Optimizing R2U-Net for Skin Cancer Segmentation with Attention and Hybrid Loss

Sreejita Das

Advisor: **Dr. Kaustuv Nag**

Department of Computer Science and Engineering
Indian Institute of Information Technology Guwahati

This report is submitted for the course of
CS300 : Project-I

IIIT Guwahati March 2025

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this report are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this or any other university. This report is my own work and contains nothing that is the outcome of work done in collaboration with others except as specified in the text and Acknowledgements.

Sreejita Das
Roll: 2201201,
3rd year, Bachelors of Technology,
Department of Computer Science and Engineering,
Indian Institute of Information Technology Guwahati.

Acknowledgements

I express my sincere gratitude to Dr. Kaustuv Nag for their expert guidance and steadfast support throughout this project. I am also thankful to the faculty of the Department of Computer Science and Engineering at the Indian Institute of Information Technology Guwahati for their valuable mentorship. Additionally, I appreciate the support of my peers during this academic endeavor. Lastly, I extend my gratitude to my parents and family for their consistent encouragement and support throughout my studies.

Abstract

Accurate segmentation of skin cancer lesions is crucial for early diagnosis but remains challenging due to class imbalance and the need for precise boundary delineation. This project advances the R2U-Net model, a neural network designed for medical image segmentation, by integrating attention gates and a hybrid loss function combining Dice and weighted Binary Cross-Entropy (BCE) losses. These enhancements sharpen the model's focus on lesion regions and optimize its performance in balancing overlap accuracy with pixel-level precision. Evaluations on subsets of the ISIC 2017 dataset reveal substantial improvements in segmentation accuracy, underscoring the approach's potential for real-world impact. This work delivers a practical solution for enhancing medical image segmentation and establishes a robust foundation for future advancements in the field.

Table of Contents

List of Figures

Chapter 1

Introduction

Medical image segmentation is a critical process in healthcare, enabling the precise identification of anatomical structures in medical images. U-Net [4] revolutionized medical image segmentation with its encoder-decoder architecture. Building on U-Net, RU-Net [1] introduced recurrent convolutional units to enhance feature accumulation over time. R2U-Net [1] further advanced this by combining recurrent and residual units, boosting feature extraction and training stability.

1.0.1 Problem Statement

Accurate segmentation of skin cancer lesions is essential for early diagnosis and treatment, yet it remains a challenge in medical imaging. Two obstacles hinder this task: **class imbalance**, and the need for **precise boundary detection** of irregular lesion edges. While R2U-Net excels in feature extraction and training stability, it struggles to capture fine details—like subtle textures—and maintain accuracy across imbalanced classes, often causing over-segmentation or missed boundaries in complex cases

1.0.2 Solution

We propose an enhanced R2U-Net framework by incorporating **attention gates** and a **hybrid loss function** that blends Dice and weighted Binary Cross-Entropy (BCE) losses. Attention gates [**?**], refine skip connections to prioritize lesion regions, sharpening the model's focus on subtle features. Simultaneously, the hybrid loss improves overlap accuracy and mitigates class imbalance more effectively than traditional loss functions.

1.1 Objective

This study aims to enhance the R2U-Net model, a neural network designed for medical image segmentation, to improve the accuracy and reliability of skin cancer lesion segmentation. The investigation focuses on two primary objectives: **first**, to integrate attention gates and a hybrid loss function—combining Dice and weighted Binary Cross-Entropy (BCE) losses—into the R2U-Net architecture, addressing challenges such as class imbalance and precise boundary detection; and **second**, to evaluate the effectiveness of these enhancements in improving segmentation performance, particularly in terms of overlap accuracy and pixel-level precision. By analyzing the impact of attention gates and the hybrid loss on R2U-Net's ability to segment skin cancer lesions, this research seeks to contribute to the development of more accurate diagnostic tools and offer insights into optimizing neural network components for medical imaging.

1.2 Motivation

The timely identification of skin cancer, notably melanoma, plays a pivotal role in enhancing patient prognosis, as early intervention markedly elevates treatment success rates. Traditional diagnostic approaches, however, depend extensively on manual segmentation of dermoscopic images, a process that is both labor-intensive and subject to inconsistencies. These challenges underscore the pressing demand for automated and precise segmentation techniques to support efficient diagnosis. This project is driven by the potential to refine the R2U-Net model—a cutting-edge convolutional neural network tailored for medical image segmentation—by overcoming its limitations in managing class imbalance and delineating intricate lesion boundaries. Through the incorporation of attention gates and a hybrid loss function, this work seeks to develop an advanced model that enhances segmentation accuracy while offering a practical, scalable tool for clinical deployment.

