a w1de range of segmentation assignments and datasets, sometimes competing with or outperforming previous fully supervised approaches. This effort seeks to enable continued research into establishing robust and general foundation models for tasks involving computer vision such as SAM, which may be flexibly adapted to new situations via prompting. The large-scale data, promptable formulation, and model architecture are all promising advances towards this aim. Several studies are looking into ways to improve SAM's capabilities by combining t may be applied in various informational educational institutions.

4.1.2 PolyFormer

PolyFormer efficiently combines multi-task predictions as output with multi-modal features as input. It is noteworthy because it presents a new innovative regression-based decoder that produces accurate 2-D coordinates free of errors of quantzation . PolyFormer exhibits promising generalization to unknown scenarios and achieves competitive performance in both RIS and REC tasks, according to experimental results. It is possible to apply this straightforward framework to tasks other than RIS and REC, according to the researchers. All things considered, PolyFormer provides a simple yet effective way to deal with the problems of image segmentation and referring expression comprehension.

4.2 Future Work

- **Increased Research on No Target Objects:** While using GRES, about 40 per cent of the samples consisting of No-Target objects were missed. Dedicated research on this area can increase the efficiency and reliability on the upcoming models. This shows that a dediceted no-target c1assifier is des1red.
- **Better Regression Models:** Polyformer indicated that using regression-based models provides better accuracy than classification-type models. Better models can be proposed in this area.
- **Extensibility of Models:** Model of polyformer can be scaled for usage in fields other than REC and RIS.
- **Minimal GPU usage:**Accessibility of VOS to devices of less GPU is the upcoming benchmark to be scaled down to devices like mobile phones with low memory.
- **Fast Moving objects in videos:** Fast moving objects in videos may lead to frames of objects with motion blur. Low-Quality Image Feature recognition techniques can be inculcated in the model for better accuracy.
- **Extending the model to 3D medical image segmentation:** The Segment Anything mode1 has shown great promize in segmenting objects from 2D natural and medical images. Most medical imaging modalities like the CT and the MRI acquire 3D volumetric data. Extending the model to directly take 3D volume as input and output semantic 3D segmentation can unlock its full potential for medical applications. This would allow analysis the 3D anatomical structures and pathological regions in a single go.
- **Using Various Faster Encoding Architectures:**By using an improved Prompt Encoder, we can improve the speed and accuracy of the model. Since SAM is used in real-time, a slight increase in speed from prompt encoding can significantly improve the results and usage. For Polyformer, we can upgrade BERT to a recent update.
- **Incorporate multi-modal inputs:**Medical diagnosis often combines information from multiple imaging modalities, which provide complementary information like CT, MRI and PET. Incorporating multi-modal inputs into the Referring Object Segmentation models can help in leveraging the strengths of different modalities and therefore improve segmentation accuracy, making it useful for segmenting lesions which may be only visible in some of the modalities.
- **Create a mobile-friendly model:**While the existing Segment Anything model is quite accurate, it includes billions of parameters and requires significant GPU resources. Creating a lighter and more efficient mobile model allows for deployment on edge devices such as tablets and phones. This allows doctors to perform early picture analysis at the point of service without relying on centralised computing. On-device segmentation can also help to develop new mobile health applications.
- **Expand to additional anatomical structures and pathologies:** Thus far, the model has been used to segment main organs and several lesions. Expanding it to more detailed anatomic and pathologic structures can help it reach its full potential.

This involves segmenting smaller organs, dividing 3D volumes into substructures, and diagnosing various diseases.

- **Publish pre-trained models:** Creating pre-trained segmentation models that is publicly available for typical medical or any other problem statement segmentation activities can help lessen the barrier to adoption. This can speed up program creation and validation studies across new datasets and healthcare organizations while requiring minimal training resources.

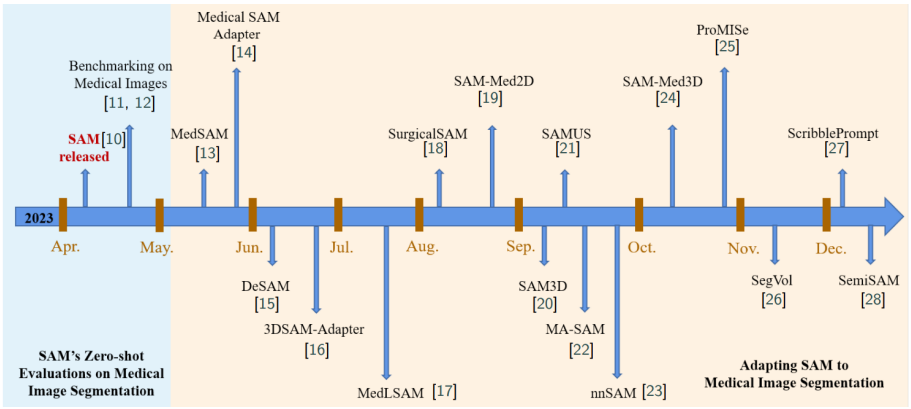


Figure 4.1: Segment Anything Mode1 for Medical 1mage Segmentation: Current App1ications and Future Directions

5. References

[1]. Zhang, Z. Shen, and R. Jiao, “Segment anything model for medical image segmentation: Current applications and future directions,” 2024.

[2]A. Kirillov et al., "Segment Anything," 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2023, pp. 3992-4003,
doi: 10.1109/ICCV51070.2023.00371.

[3] U. Doshi. (2023) Segment anything model (sam) explained. [Online]. Available: https://medium.com/@utkarsh135/segment-anything-model-sam-explained-2900743cb61e

[5] Bellver, M., Ventura, C., Silberer, C., Kazakos, I., Torres, J. (2020). RefVOS: A Closer Look at Referring Expressions for Video Object Segmentation

[6]Chang Liu, Henghui Ding, Xudong JiangGRES, (2023).GRES: Generalized Referring Expression Segmentation

[7]Jiang Liu1,, Hui Ding2, Zhaowei Cai, Yuting Zhang, Ravi Kumar Satzoda, Vijay Mahadevan, R. Manmatha (2023). PolyFormer: Referring Image Segmentation as Sequential Polygon Generation