AI02 - NIGHT VISION

LOW LIGHT MEDIA ENHANCEMENT

PROJECT MEMBER APPLICATION

2024-2025

Devisree A
EE23B023

Section A. Managerial Questionnaire

Essentials :

- AI is on the rise and it has already become essential to have AI skills. I consider myself a decent fit for the role as I am interested in learning more about AI/ML.
- I was a Deputy Coordinator in the AI club. I got a lot of exposure to various domains like neural networks, regression models, data preprocessing techniques, CV, reinforcement learning (Q-learning to be more specific) and CNNs. I know how to use standard libraries like numpy, matplotlib and pandas and train data based on various models.
- I participated in Convolve 2.0 hackathon and achieved a score of 0.90.
- Being specific to this project, I love photography and editing them. This project aligned with my interests and I'd love to understand and learn the working behind filtering and enhancement techniques involved.

Commitments/PoRs :

- I am willing to commit 6-8 hours per week on average weeks with slight flexibility during quiz weeks and a week or two before endsems.
- The other PORs I'm planning to apply for are :
    - A Project Member in one of the other CFI clubs (not very likely)
    - Saathi Mentor
    - AI club coordinator
- For a project member as a part of another CFI club, the peak will be a week or two before Research Conclave and Openhouse, similar to a project member as a part of AI club.
- I'll make sure to dedicate enough time for the projects throughout the tenure so that work doesn't pile up in the end. Still even if it does, I think I'll be able to manage time for these as it won't be close to the endsems.
- Being an AI coordinator will surely complement my learning process because of the discussion oriented and teaching nature of work.

Section B: Common Technical Questionnaire

Quantile Regression

Here is the link to the notebook for implementation of quantile loss in the given neural network.
Quantile_Loss_CFI_AI_PM_24-25_Devisree_A_EE23B023.ipynb

Section C: Project Specific Questionnaire

1.1 Retinex

a)

Let’s take a look at why retinex is widely used in deep learning models.

- Retinex algorithms split an image into illumination map and reflectance image. This separation allows the algorithm to differentiate between ambient lighting conditions and properties of the object allowing for better control over brightness.
- Manipulation of the illumination map or (and) the reflectance image is the key behind generating enhanced images with better visibility.
- They expand the dynamic range of low-light images (from a very dark image to something that is uniformly illuminated and is visually pleasing). They adjust illumination over a wide range of intensities and while preserving color contrast.
- Image details and textures are enhanced as noise is reduced.
- Methods like Histogram equalization (HE) and dehazing methods are commonly used but they do not provide results as good as retinex algorithms.

b)

- Retinex theory explains how human vision perceives the color and brightness of the objects in different light conditions.It explains the phenomenon of image constancy, where the perceived color of an object remains the same despite changes in illumination. According to Retinex theory, an image is thought to be a product of illumination map and reflectance.

$$I^c(x,y) = R^c(x,y) \times T(x,y)$$

- Illuminance refers to the light conditions that illuminate the entire scene whereas reflectance is the inherent properties of an object’s surface which is invariant to changes in illumination.
- Illumination map represents the lighting conditions across the spatial regions of an image. The low frequency components of the image are extracted to represent the overall illumination which is done by applying a low-pass filter and smoothen the image. The low frequency components are normalized to obtain the estimation of the illumination map.
- The reflectance image is calculated using the formula :

$$R^c(x,y) = \frac{I^c(x,y)}{T(x,y)}$$

- In this context, corruption refers to unwanted effects that disrupt the perception of color and brightness like noise, dark areas caused by blocking light sources, overexposed areas caused by bright light sources, reflections from glossy surfaces, etc.
- By simultaneously modifying the illumination map and reflectance image, we can get an enhanced image with improved visibility.

c)

- **Single-Scale Retinex (SSR)**
  To separate the illumination and reflectance components, SSR applies a logarithmic transformation to the input image which enhances the dynamic range of the image. SSR estimates the illumination of an input color image by applying a single Gaussian-form linear Low-Pass Filter (LPF). This LPF smooths the image to capture overall lighting conditions. The estimated illumination is subtracted from the logarithmic transformation of the input image to obtain the enhanced output image. This process enhances the dynamic range and improves visibility.