Chapter 2

Related Work

2.1 Literature Survey

2.1.1 U-Net

Ronneberger et al. [4] introduced U-Net in 2015, a pioneering model for medical image segmentation that uses an encoder-decoder architecture with skip connections to preserve spatial details, making it effective for skin cancer lesion segmentation [4].

2.1.2 3D U-Net

Çiçek et al. [5] presented 3D U-Net in 2016, extending U-Net for volumetric medical image segmentation by learning from sparsely annotated 3D data, applicable to tasks requiring spatial depth [5].

2.1.3 V-Net

Also in 2016, Milletari et al. [3] introduced V-Net, a 3D fully convolutional network with residual connections for volumetric segmentation, incorporating a Dice loss to address class imbalance in medical imaging [3].

2.1.4 RU-Net and R2U-Net

Alom et al. [1] proposed RU-Net and R2U-Net in 2018, enhancing U-Net with recurrent convolutional layers (RU-Net) and recurrent-residual units (R2U-Net) for better feature accumulation and training stability in skin cancer segmentation [1].

2.2 Primary Model : R2UNet

Recurrent U-Net (RU-Net) and Recurrent Residual U-Net (R2U-Net), introduced by Alom et al. [1], are advanced extensions of U-Net, designed to improve feature extraction for tasks like skin cancer lesion segmentation on the ISIC 2017 dataset [1]. RU-Net incorporates Recurrent Convolutional Layers (RCLs) into its encoder-decoder framework, enabling feature accumulation over time steps (e.g., $t = 2$ or $t = 3$) to capture fine details in dermoscopic images[1].

R2U-Net builds on RU-Net by integrating RCLs with residual connections, forming Recurrent Residual Convolutional Units (RRCUs), which enhance training stability and robustness while maintaining the same number of parameters as U-Net [1]. The RRCUs combine recurrent feature accumulation with residual learning, mitigating the vanishing gradient problem and improving performance on fine lesion details, though class imbalance remains a challenge [1].

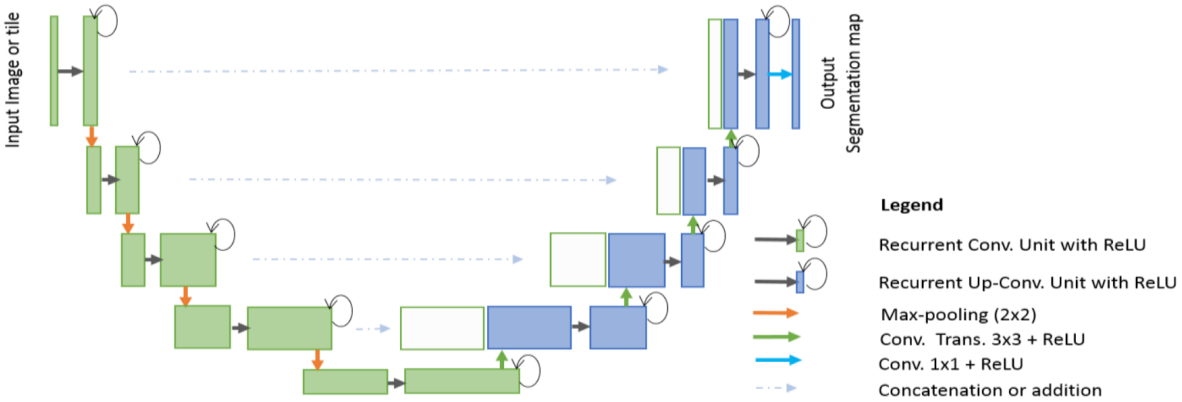

Fig. 2.1 R2U-Net architecture using recurrent residual convolutional units

2.3 Working Principle

R2U-Net processes input dermoscopic images through its encoder-decoder architecture to produce segmentation maps, leveraging Recurrent Residual Convolutional Units (RRCUs) in each convolutional block [1]. As shown in Fig. 2.2, the encoder downsamples the input image using max-pooling, while the decoder upsamples feature maps via up-convolution, with skip connections preserving spatial details.

