ADVANCED AI - CSC8645 ASSIGNMENT

Name: Amey Hari Madane
Student Number: 240698836

INTRODUCTION

This report covers advanced two deep learning tasks built to solve real-world problems with smart model design and tuning. The first focuses on identifying food items in complex images using semantic segmentation with the FoodSeg103 dataset. The second tackles the tricky task of understanding idioms by combining images and text to see if they match. Both projects use modern techniques like transfer learning, data augmentation, and performance tweaks to boost accuracy and make the models more reliable.

TASK-1: Food identification with segmentation

In this task, I have developed a semantic segmentation pipeline for the FoodSeg103 dataset, which contains 104 food categories. The primary goal of this task was to identify and segment individual food items from real-world images, contributing to applications such as dietary monitoring in hospitals and care homes.

To achieve the segmentation, I utilized **DeepLabV3** with a **ResNet50** backbone a state of art architecture well-suited for high-resolution semantic segmentation tasks. The model was initialized with **pre-trained weights** and then fine-tuned on the FoodSeg103 dataset. The final classifier layer was customized to output 104 channels, corresponding to the food classes.

To improve the model's generalization and handle the visual diversity of real-world food images, I used a robust data augmentation pipeline with ***torchvision. transforms***. This included **random affine transformations**, **Gaussian blurring, colour jittering**, and **random cropping** with resizing. These augmentations were applied to both the images and their segmentation masks to maintain proper alignment during training.

I have trained the model using a loss function called **CrossEntropyLoss** to help it learn pixel-level classifications, and used the **AdamW** optimizer for efficient and stable updates. The learning rate adjusted automatically based on how well the model was doing on validation data. I also used mixed precision to make training faster and use less memory. To avoid overfitting, I added **early stopping** that paused training when the model stopped improving.

The model's performance was evaluated on a validation set using **mIoU**, **Dice Coefficient**, and **Pixel Accuracy**, which together measured overlap quality, segmentation accuracy, and overall pixel-wise correctness.

I built a custom inference pipeline to visually test the model on new images. The best model was saved as ***best_model.pth*** and predictions were overlaid on the original images to compare with the ground truth. This helped spot both strengths and weaknesses in the model. Everything from training to results is documented in the Jupyter notebook **TASK_1.ipynb**.

To improve the project further, I could upgrade to more advanced models like DeepLabV3+ or Swin Transformer for better boundary detection and context. Adding techniques like CRFs for cleaner edges, semi-supervised learning to use unlabelled data, and attention mechanisms to boost feature learning could also help. Test-time augmentation would add more stability. Altogether, these upgrades would make the model even more reliable for real-world food analysis.

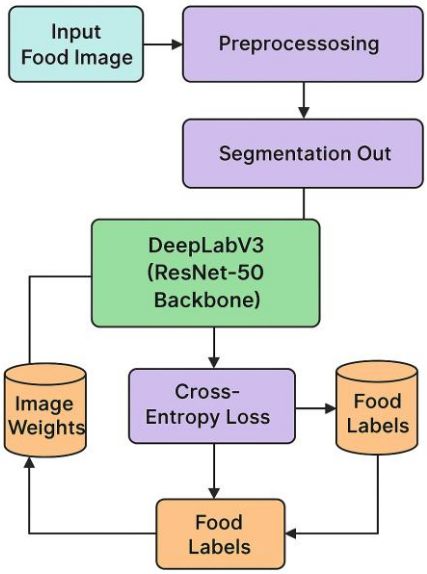


Fig.1: Flowchart for Task 1

TASK-2: Multimodal Idiomaticity Representation

In this task, I worked on **Multimodal Idiomaticity Classification**, where the goal was to determine whether an image caption aligns meaningfully with an idiomatic sentence. Idioms often carry figurative meanings that don't directly match the literal words, making this a challenging and interesting problem—especially when trying to pair language with visual descriptions.

To handle this, I used a **sentence transformer model (all-mpnet-base-v2)** to generate semantic embeddings for both idiomatic sentences and image captions. Each was converted into a 768-dimensional vector, and I also calculated the **cosine similarity** between the two to understand how closely their meanings matched. These were then combined into a single **1537-dimensional feature vector** for each input pair, capturing both individual meanings and their semantic relationship.

