VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

CSE3013 - Artificial Intelligence

Assistive Vision: Auto Caption and Speech Generation

Faculty: Prof. Rajeshkannan R

Team ID: 5

Team Members

1. Vanshika Nehra - 20BCE0599
2. Arya Jay Wadhwani - 20BCE0399
3. Arush Saxena - 20BCE2106
4. Oishi Poddar - 20BCE0187

SLOT: A1+TA1

1. Topic Definition

Visually impaired people and senior citizens are unable to identify the objects in front of them. They find it difficult to perform day to day actions because of their affected vision. With our project, we aim to provide assistive vision which will give a vivid description of the object in front of them and help them to identify it.

According to the recent census, it is said that 2.2 billion people suffer from visual disability all over the world, and require assistance for daily activities. Vision impairment poses an enormous global financial burden with the annual global costs of productivity losses associated with vision impairment from uncorrected myopia and presbyopia alone estimated to be US$ 244 billion and US$ 25.4 billion.We aim to provide a visual aid that improves daily performance, and independent living, thereby enhance the quality of life among these people.

Current technology allows applications to be efficiently distributed and run on mobile and handheld devices, even in cases where computational requirements are significant. Apps like; VoiceOver, Siri, Lookout, Be My Eyes, Blind Bargains etc have helped blind people do their daily activities.

Our main objective is to make an ML model that will analyze the picture that the user captures on their screen and voice out the components in it. We will be creating a text description of the captured image and then accurately convert the text into audio using Google Speech API.

2. Current Needs and Trends

According to WHO [11], at least 2.2 billion people have a near or distance vision impairment. In at least 1 billion – or almost half – of these cases, vision impairment could have been prevented or has yet to be addressed. The majority of people with vision impairment and blindness are over the age of 50 years; however, vision loss can affect people of all ages.

Vision impairment poses an enormous global financial burden with the annual global costs of productivity losses associated with vision impairment from uncorrected myopia and presbyopia alone estimated to be US$ 244 billion and US$ 25.4 billion.

Worldwide, more and more people are living with diabetes, particularly Type 2, which can impair vision if not spotted and treated. Nearly everyone living with diabetes will face some form of retinopathy — disease of the retina that impacts on vision quality — at some point in their lives. For those living with diabetes, regular eye checks and good diabetes control will help protect your vision.

We aim to provide a visual aid that improves daily performance, and independent living,thereby enhancing the quality of life among these people.
Scope and Applications. We will be creating a text description of captured image and then accurately convert the text into audio using Google Speech API

Main Objectives of our project include:

- To develop an application to help visually impaired people gain understanding of the environment
- To generate captions that will be converted to speech thereby helping people to listen and gain clarity
- To develop a neural network model to help generate captions along with an API to convert them back to Speech
- To consider various shortcomings based on the visual context and involve all the approaches in the project

3. Importance: A Case Study

To understand the need of this project, a study was conducted on visually imparied individuals and how they worked with Braille on clothing apparel. The purpose of this study was to survey visually impaired individuals to identify their preference of the availability of Braille labeling on clothing products as they are shopping/ selecting apparel.

Visually impaired people highlighted some challenges they face while shopping. One respondent stated *"we have a problem in color identification and we have to rely on other people to pick an item of our indicated color."* Another respondent said that they faced a challenge with size; that is, while they can feel if the size of the item is appropriate, they are unable to see how that particular size exactly fits them. Also, they cited a challenge related to price; since they cannot see the price tags they have to depend on stores' salespeople to indicate the price to them. As such, they are never certain if the price is actual or exaggerated as they are left to the "mercy" of the salespeople. This challenge mirrors that of care instructions where they have to rely on the salespeople to read the care instructions of an item for the visually-impaired people to confirm whether they are able to maintain the item. The participants also mentioned that they face a challenge with matching of the items; considering that they cannot see the design, color and size of specific items, they are unable to compare and match different items and visualize how they look in them. In light of these challenges, they provided suggestions to help address them. They suggested that stores have reader machines for labels in order to ease their shopping experience. Preferably, they should be provided with hand-held machines which they can use in scanning labels and hearing the prices and care instructions loudly thus informing their purchasing decisions.