Within each RRCU, the Recurrent Convolutional Layers (RCLs) first accumulate features over time steps (e.g., $t = 2$ or $t = 3$). For an input x_l at the l-th layer, the RCL output at time step t for a pixel at position (i, j) on the k-th feature map is:

$$O_{ijk}^{l}(t) = (w_k^f)^T * x_l^{f(i,j)}(t) + (w_k^r)^T * x_l^{r(i,j)}(t-1) + b_k,$$

where $x_l^{f(i,j)}(t)$ and $x_l^{r(i,j)}(t-1)$ are inputs to the standard and recurrent convolutions, w_k^f and w_k^r are their respective weights, and b_k is the bias [1]. This output is passed through a ReLU activation, $\mathcal{F}(x_l, w_l) = \max(0, O_{ijk}^{l}(t))$. The RRCU then applies a residual connection, computing the block's output as:

$$x_{l+1} = x_l + \mathcal{F}(x_l, w_l),$$

where x_l is the input to the RRCU [1]. The recurrent operation enhances feature accumulation, while the residual connection stabilizes training by allowing gradients to flow directly through the network, enabling R2U-Net to capture fine lesion details.

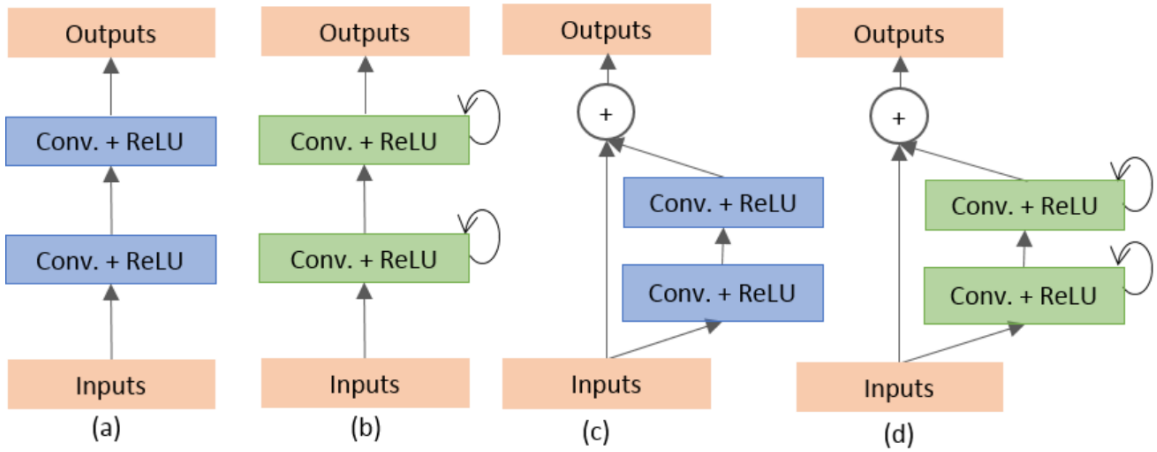

Fig. 2.2 Different variant of convolutional and recurrent convolutional units (a) Forward convolutional units, (b) Recurrent convolutional block (c) Residual convolutional unit, and (d) Recurrent Residual convolutional units (RRCU).

2.4 Notations

2.5 Dataset

The dataset used in this project is the International Skin Imaging Collaboration (ISIC) 2017 dataset, which is a widely recognized benchmark for skin lesion segmentation tasks [2]. It consists of dermoscopic images of skin lesions, annotated with ground truth masks for segmentation. The dataset is divided into three subsets: training, validation, and test sets, as summarized in Table 2.1.

Table 2.1 Distribution of the ISIC 2017 Dataset

Data	**Sample Images**	**Ground Truth**
Training	2000	2000
Validation	150	150
Test	600	600

A sample image from the ISIC 2017 dataset, along with its ground truth mask, is shown in Fig. 2.3. This dataset is particularly relevant to this project as it provides a standardized benchmark for evaluating the effectiveness of R2U-Net and its enhancements in skin cancer lesion segmentation.

Fig. 2.3 A sample dermoscopic image from the ISIC 2017 dataset (bottom) and its corresponding ground truth segmentation mask (top).

2.6 Evaluation Metrics

To assess the performance of R2U-Net for medical image segmentation, several standard evaluation metrics are employed, as outlined in Alom et al. [1]. These metrics are particularly relevant for evaluating segmentation models on the ISIC 2017 dataset, where precise delineation of skin cancer lesions is critical. The metrics used, along with their formulas, are as follows:

- **Accuracy (AC)**: Represents the proportion of correctly classified pixels (both lesion and background) in the image. It is calculated as:

$$AC = \frac{TP + TN}{TP + TN + FP + FN},$$

  where TP is True Positives, TN is True Negatives, FP is False Positives, and FN is False Negatives.

