AI02 – NIGHT VISION

LOW LIGHT MEDIA ENHANCEMENT

PROJECT MEMBER APPLICATION

Section A - Managerial Questionnaire

Essentials:

My motivation to join AI club project member is to gain valuable insights and skills that will allow me to actively contribute to projects and get a platform to study more about the projects which AI club brings.

I am fit for this job because I love how AI is able to perform everything and very curious to know all the nook and corners of the mechanism that AI is implementing in the modern world And I am well confident that this curiosity will automatically push me to learn more and be a responsible project member for the team. I have been learning various aspects AI/ML and Pytorch for the past year and I am eager to learn the math behind the ML. I am excited about the opportunity to explore AI's capabilities further and discover how it can be applied to various real-world challenges like the **Low light media enhancement.**

Commitments/PoRs:

1. I am ready to spend 2-4 hr for the AI club per day, ie 15- 16 hr per week. I can also dedicate my time in the weekends if there are any deadline or problem in projects.
2. I am not planning to take up any PoRs in the next year so far, apart from being the AI club PM :) If I find any contest or projects related to AI/ML, I might take a look into it depending upon the scenario (only if I have time to dedicate both AI club and this, if not, AI club project member is my first priority)

Section B: Common Technical Questionnaire

Quantile Regression: **Colab_link**

Section C: Project Specific Questionnaire

1.1 RETINEX

1. Several image enhancement techniques exist for low light image enhancement. Why do you think Retinex is extensively used in deep learning techniques?

With the development of deep learning, CNNs have been applied in low-light image enhancement which are processed mainly into two categories. The first category

directly employs a CNN to learn **mapping function** from the low-light image to its normal-light counterpart, thereby ignoring human colour perception. The second category is the **Retinex theory** (applied on reflectance ie underlying the scene details) where they employ different CNNs to decompose colour image, denoise the reflectance, and adjust the illumination (global lighting conditions), respectively. Retinex models separate illumination and reflectance, allowing targeted adjustments to each component.

Deep learning models benefit from Retinex-based techniques due to their ability to learn complex mappings from input images to illumination and reflectance components. When these are targeted and trained separately, the deep learning networks implements the end-to-end learning, allowing joint optimization of illumination adjustments and reflectance.

The illumination map and reflectance image allow us to manipulate lighting effects and get intrinsic scene details. Whether in photography, computer vision, or image processing, these concepts play a crucial role in enhancing visual content.

2. Elaborate on the illumination map and reflectance image. what does the term corruption imply in the context of Retinex?

   Illumination Map:

   - The illumination map represents the **global lighting conditions** in an image. It characterizes where and how much light falls on the scene and shows how it's affecting the **overall brightness** and contrast.
   - As it deals with the brightness of the images, it is typically a **grayscale image**. This is because the grayscale has only one channel in the filter (which determines the intensity) making them simpler and focus solely on the brightness across the image. Also, colour can be a problem when it comes to lighting (ie sunlight/artificial light can introduce different colour cast) and may affect the brightness and intensity of the images.
   - Bright regions in the illumination map correspond to well- lighted areas, while dark regions represent shadows or low-light regions.
   - The illumination map captures both direct and indirect lighting (e.g., ambient light, reflections).
   - In computer vision and image processing, separating the illumination map from the original image helps enhance details and improve visual quality.

   Reflectance image:

   - The reflectance image captures the underlying scene details without the influence of lighting. It isolates the intrinsic properties of the objects in the scene.

- It represents the material properties, textures, and colours of the objects.
- Bright regions in the reflectance image correspond to strong scene details, while dark regions indicate less-detailed areas.
- Separating the reflectance image allows us to enhance fine structures and correct for uneven lighting.

In the deep learning-based methods, the Retinex theory is adopted by first treating the estimated reflectance layer (can be achieved by first estimating the illumination and then doing element-wise division) as the enhanced image. By adjusting the illumination map and enhancing the reflectance image, Retinex algorithms improve image quality, especially in challenging lighting conditions. Mathematically, we can represent the observed image as the product of illumination map and reflectance image

$$\mathbf{I(x,y) = L(x,y) \odot R(x,y)}$$

Where, I(x, y) is the observed pixel intensity at position (x, y).