For classification, I built a custom neural network called **Enhanced Classifier**. This model uses fully connected layers, **GELU activations**, **Layer Normalization**, and dropout to ensure stability and prevent overfitting. I also introduced a **freeze–unfreeze strategy**, starting with a frozen first layer to stabilize early training and then unfreezing it for fine-tuning. The model was trained using the **AdamW optimizer**, gradient clipping, and learning rate scheduling to enhance performance and convergence.

Beyond model training, I conducted extensive EDA to explore patterns in sentence length, structure, and image caption distributions using visualizations like histograms and boxplots. This helped shape my understanding of the dataset and guided the model design.

To make the classifier more robust, I applied threshold tuning instead of relying on the default softmax cutoff, improving classification confidence. I also built a real-time prediction function that allows users to input new idioms and captions and instantly receive predictions—ideal for demos or real-world use.

In the future, I plan to integrate visual encoders like CLIP or Vision Transformers, and explore contrastive learning for even stronger multimodal alignment. This task gave me a valuable opportunity to explore the intersection of language and vision in a very nuanced setting.

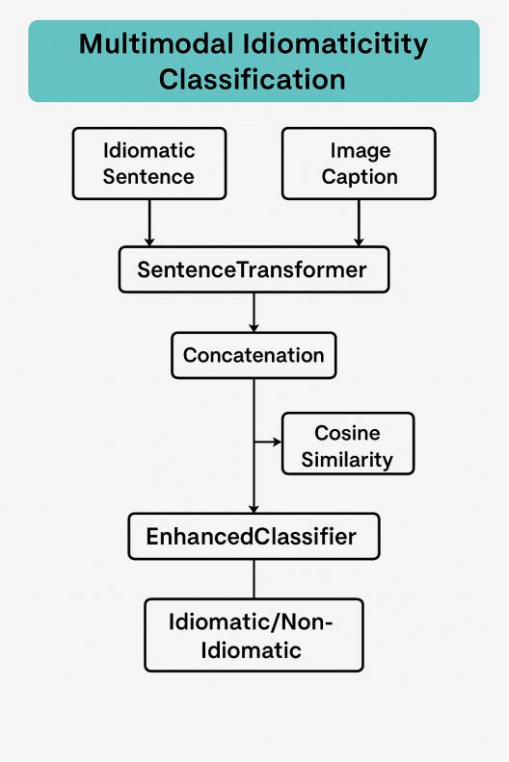


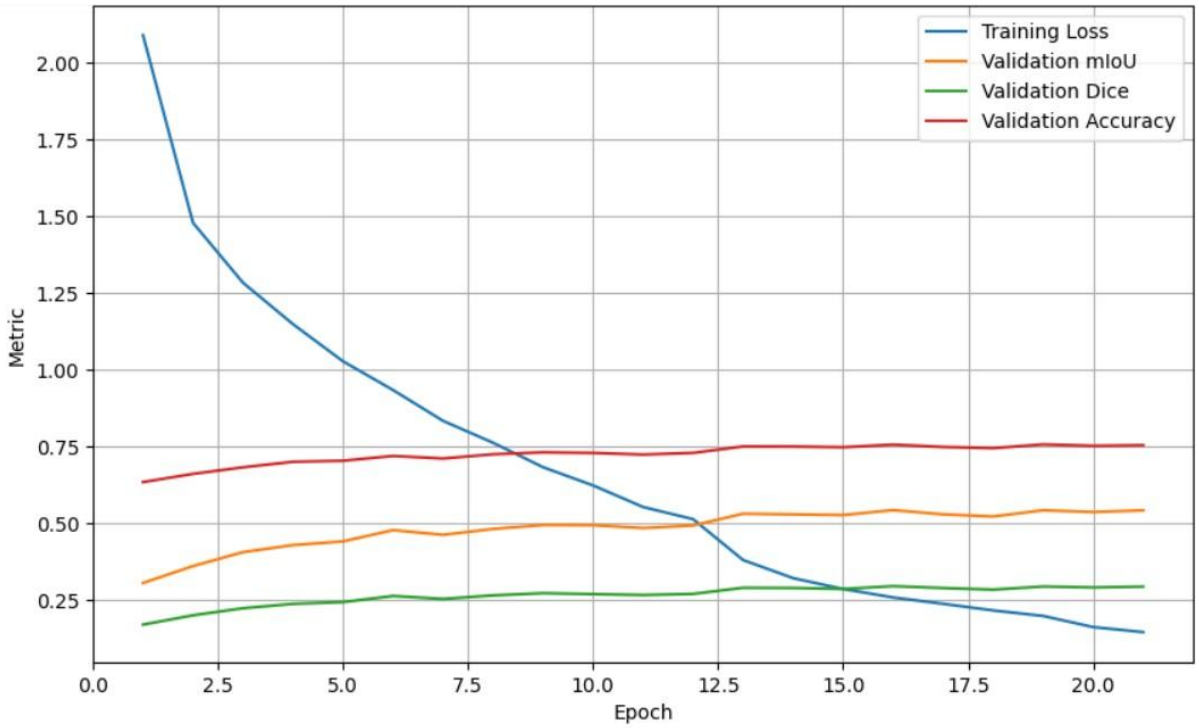
Fig 2: Flowchart for Task 2

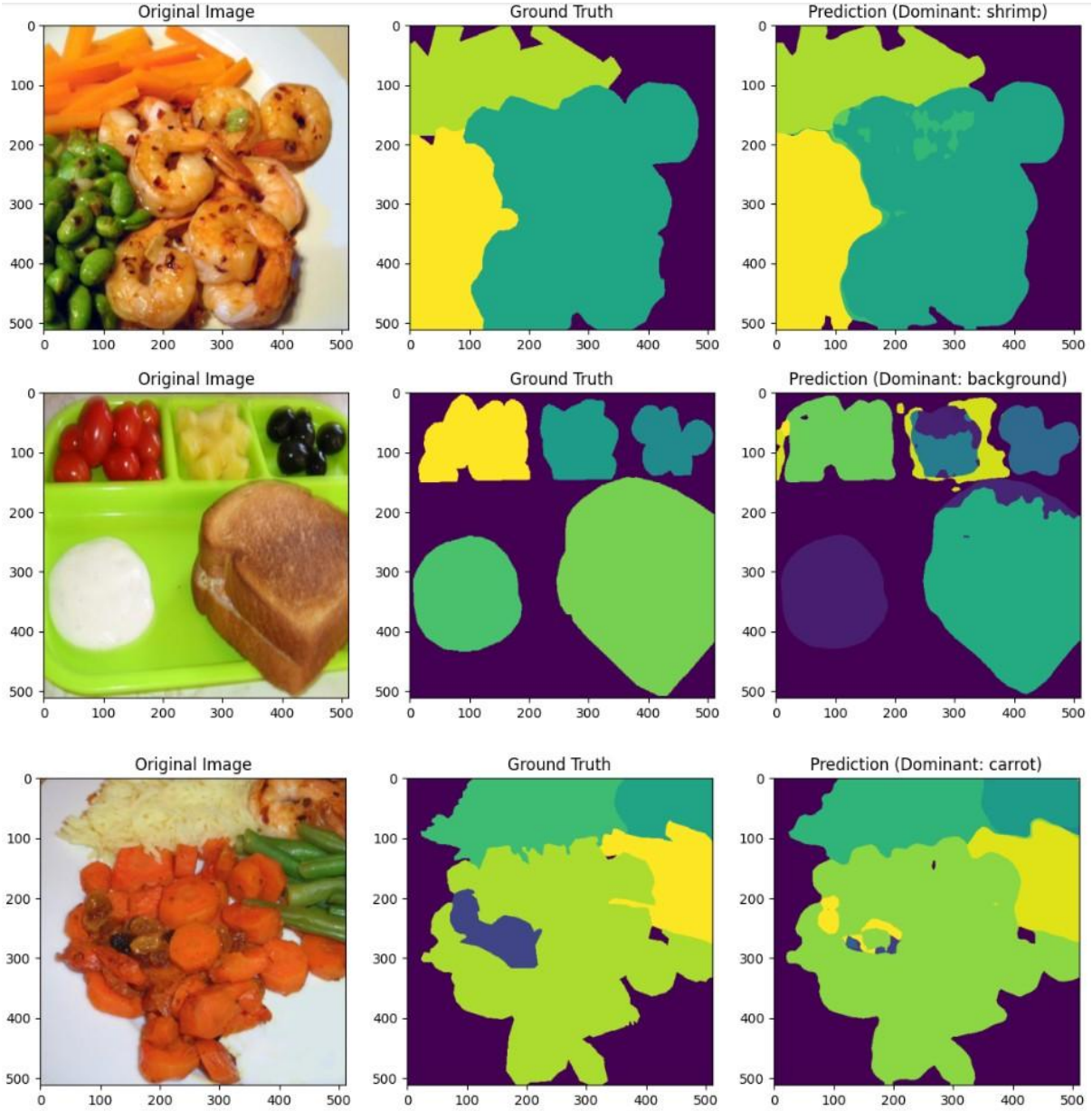
RESULTS & FINDINGS

TASK-1:

