CLIP-Driven Universal Model for Organ Segmentation and Tumor Detection

EunBin CHO

Visual Computing and Medical Imaging Lab.,
Department of Software Convergence,
Seoul Women's University

Contents

Paper Information

Authors

- Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A. Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, Zongwei Zhou

Affiliation

- Johns Hopkins University, City University of Hong Kong, Vanderbilt University, NVIDIA

Publication

- Publication Platform: arXiv
- Publication Date: Aug 17, 2023
- Github Page: https://github.com/ljwztc/CLIP-Driven-Universal-Model

Introduction

■ Research Topic

- Development of CLIP-Driven Universal Model for Medical Imaging
  - Develop universal model for segmenting organs & tumors by applying **CLIP Embeddings**
  - Ensure **generalization** & offer **high accuracy and efficiency**
  - Overcome the limitations of traditional one-hot label encoding
  - Approach CLIP embeddings to **capture semantic relationships** between organs and tumors.

Introduction

■ Research Necessity

- ◻ Limitations of Existing Medical Imaging AI Models
  - Partially labeled datasets
  - Poor generalization performance
  - One-hot label encoding lacks semantics

■ Research Objectives and Expected Effects

- ◻ CLIP-based embeddings for meaningful segmentation
  - More intuitive anatomical relationship learning and improved generalization
- ◻ Masked Back-Propagation for partial labels
  - Effective use of partially labeled datasets enables robust learning & the development of a highly generalizable model
- ◻ Universal Medical Imaging AI model
  - Highly generalizable AI model with consistent performance across diverse CT data

BackGround

- **Medical Image Segmentation**

  Identify specific organs or tumors in medical imaging(e.g CT or MRI)

  - Conventional Deep Learning Based Approaches
    - CNN based (e.g U-Net, nnU-net)
    - Transformer based (e.g Swin UNETR, TransBTS, nnFormer)

  - Partial Label Problem
    - Contain labels only for specific organ or tumor
      - Performance degradation when applied to data from other hospitals
      - Organs without labels are mistakenly recognized as background that leads to error

➡ Aims to address this issue by utilizing the **Masked Back-Propagation technique**

BackGround

■ CLIP(Contrastive Language-Image Pre-training)

A vision-language model developed by OpenAI that aligns images and text in a shared space using contrastive learning[1)]

□ Role of CLIP in Medical Imaging

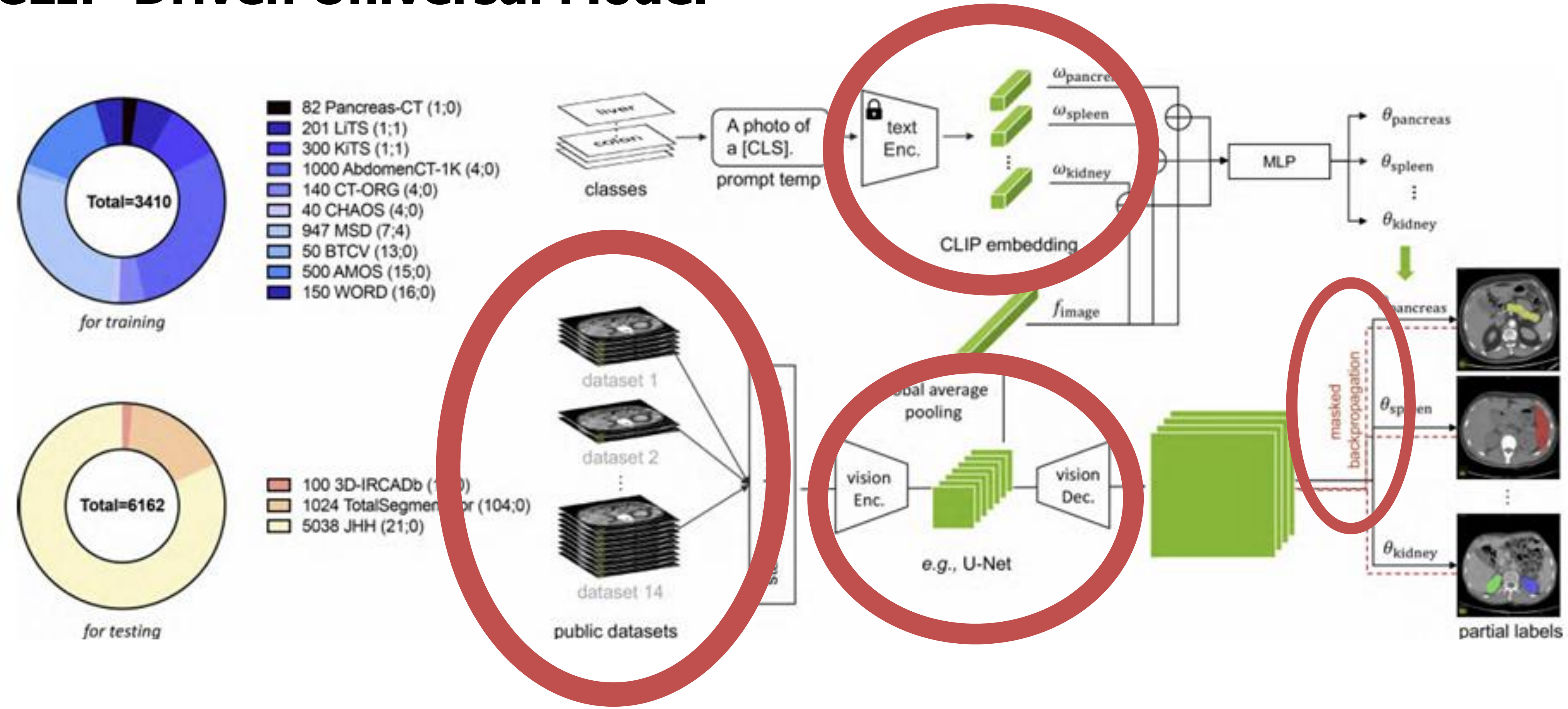
- Addressing one-hot encoding limitations & capturing organ-tumor relationships
- e.g) One-hot Encoding: “Liver” & "Liver Tumor‟ treated as independent classes
  CLIP: "A CT scan of the liver." and "A CT scan of a liver tumor.‟ → reflect close relationship
- Enhances segmentation performance by learning semantic similarity