- **Multi-Scale Retinex**
  Similar to SSR, it separates the input image into illumination and reflectance components. MSR incorporates multiple levels of filtering to capture both local and global variations in illumination.

- **Multi-Scale Retinex with Color Restoration**
  It is a variation of MSR which includes a color restoration step to correct color shifts during the enhancement process. If color shifts are detected, MSCR maps the enhanced image back to original space which involves transforming the color values of each pixel in the enhanced image from modified color space to reference color space using color transformation matrix. It then performs color restoration adjusting hue, saturation and brightness.

- **Deep Retinex Decomposition (DRD)**
  It begins by decomposing the input low-light image into illuminance and reflectance components using a deep neural network. The illumination component is enhanced to improve visibility and brightness while reflectance is enhanced to preserve details and texture. It then combines both the components preserving the overall appearance of the original image.

  Reference to research papers:

  Retinex in deep learning algorithms : 2202.05972 (arxiv.org)
  SSR : Color Image Enhancement Using Single-Scale Retinex Based on an Improved Image Formation Model (eurasip.org)
  MSR : IEEE Xplore Full-Text PDF:
  MSRCR : An automated multi Scale Retinex with Color Restoration for image enhancement (researchgate.net)
  DRD : 1808.04560 (arxiv.org)

1.2 U-Net

a)

- The U-Net architecture is used for semantic segmentation. Semantic segmentation is used to classify each pixel of an image into various predefined classes. It is different from instance segmentation and object classification in the sense that it does not identify a particular object and does not provide labels to identified objects as output.
- In image classification, we give an image as input and it gives the class it belongs to as output. It doesn't make any modifications to the original image.
- In object detection, it identifies boundaries of relevant objects and bounds them in boxes.
- Semantic segmentation is used for a different purpose. It takes an image as input and masks certain regions that belong to a particular class and is given as output.
- In u-net architecture, the encoder part scales down the image (while maxpooling) while capturing important features. It has recognized the features we want it to have recognized but the feature map obtained doesn't have the same dimensions as the input image. We need to scale up this feature map to restore the dimensions of the image.
- We can't directly work with the scaled down version of the feature map. As it is scaled down, it would have failed to capture detailed features but would have just captured general features. So, scaling up increases the accuracy of the classification performed when skip connections are used.

b)

The encoder and decoder are symmetric in their structure. At each stage of the decoder path, skip connections are introduced from the corresponding stage of the encoder path. The 'connection' is essentially concatenating the feature map from the encoder and the corresponding upsampled feature map in the decoder. So while performing convolution in the upcoming stages, the layers from the upsampled feature map as well as the connected layers contribute to the final output.

c)

As I mentioned previously, an encoder captures only general features. I can think of two reasons for this :

I. During maxpooling, the intensity of the brightest pixel from a group of 4 pixels is considered. Thus, it misses out on detailed information and captures generic features only each time maxpooling is performed.

II. While performing 3x3 convolution, the convolutional kernel goes over the edge pixels lesser than the ones in the middle hence failing to extract features at the edge of an image more effectively each time convolution is performed.

To produce more accurate and detailed segmentation masks skip connections are used.

But the purpose and architecture of skip connections in ResNet are different from that of in U-Net :

- ResNet was primarily designed to address the problem of vanishing gradients in deep neural networks (CNN). This problem again arises due to several reasons :

        - Usage of activation functions such as ReLU, sigmoid and tanh cause the gradient to become very small for large values of input.
        - If the network has many layers, the gradients can get smaller and smaller as they propagate through these layers during backpropagation.