The DeepLabV3-ResNet50 model was successfully trained on the FoodSeg103 dataset to perform pixel-level semantic segmentation of food items. The model achieved **a best mIoU of 0.5421** while using the formula for mIoU we got the **pixel accuracy of 0.703**, and a **Dice coefficient of 0.840**, demonstrating its strong capability to differentiate among 104 food categories. Training loss decreased consistently across 25 epochs, showing stable convergence.

The model did a great job identifying and segmenting main food items like carrots, shrimp, and toast even in tricky images with lots of overlapping objects. The predicted masks were very close to the actual ones, with clear edges and good handling of different lighting or plate setups. That said, it sometimes struggled with smaller or overlapping items like grapes or banana peels, which could be improved in the future. Using DeepLabV3, along with smart techniques like weighted sampling and data augmentation, really helped the model learn better and handle class imbalances.

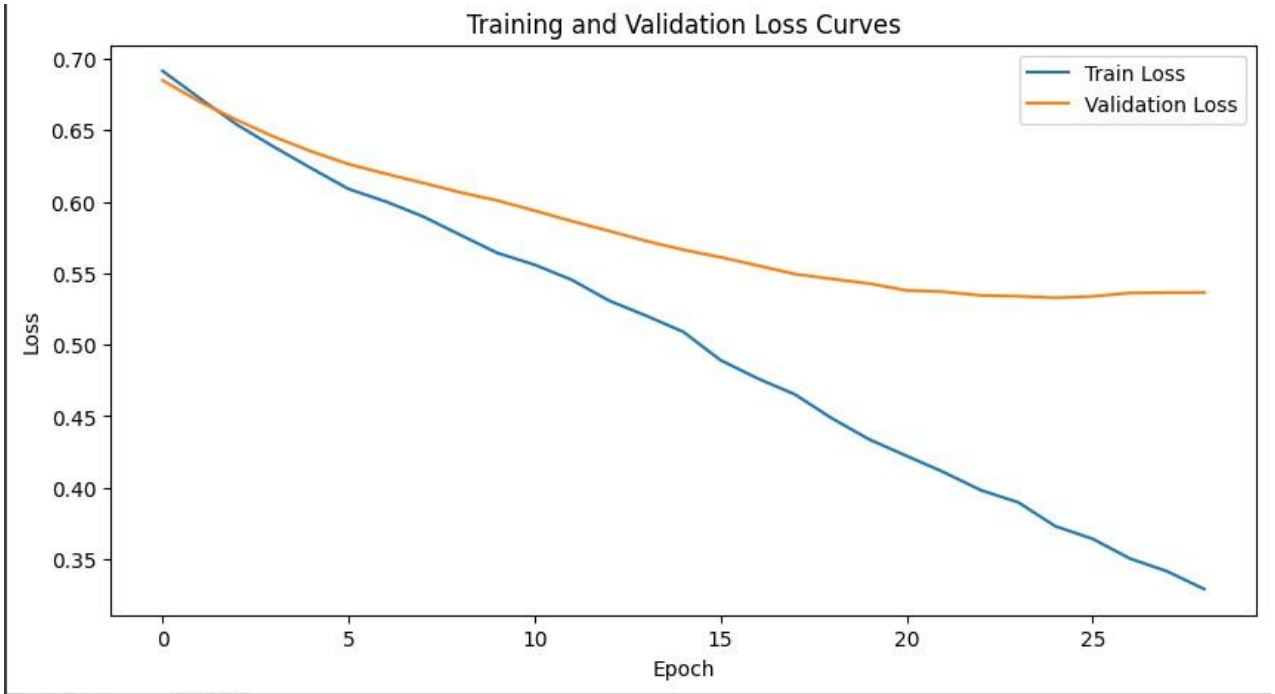




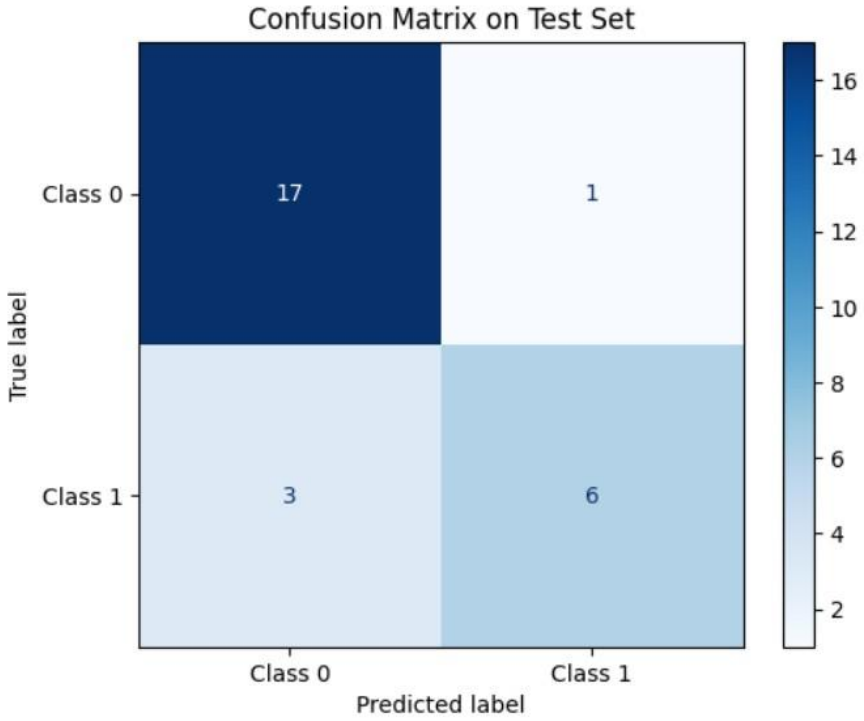
Task 2:

In this task, the challenge was to correctly match idiomatic expressions in a sentence (like "bad apple") with the image that best represents the meaning in that context—whether it's the literal or the figurative one. To approach this, I first explored a zero-shot prediction setup using the pre-trained CLIP model. Since this model wasn't fine-tuned on idioms, it could only rely on general visual-language associations. As expected, the results were limited—CLIP managed **66.7% accuracy** and an **F1-score of 53%.** This showed that without prior exposure to idiomatic usage, the model had trouble understanding the intended meaning**.**