L(x, y) represents the illumination map.

R(x, y) represents the reflectance image.

Θ represents the element wise multiplication

Corruption: It refers to the undesirable effects or distortions present in a low-light image. Low-light images often suffer from poor visibility, reduced contrast, and loss of details due to insufficient illumination. These images may exhibit noise, shadows, and uneven lighting, making them challenging to interpret or enhance. In the traditional cognitive methods, they rely on Retinex theory where they treat the reflectance component as the solution and assume that the low light images are corruption-free, leading colour distortion and amplified noise when lighting up under exposed photos.

To overcome this, a transformer kind of ideology called**, Retinexformer** was introduced where firstly, we formulate a simple yet principled **One-stage Retinex-based Framework (ORF)**. This ORF estimates the illumination information and uses it to light up the low-light images. Then ORF employs a **corruption restorer** to suppress noise, artifacts, under/overexposure, and colour distortion.

3) Different deep learning models currently used Retinex based techniques, albeit each model deploys it differently. List out the different ways in which the illumination map and reflectance image is processed in these different models.

In traditional model-based methods they applied **Retinex theory** to **decompose** the image into reflectance and the illumination layers. Such a Retinex decomposition can be achieved by traditional image processing techniques, such as filtering. However, such methods did not take proper image priors into considerations, and thus the

performance was not satisfactory. A more preferable way in using Retinex theory was formulated by treating the Retinex decomposition as an **optimization problem**. However, since the solution to Retinex decomposition is mathematically non-unique, it is crucial to design proper priors to make the optimization well-defined.

Inspired by the traditional methods, Deep learning-based models adopted the decomposition in a progressive way by treating the reflectance image as the enhanced image with the help of CNN. Throughout the period, it introduced additional feature to extract the illumination map, constructed a pipeline which first decomposes the image and then adjusts both illumination and reflectance layers. Later, they formulated a principled **One-stage Retinex-based Framework (ORF)**. ORF first estimates the illumination map information to light up the low-light image and then restores the corruption. And an **Illumination-Guided Transformer (IGT)** that utilizes illumination representations to non-local interactions of regions with different lighting conditions. By plugging IGT into ORF, we obtain our algorithm, **Retinexformer.**

Retinexformer: One-stage Retinex-based Transformer for Low-light Image Enhancement, Yuanhao Cai 1 , Hao Bian 1 , Jing Lin 1 , Haoqian Wang 1,* , Radu Timofte 2 , Yulun Zhang 3,

Low-light Image Enhancement by Retinex Based Algorithm Unrolling and Adjustment. Xinyi Liu, Qi Xie, Qian Zhao, Hong Wang and Deyu Meng

Progressive Retinex: Mutually Reinforced Illumination-Noise Perception Network for Low Light Image Enhancement. Yang Wang1 ,Yang Cao1 ,Zheng-Jun Zha1∗ Jing Zhang3 ,Zhiwei Xiong1 ,Wei Zhang2 ,Feng Wu1

1.2 U-NET

a) What is the purpose of scaling the image back up in U-Net?

We need to scale up the image generated from the encoder back to the same dimensions as the original input image. In the U-Net architecture, the downsampling path (encoder) reduces the spatial dimensions of the input image through successive convolutions and max pooling operations.

To generate the final segmentation map, the U-Net model employs an **upsampling path (decoder)**, which gradually increases the spatial dimensions of the feature maps. This upsampling process involves **transpose convolutions** or upsampling followed by convolutional layers. By scaling the feature maps back up to the original input image size, the model ensures that the segmentation map aligns with the input image.