This Case Study further gave us motivation on taking up this project so we can make a difference in the lives of these people.

4. General Process

We will be using the Flickr Dataset which has images and captions embedded in them. Kaggle will present this dataset which will be loaded on Google Colab using Pandas as the library. First we performed data preprocessing which involved converting the captions into proper generated sentences with all quotations and unnecessary words removed using the basic split function. We further removed the stopwords and found the keywords using regex and frequency counter of words that give us the most important partially generated caption. We then use the OpenCV library in Python to read the image that is given in the dataset.

ResNET50 model is used here with the Input Layer of the shape (None,224,224,3) connected to the Zero Padding convolution layer with output shape (None,230, 230,3) passed onto the conv2D layer(None,112,112,64) with Batch Normalization as well. A relu activation function is used in this layer which is then passed onto the 2 2D pooling, one with zero padding with an output shape of (None,114,114,64) and (None,56,56,56). We then pass onto the 2nd Convolution layer with various normalization and activation functions such as relu and out with Batch Normalization. 5 Layers like these are repeated with various output shapes based on the RGB configurations. We finally put the predictions onto a Dense Layer before which GlobalAveragePooling was performed.

Using these predictions, we again convert the predictions back into the sentences and use the Google Text to Speech API as well.

5. Literature Survey

- **Building up Multi-Layered Perceptrons as Classifier System for Decision Support [1]**
  This paper focuses on some application issues in multi—layered perceptron's research. The following problem areas are discussed: (1) the classification capability of multi-layered perceptrons; (2) the self-configuration algorithm for facilitating the design of the neural nets7 structure; and, finally (3) the ap-plication of the fast BP algorithm to speed up the learning procedure. Some experimental results with respect to the application of multilayered perceptron's as classifier systems in the comprehensive evaluation of Chinese large cities are presented.

- **Network Compression via Mixed Precision Quantization Using a Multi-Layer Perceptron for the Bit-Width Allocation [2]**
  Deep Neural Networks (DNNs) are a powerful tool for solving complex tasks in many application domains. The high performance of DNNs demands significant computational resources, which might not always be available. Network quantization with mixed-precision across the layers can alleviate this high demand. However, determining layer-wise optimal bit-widths is non-trivial, as the search space is exponential. This article proposes a novel technique for allocating layer-wise bit-widths for a DNN using a multi-layer perceptron (MLP). The Kullback-Leibler(KL) divergence of the softmax outputs between the quantized and full precision network is used as the metric to quantify the quantization quality. We explore the relationship between the KL-divergence and the network size, and from our experiments observe that more aggressive quantization leads to higher divergence, and vice versa. The MLP is trained with layer-wise bit-widths as labels and their corresponding KLdivergence as the input. The MLP training set, i.e. the pairs of the layer-wise bit-widths and their corresponding KL-divergence, is collected using a Monte Carlo sampling of the exponential search space. We introduce a penalty term in the loss to ensure that the MLP learns to predict bit-widths resulting in the smallest network size. We show that the layer-wise bit-width predictions from the trained MLP result in reduced network size without degrading accuracy while achieving better or comparable results with SOTA work but with less computational overhead. Our method achieves up to 6x, 4x, 4x compression on VGG16, ResNet50, and GoogLeNet respectively, with no accuracy drop compared to the original full precision pretrained model, on the ImageNet dataset.

- **Digital modulation classification using multi-layer perceptron and time-frequency features [3]**
  Considering that real communication signals corrupted by noise are generally nonstationary, and time-frequency distributions are especially suitable for the analysis of nonstationary signals, time-frequency distributions are introduced for the modulation classification of communication signals. The extracted time-frequency features have good classification information, and they are insensitive to signal to noise ratio (SNR) variation. According to good classification by the correct rate of a neural network classifier, a multilayer perceptron (MLP) classifier with better generalization, as well as, addition of time-frequency features set for classifying six different modulation types has been proposed. Computer simulations show that the MLP classifier outperforms the decision-theoretic classifier at low SNRs, and the classification experiments for real MPSK signals verify engineering significance of the MLP classifier.