➡ Designed for precise organ and tumor segmentation using CLIP-based label embeddings

[1)] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.

OverView

■ CLIP-Driven Universal Model



Liu, Jie, et al. "Clip-driven universal model for organ segmentation and tumor detection." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.

OverView

■ CLIP-Driven Universal Model

□ Training & Test Dataset

- Training Data: 14 public datasets with 3,410 CT scans containing partial labels for different organs and tumors.
- Test Data: 3 additional datasets with 6,162 CT scans for model validation

□ Text Branch

Generate CLIP-based text embeddings for **segmentation reflecting organ-tumor relationships**

- Method:
  - Utilize CLIP's text encoder with prompts in the format *"A photo of a [CLS]."*
    e.g) "A photo of a liver", "A photo of a kidney tumor"
  - Generate CLIP embeddings for each class
  - Pass the embeddings through an MLP network before sending them to the Vision Branch.

OverView

■ CLIP-Driven Universal Model

□ Vision Branch

Takes CT scans as input and **outputs organ and tumor segmentation results**

- **1) Standardized Processing**
  - Convert CT scans into a standardized format to enable training with multiple datasets
- **2) Vision Encoder**
  - Extract features from CT images using CNN- or Transformer-based networks (e.g., U-Net)
- **3) Global Average Pooling**
  - Summarize overall image features to prepare for integration with CLIP-based text embeddings
- **4) Vision Decoder**
  - Generate the final segmentation masks for each organ and tumor

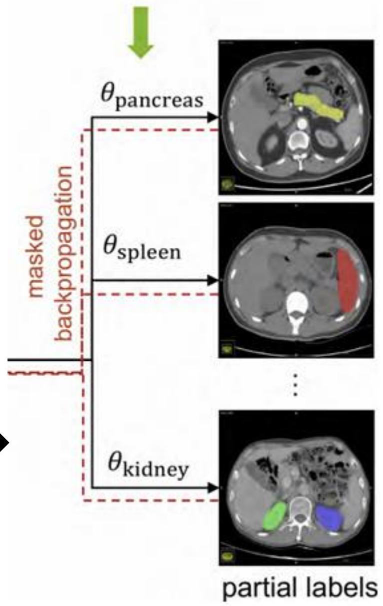
OverView

■ CLIP-Driven Universal Model

□ Masked Backpropagation

- Existing Partial Label Problem
  - e.g BTCV datasets labels liver, while the WORD dataset does not include liver labels.
- **Solution: Masked Backpropagation**
  - Compute loss **only for classes with labels** in each dataset
  - Exclude unlabeled classes from loss calculation to prevent unnecessary gradient updates
    - » allows effective training by combining **partially labeled datasets**

Visually represents the **Masked Backpropagation** technique,
where only specific classes (e.g., pancreas, spleen, kidney) are activated for training.