If we don't scale the feature maps, then this might lead to mismatched dimensional error, loss in special information, lack of spatial precision by segmentation map and give errors.

b) How do skipped connections work in U-Net?
The skipped connections allow the information to flow directly across different layers. In the U-NET architecture, skip connections connect layers in the encoder part (which captures high-level features) to corresponding layers in the decoder part (which reconstructs the output). It is able to connect exactly to the right dimensions because the architecture is designed in a way that there are 4 layers in both encoder and decoder and the skipped connections are connected with their respective right spatial dimensions. By doing this, it helps us to use the contextual feature information collected in the encoder blocks to generate our segmentation map and to use our high-resolution features learned from them (through skip connections) and project it to our feature map.

c) What problems do skipped connections solve in U-Net? Is this the same as the problem solved by skipped connections in ResNet? Why or why not?
The problems that skipped connections addresses are:

- **Feature Reusability:** By connecting layers in the encoder (which captures high-level features) to corresponding layers in the decoder (which reconstructs the output), U-Nets reuse fine-grained details learned in the encoder during image reconstruction.
- **Spatial Information Recovery**: Long skip connections pass features from the encoder path to the decoder path, helping recover spatial information lost during downsampling.

These connections enhance the model's ability to capture both local and global context, giving us better predictions in tasks like semantic segmentation, image-to-image translation, and medical image segmentation.

When it comes to ResNets, it uses skip connections to allow information to flow directly across different layers. These connections enable the network to learn effectively even in deep architectures.

Unlike U-Nets, where skip connections connect same-sized parts in the downsampling and upsampling paths, ResNets use skip connections within **residual blocks (ResBlocks)** to address both degradation and the **vanishing gradient problem.**

Although, their purpose differs.

- U-Nets focus on feature reusability and spatial information recovery for image reconstruction.

- ResNets primarily address the **vanishing gradient problem**, allowing very deep networks to learn effectively.

d) Why are U-Nets used in low light image enhancement?
There are various issues in low light image enhancements, which includes low brightness, low contrast (the difference between the light and dark regions is minimal), colour distortion (unnatural colours) and significant noises. Enhancing such images while preserving important details is challenging.
The U-NET has one of the best architecture which is used for semantic segmentation and image-to-image translation tasks. With their skip connections, U-net is able to combine the location information from the input image that we applied in downsampling path with the contextual information in the upsampling path to finally obtain a general information combining localization and context, which is important to generate a good segmentation map.
Also, recent advancements in UNet based architecture include **attention mechanism** which allow the network to focus on relevant regions, emphasizing low light features while suppressing the noise. The proposed **attention gate (AG)** is incorporated into the standard U-net network to enhance model accuracy and sensitivity to foreground pixels while massive computation is not required overhead.

So, with the help of skip connections and advanced attention mechanism, U-Net architecture is widely used for critical low light image enhancement tasks which includes night time video, surveillance, autonomous vehicles, deep sea explorations, etc.

Extreme Low-Light Image Enhancement for Surveillance Cameras Using Attention U-Net by Jangwoo Kwon

1.3 ATTENTION IS ALL YOU NEED

a) Give an intuitive explanation of attention modules in computer vision.
1. Channel Attention: This module selectively focuses on different (particular) channels output (i.e feature maps) of convolutional features. By adjusting channel weights, it emphasizes relevant features while suppressing less informative ones.
2. Spatial Attention: This module focuses on specific spatial regions within an image. Spatial attention can be applied at different scales, allowing the network to zoom in on critical regions.

3. Temporal Attention: It helps models attend to relevant frames or time steps in videos. For example, in action recognition, temporal attention focuses on critical moments during an activity.

Attention mechanism was first found for NLP and later introduced in computer vision but it follows the same mechanism. When processing an image, these modules allow the network to selectively focus on specific regions. Here's how it happens:

Feature Extraction: the model first scans the image and extracts the information that might be important for the tasks. These features represent different aspects of the scene, such as edges, textures, or objects.

Attention Mechanism: Instead of treating all features equally, attention mechanisms dynamically assign weights to that specific information. Just like in the NLP, where the sentences are sliced into **tokens** and each tokens have a separate **vector** with distinct values (query, key and value vectors), in CV the pixels are made into tokens and have different vectors.