- **A Reliable Localization Algorithm Based on Grid Coding and Multi-Layer Perceptron [4]**
  The traditional RSS-based fingerprint localization algorithm needs RSS values from all access points (AP) at each reference point (RP). In the large-scale indoor environment, the increasing number of APs will lead to establishing a large-scale fingerprint database, which occupies a lot of storage space. In this paper, we propose a new reliable localization algorithm, which firstly utilizes quantized RSS to encode the monitoring region which has been divided into grids, so as to specify the grids that the interested target

appears roughly. Then, we utilize Multi-Layer Perceptron (MLP) to train the grid regions in which the beacon's deployment is non-isomorphic and obtain the accurate localization result. Due to the same deployment of isomorphic regions, it is imperative to train only one model to replace the others, which greatly reduces the computation of neural networks. It can be concluded from the experimental results that compared with the traditional MLP-based fingerprint localization algorithm, the proposed algorithm reduces the size of fingerprint database over 80% with guarantee of localization accuracy. Moreover, our algorithm can obtain better localization accuracy compared with the other latest quantization based localization algorithm.

- **Convolution in Convolution for Network in Network [5]**
  Network in network (NiN) is an effective instance and an important extension of deep convolutional neural networks consisting of alternating convolutional layers and pooling layers. Instead of using a linear filter for convolution, NiN utilizes shallow multilayer perceptron (MLP), a nonlinear function, to replace the linear filter. Because of the powerfulness of MLP and 1 × 1 convolutions in spatial domain, NiN has stronger ability of feature representation and hence results in better recognition performance. However, MLP itself consists of fully connected layers that give rise to a large number of parameters. In this paper, we propose to replace dense shallow MLP with sparse shallow MLP. One or more layers of the sparse shallow MLP are sparsely connected in the channel dimension or channel–spatial domain. The proposed method is implemented by applying unshared convolution across the channel dimension and applying shared convolution across the spatial dimension in some com-putational layers. The proposed method is called convolution in convolution (CiC). The experimental results on the CIFAR10 dataset, augmented CIFAR10 dataset, and CIFAR100 data set demonstrate the effectiveness of the proposed CiC method.

- **Efficient Convolution Neural Networks for Object Tracking Using Separable Convolution and Filter Pruning [6]**
  Object tracking based on deep learning is a hot topic in computer vision with many applications. Due to high computation and memory costs, it is difficult to deploy convolutional neural networks (CNNs) for object tracking on embedded systems with limited hardware resources. This paper uses the Siamese network to construct the backbone of our tracker. The convolution layers used to extract features often have the highest costs, so more improvements should be focused on them to make the tracking more efficient. In this paper, the standard convolution is optimized by the separable convolution, which mainly includes a depth wise convolution and a pointwise convolution. To further reduce the calculation, filters in the depth wise convolution layer are pruned with filter variance. As there are different weight distributions in convolution layers, the filter pruning is guided by a hyper-parameter design. With the improvements, the number of parameters is decreased to 13% of the original network and the computation is reduced to 23%. On the NVIDIA Jetson TX2, the tracking speed increased to 3.65 times on the CPU and 2.08 times on the GPU, without significant degradation of tracking performance in VOT benchmark.

- **A Deep Neural Framework for Image Caption Generation Using GRU-Based Attention Mechanism [7]**
  Image captioning is a fast-growing research field of computer vision and natural language processing that involves creating text explanations for images. This study aims to develop a system that uses a pre-trained convolutional neural network (CNN) to extract features from an image, integrates the features with an attention mechanism, and creates captions using a recurrent neural network (RNN). To encode an image into a feature vector as graphical attributes, we employed multiple pre-trained convolutional neural networks. Following that, a language model known as GRU is chosen as the decoder to construct the descriptive sentence. In order to increase performance, we merge the Bahdanau attention model with GRU to allow learning to be focused on a specific portion of the image. On the MSCOCO dataset, the experimental results achieve competitive performance against state-of-the-art approaches.