Results

	Existing Model (e.g Swin UNETR)	**Universal Model (proposed model)**	**Improvement**
MSD (Dice Score)	~82–84%	87.39%	+3~5%
BTCV (Dice Score)	~80–82%	86.13%	+4~6%
False Positives	Relatively high	Reduced	Lower false positives
Sen.	High FP risk in tumor detection	Maintains high sensitivity while reducing FP	More accurate tumor detection
Harm.	Lower (92.26% for pancreatic tumors)	92.59% for pancreatic tumors	+0.33%
Computation Speed	Baseline	6× faster	6× faster than Swin UNETR, 19× faster than nnU-Net
Generalization	Large performance variation	Maintains high performance	Strong generalization
Transferability	Optimized for specific datasets	Various diseases and datasets	More versatile

Interpretation

- **Addressing the Partial Label Problem**
  - Masked Back-Propagation technique
  - Utilization of 14 public datasets (3,410 CT scans)
    - Maximize generalization performance by training on diverse datasets

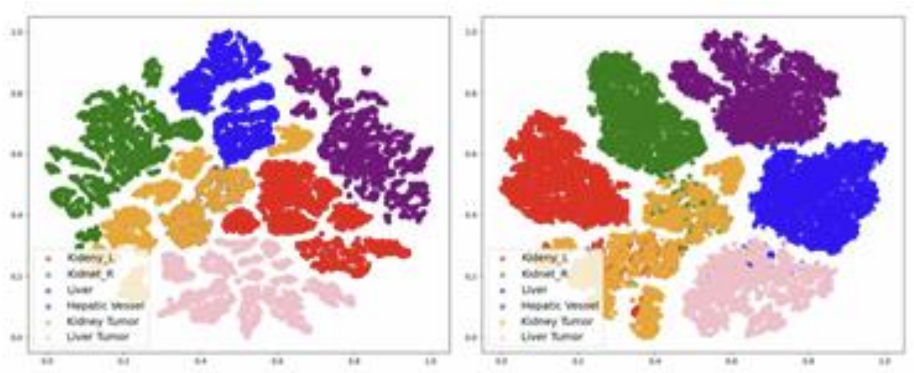
- **Leveraging CLIP Embeddings for Semantic Understanding**
  - CLIP-based embeddings
    - t-SNE visualization
      - shows that CLIP embeddings form a better feature space than one-hot encoding

- **Scalability and Efficiency of the Model**
  - Compatibility with various backbones (CNN, Transformer, etc.)
  - High performance relative to computational cost

Conclusion

- **Addressing the Partial Label Problem**
  - Universal Model trains from partially labeled diverse datasets
  - Use Masked Backpropagation
    - exclude loss calculation for unlabeled regions & enable better generalization

- **Effectiveness of CLIP Embedding**
  - CLIP embeddings → organs and tumors naturally cluster based on similarity


(a) One-hot label encoding (b) CLIP label encoding

Conclusion

- **Scalability and Computational Efficiency**
  - Designed to support both CNN (e.g., U-Net) and Transformer (e.g., Swin UNETR) backbones
    - enable application across various architectures
  - With reduced computational cost and 6× faster processing

- **Generalizability Across Different Datasets**
  - A generalizable medical AI model (Foundation Model)
    - maintains consistent performance across diverse environments

Considerations & Adaptation

- **Medical Prompt Design**
  - Generate appropriate CLIP-based text embeddings for anatomical structures
  - Text embeddings should capture relationships between the meniscus and surrounding tissues (bones, cartilage, etc.)
    - Need to experimentally optimized
      - e.g "A 3D MRI scan of the medial meniscus“"A segmented meniscus in a knee MRI image“

- **Data Preprocessing**
  - meniscus segmentation relies on MRI data, might require different preprocessing methods
  - Resolution and contrast characteristics of MRI must be considered

Considerations & Adaptation

- **Masked Backpropagation**
  - Some datasets might label only specific structures (Medial/Lateral Meniscus)
  - Masked Backpropagation enables effective learning on partially labeled datasets
    - experimental evaluation is needed to assess how Masked Backpropagation improves

- **Apply Various Backbone Models**
  - Compare CNN (U-Net) and Transformer (Swin UNETR) for Meniscus segmentation

- **Model Performance Evaluation**
  - Dice Similarity Coefficient (DSC) & Normalized Surface Distance (NSD)
    - use the same metrics to compare with existing models

Considerations & Adaptation

- **Key Experiments**
  - Evaluate how CLIP-based embeddings enhance Meniscus MRI segmentation
  - Compare with traditional One-Hot Encoding-based models
  - Verify whether Masked Backpropagation addresses the partial label problem