Weighted combinations: Those vectors are used to compute attention score which are calculated by measuring the compatibility between query and key vectors. Features that are more relevant to the task receive higher attention scores. And thus, they are assigned with higher weights. This weighted combination emphasizes important regions and suppresses less relevant ones.

b) List out any challenges we might face in implementing attention mechanisms?
While calculating the attention scores and using SoftMax functions, the volume of weights matrix for both query, keys and the context size can be huge for large language models (where attention overhead increases) and scaling it up can be non-trivial and time consuming. It can be computationally intensive especially significant inputs.
Sometimes attention weights can be sensitive to input variations or other noises which may not detect the attention object we tend to focus on.

c) Where do you think this technique would be of use in low-light video enhancement? Justify. (Brownie points if you can brief about how attention models are implemented in other domains of AI)
Attention U-net network, which is the integration of a self-attention gate and the standard U-net, and that can be applied to extreme low-light image enhancement. As mentioned about the Attention Gates in UNet Network, they can steadily decrease features responses in irrelevant background regions. also, it does not require training of multiple models and a large number of extra model parameters. These are some modules which are used:

1. Low light videos often suffer from noise, low contrast and loss of details. Spatial attention mechanism can adaptively adjust the luminance information in different regions of the frame. The application is like enhancing specific regions (such as face or moving objects which are at long distance), while maintaining the overall image quality.
2. Chanell attention modules (mentioned in (a) part) can also alter the filters inside the channel in convolution layers depending upon the situation. In case of low light videos, these modules can selectively enhance or suppress colour channels to the correct chromatic aberrations.
3. The low light visual lack visual details, but other modalities like audio or depth may provide additional informations. Thus, the multi-modal attentions combine the visual and non-visual cues (for example audio-visual synchronization) at appropriate frame rate to get more information on the quality.

One such field where attention mechanism is used is **reinforcement learning**. In RL, the attention mechanisms allow the agent to focus on relevant parts of the environment or input sequence. The **Policy gradient methods** play a crucial role in enhancing attention mechanisms across various domains especially when it comes to RL. They directly optimize policies by updating policy parameters to maximize expected returns.

1.4 Coding Questionnaire

CNN architecture: **CNN_Architecture**

Section D: APPROACH

1) What are your thoughts about approaching the problem? Try to come up with a step-by-step solution pipeline for the problem?

In my learning journey of this application, I first made myself clear with the all the concepts related to the project before diving into writing code/app-ing. Understanding the big picture helped me connect the answers and gain deeper insights. As I progressed, I started to connect the dots and I grinded more into each questions separately. I realized that solving AI problems is a step-by-step process, and often, AI itself provides solutions. There are tons of information lying around the internet and all that matters is what/how we grasp them, connect them and make it look simpler. #JustLikeLinearRegression :)

2) According to you, what machine learning domains does this problem explore? Just listing them out is enough.
    - Computer Vision
    - Deep learning
    - NLP
    - Robotics

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

3) What are some of the problems the project might face? Come up with innovative ways to solve them.
   - In this project, we might encounter multiple computations and complex mathematical operations. So, we need strong internet and powerful GPU which accelerates the execution of complex algorithms and giving results faster.
   - Enhancing low-light images often involves working with large data sets, which can strain the computational capabilities of CPUs alone. GPUs can efficiently handle the processing of large volumes of image data, enabling faster training, testing, and models.
   - High-quality low-light images or datasets are fundamental for training and testing image enhancement algorithms. The datasets should not be totally black out or noisy so that enhancement/training dataset becomes uncertain.
4) What do you think are some of the applications of this project? Propose ways to take the project forward.

We can introduce new sub-projects which ideates on different enhancement activities like colour enhancement, converting high light image to low light, and also publish a research paper on the mechanism and we worked on it. We can also seek guidance and mentorship from faculties, researchers, or industry experts with expertise in image processing and computer vision. We can collaborate with them to gain insights, receive feedback, and explore potential research directions.