- **Sensitivity (SE)**: Also known as the true positive rate, this metric evaluates the model's ability to correctly identify lesion pixels. It is given by:

$$SE = \frac{TP}{TP + FN},$$

- **Specificity (SP)**: Also referred to as the true negative rate, this metric assesses the model's ability to correctly identify background pixels, helping to minimize false positives in segmentation. It is expressed as:

$$SP = \frac{TN}{TN + FP},$$

- **Dice Coefficient (DC)**: Measures the overlap between the predicted segmentation and the ground truth, emphasizing the similarity between the two regions. The DC is defined as:

$$DC = 2\frac{|GT \cap SR|}{|GT| + |SR|},$$

  where GT refers to the ground truth and SR refers to the segmentation result.

- **Jaccard Similarity (JS)**: Also known as the Intersection over Union (IoU), this metric quantifies the ratio of the intersection of the predicted and ground truth regions to their union, providing a measure of spatial overlap. It is calculated

as:

$$JS = \frac{|GT \cap SR|}{|GT \cup SR|},$$

Together, these metrics provide a comprehensive evaluation of R2U-Net performance, balancing the need for accurate lesion detection with the minimization of false positives and negatives, which is essential for clinical applications in skin cancer segmentation.

2.7 Notations

The negative log-likelihood at the initial step is defined as $L_0 = -\log p_\theta(x_0|x_1)$ in eq(**??**). Subsequently, the divergence between the approximate posterior and the conditional distribution is captured by $L_{t-1} = D_{\mathrm{KL}}(q(x_{t-1}|x_t, x_0) \,||\, p_\theta(x_{t-1}|x_t))$ in eq(**??**). The overall divergence for the entire diffusion process is represented as $L_T = D_{\mathrm{KL}}(q(x_T|x_0) \,||\, p(x_T))$ in eq(**??**).

2.8 Loss function

To represent the mean $\mu_\theta(x_t, t)$, we propose a specific parameterization motivated by the analysis of L_t. With $p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 I)$, we can write:

$$L_{t-1} = \mathbb{E}\left[\frac{1}{2\sigma_t^2} \parallel \bar{\mu}_t\left(\boldsymbol{x}_t, \boldsymbol{x}_0\right) - \mu_\theta\left(\boldsymbol{x}_t, t\right) \parallel^2\right] + C$$

$$L_{simple} = \mathbb{E}_{t, x_0, \epsilon}\left[\parallel \epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t\right) \parallel^2\right]$$

2.9 Notations

The loss at time step $t-1$, denoted as L_{t-1}, is expressed as the expected squared Euclidean distance between $\bar{\mu}_t(x_t, x_0)$ and $\mu_\theta(x_t, t)$, normalized by $2\sigma_t^2$. The expectation is taken over relevant variables. The term C is an additional constant.

The simple loss, denoted as L_{simple}, is defined as the expected squared Euclidean distance between ϵ and $\epsilon_\theta\left(\sqrt{\bar{\alpha}_t x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t}\right)$, where the expectation is taken over x_0, t, and ϵ.

2.10 Training

Algorithm 1 Training Algorithm

1: **while** not converged **do**
2: $x_0 \sim q(x_0)$
3: $t \sim \text{Uniform}(1, \ldots, T)$
4: $\epsilon \sim \mathcal{N}(0, 1)$
5: Take gradient descent step on $\nabla_\theta \left\| \epsilon - \epsilon_\theta \left(\sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, t \right) \right\|^2$
6: **end while**

The algorithm repeatedly samples an initial value, time step, and random noise. Then, it performs a gradient descent step on a function involving the gradient with respect to parameters. This process continues until the convergence criterion is met. The specific details of the convergence criteria, the functions involved, and the meaning of the parameters like α_t and $\bar{\alpha}_t$ would need to be provided in a more detailed context or algorithmic description.

2.11 Sampling

Algorithm 2 Sampling Algorithm

1: $x_T \sim \mathcal{N}(0, I)$
2: **for** $t \leftarrow T$ to 1 **do**
3: $z \sim \mathcal{N}(0, I)$ if $t > 1$, else $z = 0$
4: $\boldsymbol{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\boldsymbol{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \alpha_t}} \epsilon_\theta(\boldsymbol{x}_t, t) \right) + \sigma_t \boldsymbol{z}$
5: **end for**
6: **return** X

The algorithm initializes with a sample, and then iteratively generates new samples based on a recursive formula that involves noise and previous samples. The parameters α_t and σ_t might be specified or updated at each iteration.