    The vanishing gradient problem can hinder the neural networks from learning effectively.

    - When gradients become small as they propagate backward, skip connections ensure that they aren't diminished significantly.
    - Structural differences :
        - Skip connections can be introduced between any output and input layer but typically they are present within residual blocks. Symmetric architecture as in u-net doesn't exist here.
        - The output and input layers are added unlike in u-net where they are generally concatenated.

d)

- Modified U-Net architecture is used for low-light image enhancement. The three improvements are :
    - Recurrent residual convolution
    - Dilated convolution
    - Reduced depth of the network
- The forward convolution block is replaced with Recurrent Residual Convolutional Units (RRCU) which consists of recursive connections where output of one unit at one stage is fed back into the unit for further processing. This improvement was proposed because RRCU facilitates learning of residual information which is the difference between input and output of each unit. This residual learning helps it to focus on learning the enhancement rather than directly predicting the enhanced image from input.
- In dilated convolution, the spacing between filter elements is increased by inserting gaps between them, resulting in a larger receptive field. In Low Light Image Enhancement (LLIE), capturing contextual information from a wider area is essential for better understanding of the scene. Dilated convolution incorporated global context while preserving spatial resolution. It also helps to focus on important image details while reducing the impact of noise present in low light images.
- Incorporating multi-ways dilated blocks (dilated convolutions with multiple dilation rates simultaneously) in the bottle neck enhances the performance.
- Dilation rates of 5 and 2 were also introduced in the encoder which decreased one maxpooling operation which leads to better accuracy because of increased dimensions of the smallest feature map.

1.3 Attention Is All You Need

a)

We humans can focus on the scenes and aspects relevant for us while ignoring all irrelevant objects in the surroundings. For example, if you are searching for your friend in a fair, you would look at the faces of different people in the crowd and look for that face which resembles that of your friend. You wouldn't pay attention to other aspects like trees around, shopkeepers, balloons, sky etc. Attention mechanisms were introduced in CV to imitate this ability of the human visual system.

There are different categories of attention mechanisms:

- Channel attention (on what do you want to pay attention to)
- Spatial attention (where do you want to pay attention to)
- Temporal attention (when you do want to pay attention in videos)

Basically, we generate a query and then assign different attention to different spatial regions of the image based on their response to that query.

b)

Challenges we might face while implementing attention based models in CV:

- **Overfitting**
  Attention mechanisms may overfit to specific datasets or images and fail to generalize well for other tasks. Factors that lead to overfitting may include:
  - Insufficient data due to which they learn very specific patterns from the available data
  - Complexity of the models lead to overfitting of noise and irrelevant details

- **Complexity**
  Computational complexity increases for images as compared to its application for word embeddings because of the spatial dimensionality of images. For self-attention mechanism, we need to compute attention weights for all pairs of spatial locations with a very large dimension. Increased channels of an image also increases the complexity.

- **Channel representation**
  In channel attention mechanisms, each channel is represented by a scalar value which signifies the relevance of the corresponding channel.It reduces the spectral information contained in an image to a numerical value, leading to information loss.

c)

Firstly let's understand the challenges faced in low light video enhancement (LLVE) over low light image enhancement (LLIE):

- LLVE operates on videos so it should take into account the temporal dependencies between consecutive frames. LLIE implements attention mechanisms without taking into consideration the temporal dependency.
- LLVE should be capable of temporal filtering and motion estimation.
- LLVE techniques must be much faster compared to LLIE due to additional processing required to analyze and enhance video data.

Attention mechanisms (especially temporal attention mechanisms) are used in LLVE. Temporal attention mechanisms selectively focus on relevant frames in a video sequence and improve video enhancement performance.