- **Contextual LSTM for Large Scale NLP Tasks [8]**
  Documents exhibit sequential structure at multiple levels of abstraction (e.g., sentences, paragraphs, sections). These abstractions constitute a natural hierarchy for representing the context in which to infer the meaning of words and larger fragments of text. In this paper, we present CLSTM (Contextual LSTM), an extension of the recurrent neural network LSTM (Long-Short Term Memory) model, where we incorporate contextual features (e.g., topics) into the model. We evaluate CLSTM on three specific NLP tasks: word prediction, next sentence selection, and sentence topic prediction. Results from experiments run on two corpora, English documents in Wikipedia and a subset of articles from a recent snapshot of English Google News, indicate that using both words and topics as features improves performance of the CLSTM models over baseline LSTM models for these tasks. For example, on the next sentence selection task, we get relative accuracy improvements of 21% for the Wikipedia dataset and 18% for the Google News dataset. This clearly demonstrates the significant benefit of using context appropriately in natural language (NL) tasks. This has implications for a wide variety of NL applications like question answering, sentence completion, paraphrase generation, and next utterance prediction in dialog systems.

- **A CRNN-GRU based reinforcement learning approach to audio captioning [9]**
  Audio captioning aims at generating a natural sentence to describe the content in an audio clip. This paper proposes the use of a powerful CRNN encoder combined with a GRU decoder to tackle this multimodal task. In addition to standard cross-entropy, reinforcement learning is also investigated for generating richer and more accurate captions. Our approach significantly improves against the baseline model on all shown metrics achieving a relative improvement of at least 34%. Results indicate that our proposed CRNNGRU model with reinforcement learning achieves a SPIDEr of 0.190 on the Clotho evaluation set1 . With data augmentation, the performance is further boosted to 0.223. In the DCASE challenge Task 6 we ranked fourth based on SPIDEr, second on 5 metrics including BLEU, ROUGE-L and METEOR, without ensemble or data augmentation while maintaining a small model size (only 5 Million parameters).

- **Adversarial Robust Transfer Learning [10]**
  Observing that robust networks contain robust feature extractors. By training classifiers on top of these feature extractors, we produce new models that inherit the robustness of their parent networks. We then consider the case of "fine tuning" a network by re-training end-to-end in the target domain. When using lifelong learning strategies, this process preserves the robustness of the source network while achieving high accuracy. By using such strategies, it is possible to produce accurate and robust models with little data, and without the cost of adversarial training.

6. Literature Review Comparison Table

Title	**Method/Algorithm**	**Challenges**	**Observations**
Building up multi-layered perceptrons as classifier system for decision support [1]	Self-configuration algorithm and Fast Back-propagation algorithm	(1) the classification capability of multi-layered perceptrons; (2) the self-configuration algorithm for facilitating the design of the neural nets7 structure; (3) the application of the fast BP algorithm to speed up the learning procedure.	Using the self-configuration algorithm and the fast BP algorithm, a network with the structure of (33-9-3-3) is obtained and the iteration number is 5938 when convergence.