This methodology presents high-quality image samples through diffusion models, revealing connections with variational inference, denoising score matching, annealed Langevin dynamics, autoregressive models, and progressive lossy compression. Although diffusion models show significant inductive biases when applied

to image data, there is a curiosity analyzing how effectively they perform with different types of data as well as the way they might be integrated into various generative models and machine learning systems.

In this study, another paper is examined[**?**], which addresses limitations in the original methodology by enhancing sampling efficiency through learned variances, achieving competitive log-likelihoods, and demonstrating seamless scalability with model capacity and training compute. These refinements collectively contribute to a more robust and versatile probabilistic model.

$$L_{\text{vlb}} = L_0 + L_1 + \ldots + L_{T-1} + L_T \tag{2.1}$$

Sampling arbitrary steps during training in a forward noising process, parameterizing the mean function $\mu_\theta\left(\boldsymbol{x}_t, t\right)$ using neural networks.

$$\begin{aligned} \mu_\theta\left(\boldsymbol{x}_t, t\right) &= \frac{1}{\sqrt{\alpha}}\left(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta\left(\boldsymbol{x}_t, t\right)\right) \\ L_{simple} &= \mathbb{E}_{t,x_0,\epsilon}\left[\parallel \epsilon - \epsilon_\theta\left(\boldsymbol{x}_t, t\right) \parallel^2\right] \end{aligned} \tag{2.2}$$

We define a new hybrid objective L_{hybrid} using (2.1) and (2.2).The hybrid objective achieves better log-likelihoods on the training set. With this important sampled objective, we can achieve our best log-likelihoods by optimizing L_{vlb}

$$L_{\text{hybrid}} = L_{\text{simple}} + \lambda L_{\text{vlb}} \tag{2.3}$$

We use the variational lower-bound (VLB) to modify log-likelihoods and show the competitiveness of Denoising Diffusion Probabilistic Models (DDPMs) on high-diversity datasets such as ImageNet. By using a simple reparameterization and a hybrid learning objective that combines VLB with a more straightforward method, our models outperform log-likelihoods obtained from direct optimization. Interestingly, the latter shows increased noise on gradients during training, which is reduced by applying a straightforward importance sampling strategy to achieve better log-likelihoods.

Our models remarkably attain comparable sample quality with fewer sampling steps upon integrating learned variances. In comparison, DDPM requires a large number of forward passes; this reduces the sampling procedure to as little as 50 passes, improving its practical usefulness. This accomplishment is in line with

ongoing initiatives to investigate other strategies for accelerating sample procedures. Although diffusion models are already at the cutting edge, they are still inferior to GANs when it comes to difficult generation datasets. In this study, another paper is examined [**?**] GANs, with their adversarial architecture, achieve diverse and high-quality samples, but face a trade-off between diversity and fidelity. To bring these advantages to diffusion models, efforts focus on enhancing architecture and devising a scheme for balancing diversity and fidelity.

The referred to architecture improvements considerably increase FID. shows the tunability of a single hyperparameter for the diversity-fidelity trade-off by introducing a strategy that uses classifier gradients to control diffusion model sampling. Significant is the significant rise of the gradient scale factor without adversarial consequences. Our enhanced architecture performs exceptionally well in conditional synthesis and unconditional picture synthesis when classifier guidance is used. Interestingly, FIDs comparable to BigGAN are maintained with as few as 25 forward passes under the classifier direction. We present the combined efficacy for better outcomes on ImageNet at 256x256 and 512x512 resolutions by comparing our models to upsampling stacks.

Emphasizing the empirical superiority of the mean-squared error objective L_{simple} over the variational lower bound L_{vlb}, showcasing its practical effectiveness. This approach employs Langevin dynamics within a denoising model, demonstrating remarkable performance in generating high-quality image samples. The term "diffusion models" is used broadly to encompass both categories of models. To address limitations in scenarios with fewer diffusion steps, we propose a neural network parameterization for adjusting the variance $\Sigma_\theta(\boldsymbol{x}_t, t)$. The neural network output v is then integrated into an interpolation scheme.

$$\Sigma_\theta(\boldsymbol{x}_t, t) = \exp(v \log \beta_t + (1 - v) \log \tilde{\beta}_t) \tag{2.4}$$

2.12 Notations

$\Sigma_\theta(\boldsymbol{x}_t, t)$: Covariance matrix parameterized by θ. v: Balance parameter in the equation. β_t: Variable. $\tilde{\beta}_t$: Determines the final form of the covariance matrix.

