Multi-Modal Learning using AI

Referred Object Segmentation / Prompt Segmentation

Hiya Kwatra

Department of Electrical Engineering
IIT Ropar

May, 2024

Motivation

The goal of multimodal models is to improve computer-human communication by processing information in several forms, such as text, graphics, and audio.These models integrate many modalities to offer a more thorough comprehension of the data. Their proficiency in cross-modal learning enhances activities like captioning images. Multimodal models play a crucial role in real-world applications such as driverless cars by analysing several kinds of data. Their capacity to solve problems with a single kind of data and their contribution to the field of AI research make them effective instruments for developing computer systems that are more intelligent and flexible.

Referred Object segmentation

Conventional computer vision tasks, such as segmentation and detection, have limited their scalability and practicality by dealing with a pre-defined set of categories. It makes sense to replace the pre-defined categories with natural language expressions (NLE) in order to address the issues mentioned above. As an example, the phrase "the kid running after the butterfly" requires localizing only the child pursuing the butterfly and not the other children. In fact, this is how humans interact with objects in their environment. Previous approaches to the grounding problem involve forecasting a bounding box surrounding the object of interest or a segmentation mask that corresponds to the object of interest. We will talk about the recent works to increase the accuracy and efficiency of this task through various methods.

- 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.
- 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.
- While existing RES datasets and methods made encouraging progress, they were subject to strong constraints of only supporting single-target expressions.
- 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.

- 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.

- 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.
- 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.
- Its predictive ability improves with larger network sizes, training data, and aggressive cues during inference.

RefVOS: Referred Video Object Segmentation

- RefVOS, is a novel model competitive for language-guided image segmentation and state-of-the-art for language-guided VOS.
- It aims to individually produce embeddings for the image (visual features) and text separately using:-
  1. **Visual Encoder:**
     using DeepLabv3
  2. **Language Encoder:**
     using BERT fined tuned to REs of RefCOCO dataset (simple annotations).
- These features are then combined to get a multimodal feature space.

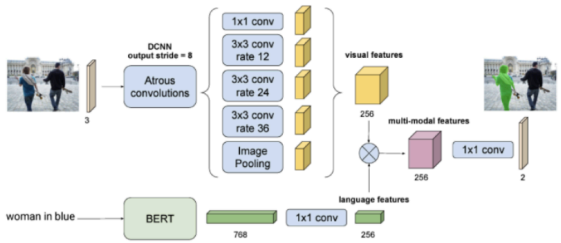

Figure 1: Architeture of RefVos

GRES: Generalized Referring Expression Segmentation

- Introduction of new dataset benchmark that permits expressions that indicate any number of target objects
- Introduction of new dataset gRefCOCO to accommodate the following:-
  1. no-target samples
     the expression does not match any object in the image
  2. multi-target samples
     the expression points to two or more target instances in the image
- Allowing both no-target and multi-target expressions strengthens the models resilience and dependability in real-world scenarios where any kind of expression can happen at any time, such as when users purposefully or unintentionally typed the wrong sentence

GRES

Figure 2: Display of comparison on RefCOCO and gRefCOCO

Referring Image Segmentation as Sequential Polygon Generation

- It receives as input text query tokens and image patches and outputs a sequence of polygon vertices. Structured output predictions are made possible by PolyFormer, which conditions each vertex prediction on vertices that have already been predicted.
- Includes the following referring image benchmarks:-
  1. **Polyformer:** PolyFormer predicts the vertices of polygons bounding the object and the corner points of bounding boxes.The polygons serve as the basis for creating the segmentation mask.
  2. **Regression-based Transformer Decoder:** It directly predicts the continuous coordinate values rather than using quantization (not ideal for geometric localization).
  3. **2D Coordinate Embedding** Improves 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.

- Multi-Task Prediction: PolyFormer efficiently combines multi-task predictions and multi-modal features, showcasing a regression-based decoder for accurate 2-D coordinates.
- Generalization and Performance: PolyFormer demonstrates promising generalization in unknown scenarios and competitive performance in image segmentation and referring expression comprehension.
- Outstanding zeros-hot capabilities: 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.

- **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:** Most medical imaging modalities like CT and 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, allowing analysis of 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.
- **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.

SAM in medical Imaging Segmentation

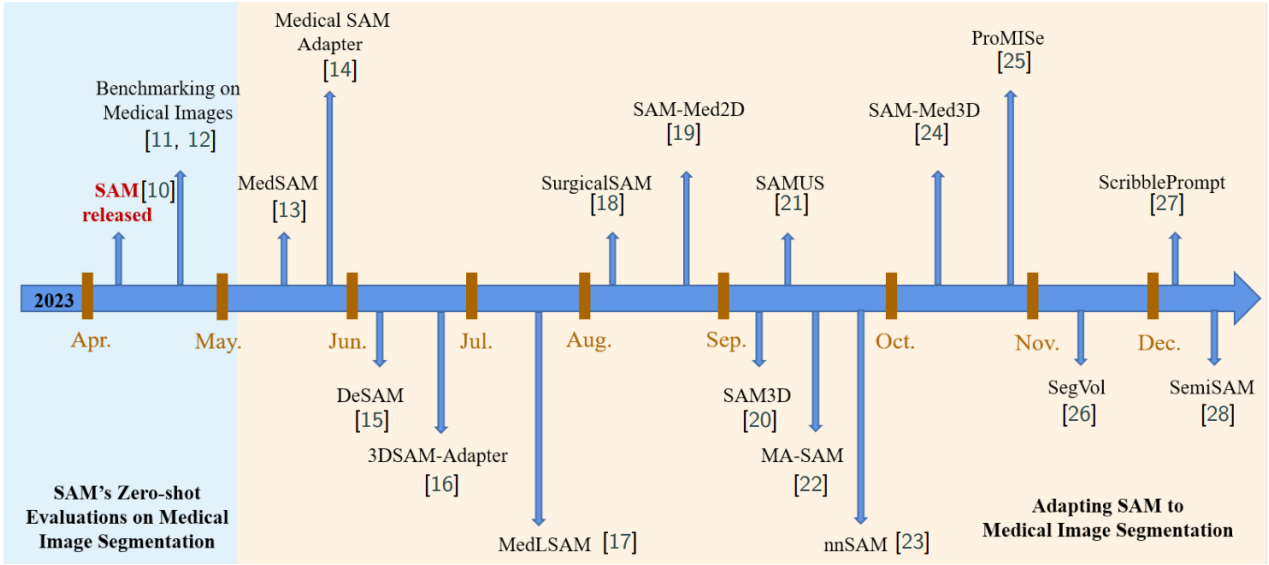

Figure 3: Segment Anything Mode1 for Medical 1mage Segmentation: Current App1ications and Future Directions

[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

Thank You