Attention mechanisms are also used in other domains of AI. It is widely used in Natural Language Processing (NLP). For example, it is used in translation of text from one language to the other and also guessing the next word in the sequence. By incorporating attention, it can selectively remember only the most relevant information and capture dependencies within data.
For example, if it has to translate a very long sentence or a paragraph, let's say, then only translation of that word is not enough but its context is also important. By using attention mechanisms, it can understand the context by capturing dependencies with the previous words and can translate better.

1.4 Coding Questionnaire

I have implemented the architecture for VGG19, AlexNet and LeNet. I have trained AlexNet on the CIFAR-10 dataset.
Here is the attached link of the collab notebook:
CNN Architecture implementation.ipynb

Section D: Approach

1)
We are working on a Low Light Media Enhancement project. Let me first give the pipeline of the project.

- **Data collection**
  We need to gather a large amount of data. It must be a diverse dataset of low-light images and videos covering all ranges of lighting conditions, objects, perspectives and scenes.

- **Preprocessing data**
  It is required to crop and resize the images. All the images in the dataset should be of the same dimensions. We can also normalize the images. It essentially means setting mean to zero and the standard deviation to a desired value. It leads to numerical stability by scaling the features.

- **Model selection**
  We have already seen that various models can be used for low-light media enhancement. Few examples are:

- Histogram equalization : It enhances the contrast of an image by redistribution of pixel intensities. The transformation is derived from the cumulative distribution function of the histogram of pixel intensity values. The pixel intensity values are redistributed to achieve a more uniform histogram.
- Gamma correction : It adjusts the brightness and contrast of an image to match characteristics of the display of the device. It applies a non-linear transformation to the pixel intensities by raising them to a value between 0 and 1.

  Other models may include retinex based deep learning networks as discussed above.

➢ **Training the model**
We need to divide the dataset into training, validation and test sets. Train the model selected using training data. Regularization of training process is important to prevent overfitting.

➢ **Validation and evaluation**
Evaluate the model on validation set and fine-tune the architecture and hyperparameters based on validation results. Commonly used evaluation metrics are PSNR(Peak SIgnal-to-Noise ratio) which is calculated based on MSE between the pixel values of reference image and reconstructed image, SSIM (Structural Similarity Index) which considers luminance, contrast and structure and produces value between -1 and 1.

➢ **Testing**
We need to evaluate the final model on the test set and measure its performance.

➢ Now, after the model is ready and is generalized for wide range of data, we can integrate it into the system architecture

2)

ML domains explored by the project are :

- Data preprocessing techniques
- Training of a model based on various evaluation metrics
- Deep learning architectures
- Retinex algorithms
- Attention Mechanisms
- Image and video processing

3)

The problems the project might face are :

- It might not be possible for us to generate a large number of images and videos so that we ensure it’s generalized well for all applications.

- Low-light images have high noise levels and reducing such high noise will be a challenge.
- Many LLIE and LLVE techniques are computationally very complex because of high processing power and large dataset of high resolution images.

Solutions to these challenges :

- We can perform data augmentation like rotate, flip, orientation etc. to generate more data from given data
- We should make use of deep learning models with retinex algorithms for best results.
- To tackle the main problem of enhancing low light videos, we need to apply temporal filtering techniques to understand temporal dependence between frames in a video sequence. We also need to implement algorithms to stabilize frames to compensate for motion blur.
- Majorly we need to test various models and evaluate them on various metrics and find the most optimal model.

4)

Some applications of the project :

- **Surveillance cameras** : They can be used in surveillance cameras to enhance footages in low-lighting environments. These can be used for military and defense applications
- **Medical applications** : It can be used for medical imaging like endoscopy where a long tube is inserted into the body where lighting conditions are poor. It helps with accurate medical diagnosis.
- **Photography and Videography** : Capturing high quality shots in low light using traditional techniques will not result in good resolution images. This project can be implemented on the captured images and videos.

To take this project forward we must at each step document the processes we follow – the dataset used, every model used and performance of each model evaluated using various evaluation metrics. We need to implement our model for various applications and make sure it generalizes well again.