Network Compression via Mixed Precision Quantization Using a Multi-Layer Perceptron for the Bit-Width Allocation [2]	The proposed method uses a Multi-Layer-Perceptron (MLP) where the input to the MLP is the KL-divergence between the softmax output of the full precision and the quantized network and the output is the bit-width configuration for the compressed network.	Training a neural network in discrete space is challenging and the convergence is slow. There are challenges of quantization too which include quantifying deviations and the many to one problem. The proposed solution consists of three main steps: sampling the search space to create a custom training set for the MLP, MLP model training, and prediction of the bit-width configuration for the compressed network	The proposed method compresses VGG16 up to 6x when compared to the full precision model (32-bit precision weights and activations) with no accuracy drop. On ResNet50 and GoogLeNet, the method achieves up to 4x compression compared to the full precision model (32-bit precision weights and activations) while maintaining the inference accuracy. It gives better performance than uniform bit-width allocation on VGG16, ResNet50 and GoogLeNet.
Digital modulation classification using multi-layer perceptron and time-frequency features [3]	The proposed approach is based on the time-frequency features of digitally modulated signals and the use of a more robust artificial neural network, commonly referred to as an MLP.		The proposed MLP classifier with time-frequency features improves the probability of correct classification in a noisy environment.
A Reliable Localization Algorithm Based on Grid Coding and Multi-Layer Perceptron [4]	a new reliable localization algorithm, which firstly utilizes quantized RSS to encode the monitoring region which has been divided into grids, so as to specify the grids that the interested target appears roughly.	The traditional RSS-based fingerprint localization algorithm needs RSS values from all access points (AP) at each reference point (RP). In the large-scale indoor environment, the increasing number of APs will lead to establishing a large-scale fingerprint database, which occupies a lot of storage space.	It can be concluded from the experimental results that compared with the traditional MLP-based fingerprint localization algorithm, the proposed algorithm reduces the size of fingerprint database over 80% with guarantee of localization accuracy. Moreover, our algorithm can obtain better localization accuracy compared with the other latest quantization based localization algorithm.
Convolution in Convolution for Network in Network [5]	The proposed method is implemented by applying unshared convolution across the channel dimension and applying shared convolution across the spatial dimension in some com-putational	NiN utilizes shallow multilayer perceptrons (MLP). MLP itself consists of fully connected layers that give rise to a large number of parameters.	The experimental results on the CIFAR10 dataset, augmented CIFAR10 dataset, and CIFAR100 data set demonstrate the effectiveness of the proposed CiC method.

	layers. The proposed method is called convolution in convolution (CiC)	There is a need for the dense shallow MLPto be replaced with sparse shallow MLP.	
Efficient Convolution Neural Networks for Object Tracking Using Separable Convolution and Filter Pruning [6]	In this paper, the standard convolution is optimized by the separable convolution, which mainly includes a depth wise convolution and a pointwise convolution. To further reduce the calculation, filters in the depth wise convolution layer are pruned with filter variance.	Object tracking based on deep learning is a hot topic in computer vision with many applications. Due to high computation and memory costs, it is difficult to deploy convolutional neural networks (CNNs) for object tracking on embedded systems with limited hardware resources	With the improvements, the number of parameters is decreased to 13% of the original network and the computation is reduced to 23%. On the NVIDIA Jetson TX2, the tracking speed increased to 3.65 times on the CPU and 2.08 times on the GPU, without significant degradation of tracking performance in VOT benchmark
A Deep Neural Framework for Image Caption Generation Using GRU-Based Attention Mechanism [7]	CNN and RNN Architecture and Algorithm with attention mechanisms followed by a GRU model merged with the Bahdanau attention model on the MSCOCO dataset.	Such a model requires a high hardware overhead along with its requirements posed a challenge. Also, the image had to be pre processed before feeding it into the Neural Network and that required a different kind of segmentation knowledge and survey filtering based caption generative dataset. Forming that was a big challenge as well.	To create the descriptive sentence, a language model called GRU was chosen as the decoder. Meanwhile, combining the Bahdanau attention model with GRU allowed learning to be focused on a specific portion of the image in order to improve performance. The entire model can be fully trained using stochastic gradient descent, which simplifies the training procedure. Experiments show that the suggested model is capable of automatically generating appropriate captions for images.
Contextual LSTM for Large Scale NLP Tasks [8]	CLSTM Algorithm is trained on two wiki documents along with an additional proposed RNN LM for adding extra context. NN-HMM Model is also proposed as an alternative where each node of a CNN corresponds to a higher level feature.	Overall there are 5000 sentence sequences in the final dataset. For each sequence prefix AiBiCi, the model has to choose the best next sentence Di from the competing set of next sentences. The average accuracy of the baseline LSTM model on this dataset is 52%, while the average accuracy of	Using contextual features in a CLSTM model can be beneficial for different NLP tasks like word prediction, next sentence selection and topic prediction. For the word prediction task CLSTM improves on state-of-the-art LSTM by 2-3% on perplexity, for the next sentence selection task CLSTM improves on LSTM by