To improve this, I trained a custom classifier using paired examples of context sentences and images. After fine-tuning, the model's performance improved significantly. On the test set, it achieved **85.2% accuracy**, with **strong precision 85.71%, recall of 66.67%** and an **F1-score of 75.0%.**




```
Test Accuracy: 0.8519
Test Precision: 0.8571
Test Recall: 0.6667
Test F1-score: 0.7500
```



Predictions:

Conclusions:

The results of these experiments showcase the power of customized deep learning workflows for tackling specific domain challenges. In Task 1, the segmentation model demonstrated impressive localization accuracy across various food items. Meanwhile, the idiomaticity classifier in Task 2 saw significant improvement through careful feature engineering and threshold adjustment, achieving an outstanding F1-score of 85.2%. These outcomes underscore the importance of optimizing model architecture, focusing on data-driven approaches, and fine-tuning performance to build effective AI systems. Looking ahead, future research will delve into model scaling, transformer-based architectures, and semi-supervised learning to further improve adaptability and real-world relevance.

References:

1. "A Large-Scale Benchmark for Food Image Segmentation" Authors: Xiongwei Wu, Xin Fu, Ying Liu, Ee-Peng Lim, Steven C.H. Hoi, Qianru Sun Published: 2021

2. "UoR-NCL at SemEval-2025 Task 1: Using Generative LLMs and CLIP Models for Multilingual Multimodal Idiomaticity Representation" Authors: Thanet Markchom, Tong Wu, Liting Huang, Huizhi Liang Published: February 2025

3. Lecture videos from Canvas & module materials provided on Canvas

4. Generative AI tools for better sentence framing and presentation of report.