2.13 Algorithm

Algorithm 3 Classifier-Guided Diffusion Sampling

Require: Diffusion model parameters $(\mu_\theta(x_t), \Sigma_\theta(x_t))$, classifier $p_\phi(y|x_t)$, gradient scale s

function CLASSIFIERGUIDEDDIFFUSIONSAMPLING(y, s)
  $x_T \leftarrow$ sample from $N(0, I)$
  for $t \leftarrow T$ **to** 1 **do**
    $\mu, \Sigma \leftarrow \mu_\theta(x_t), \Sigma_\theta(x_t)$
    $x_{t-1} \leftarrow$ sample from $N\left(\mu + s\Sigma\nabla_{x_t} \log p_\phi(y|x_t), \Sigma\right)$
  end for
  return x_0
end function

Algorithm 4 Classifier-Guided DDIM Sampling

Require: Diffusion model $\epsilon_\theta(x_t)$, classifier $p_\phi(y|x_t)$, gradient scale s

function CLASSIFIERGUIDEDDDIMSAMPLING(y, s)
  $x_T \leftarrow$ sample from $N(0, I)$
  for $t \leftarrow T$ **to** 1 **do**
    $\hat{\epsilon} \leftarrow \epsilon_\theta(x_t) - \sqrt{1-\bar{\alpha}_t}\nabla_{x_t} \log p_\phi(y|x_t)$
    $x_{t-1} \leftarrow \sqrt{\bar{\alpha}_{t-1}}\left(x_t - \sqrt{\frac{\bar{\alpha}_t}{1-\bar{\alpha}_t}}\hat{\epsilon}\right) + \sqrt{1-\bar{\alpha}_{t-1}}\hat{\epsilon}$
  end for
  return x_0
end function

Fig. 2.4 Samples from BigGAN-deep with truncation 1.0 (FID 6.95, left) vs samples from our diffusion model with guidance (FID 4.59, middle) and samples from the training set (right).

Chapter 3

Proposed Model[?]

3.1 Finetuning Neural Networks

The standard approach to enhance a neural network is to train it again using more data. But this simple continuance may result in disastrous forgetting, mode collapse, and overfitting. A great deal of research has gone into creating solutions that are optimized to lessen these problems in order to overcome these concerns.

3.2 HyperNetwork

This mechanism involves training a small recurrent neural network to influence the weights of a larger one. HyperNetworks implementation for Stable Diffusion to alter the artistic style of the output images.

3.3 Adapter

In Natural Language Processing (NLP), methods that use adapters are often used to customize transformer models that have already been trained for different applications by adding new module layers. Adapters are used in incremental learning and domain adaption in computer vision. This method is frequently used in conjunction with CLIP to apply backbone models that have already been trained to a variety of tasks. Notably, adapters have proven successful in ViT adapters and vision transformers. T2IAdapter adjusts Stable Diffusion to external circumstances in parallel.

3.4 Additive Learning

The method uses prunes, hard attention, or learning weight masks to introduce a few new parameters and freezes the original model weights to prevent forgetting. By using a side branch model and combining its outputs with a preset schedule, Side-Tuning learns new capabilities in addition to a frozen model and an additional network.

3.5 Low-Rank Adaptation (LoRA)

This prevents catastrophic forgetting by learning the offset of parameters with low-rank matrices, based on the observation that many over- parameterized models reside in a low intrinsic dimension subspace.

3.6 Zero-Initialized Layers

References

[1] Alom, M. Z., Hasan, M., Yakopcic, C., Taha, T. M., and Asari, V. K. (2018). Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. *arXiv preprint arXiv:1802.06955*. 1, 3, 4, 5, 7

[2] ISIC Archive (2017). ISIC 2017: Skin Lesion Analysis Towards Melanoma Detection. https://challenge.isic-archive.com/landing/2017. 6

[3] Milletari, F., Navab, N., and Ahmadi, S.-A. (2016). V-net: Fully convolutional neural networks for volumetric medical image segmentation. *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571. 3

[4] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241. 1, 3

[5] Çiçek, , Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). 3d u-net: Learning dense volumetric segmentation from sparse annotation. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 424–432. 3