		the CLSTM model using word + sentence-level topic features is 63% (as shown in Table 3). So the CLSTM model has an average improvement of 21% over the LSTM model on this dataset.	≈20% on accuracy on average, while for the topic prediction task CLSTM improves on state-of-the-art LSTM by ≈10% (and improves on BOWDNN by ≈7%).
A CRNN-GRU based Reinforcement Learning approach to Audio Captioning [9]	This paper proposes the use of a powerful CRNN encoder combined with a GRU decoder to tackle this multimodal task. In addition to standard cross-entropy, reinforcement learning is also investigated for generating richer and more accurate captions	The improvement in ROUGEL and METEOR is not as significant as other metrics. The improvement in ROUGEL and METEOR is not as significant as other metrics. There is only a slight difference between this submission and the submission ranking the third (0.194 / 0.196). Even though the prediction accurately describes the audio event, it is not as detailed as the human annotations. The human annotations may contain specific descriptions like “vibrating”“buzzing” while the model prediction only generates “running”. Due to the limited information in audio as well as the direct optimization towards CIDEr metric, the model chooses to output a correct, yet general description of the audio events.	A novel audio captioning approach utilizing a CRNN encoder front-end as well as a reinforcement learning framework. Audio captioning models are trained on the Clotho dataset. The results on the Clotho evaluation set suggest that the CRNN encoder is crucial to extract useful audio embeddings for captioning while reinforcement learning further improves the performance significantly in terms of all metrics
Adversarial Robust Transfer Learning [10]	By training classifiers on top of these feature extractors, we produce new models that inherit the robustness of their parent networks. We then	When the goal is to produce a model that is not only accurate but also adversarially robust, data scarcity	By using such strategies, it is possible to produce accurate and robust models with little data, and without the cost of

	consider the case of “fine tuning” a network by re-training end-to-end in the target domain.	and computational limitations become even more cumbersome.	adversarial training.

7. Conclusion

In this study, we compared the base papers on the basis of method/algorithm, challenges faced as well as the observed traits of these papers. Furthermore, as for the research section, we introduced a single joint model for automatic image captioning based on CNN and GRU with an attention network. One encoder-decoder architecture is used in the suggested model. As the encoder, various pre-trained convolutional neural networks are used to encode an image into a compact representation as graphical characteristics. Then, to create the descriptive sentence, a language model called GRU was chosen as the decoder. Meanwhile, we combined the Bahdanau attention model with GRU to allow learning to be focused on a specific portion of the image in order to improve performance. The entire model can be fully trained using stochastic gradient descent, which simplifies the training procedure.

8. References

[1] Jun, C., Fan, Z., & Shan, F. (1995). Building up multi-layered perceptrons as classifier system for decision support. *Journal of Systems Engineering and Electronics*, *6*(2), 32-39.

[2] Soufleri, E., & Roy, K. (2021). Network Compression via Mixed Precision Quantization Using a Multi-Layer Perceptron for the Bit-Width Allocation. *IEEE Access*, *9*, 135059-135068.

[3] Ye, Y., & Wenbo, M. (2007). Digital modulation classification using multi-layer perceptron and time-frequency features. *Journal of Systems Engineering and Electronics*, *18*(2), 249-254.

[4] Sun, Z., Zhang, Y., & Ren, Q. (2020). A reliable localization algorithm based on grid coding and multi-layer perceptron. *IEEE Access*, *8*, 60979-60989.

[5] Pang, Y., Sun, M., Jiang, X., & Li, X. (2017). Convolution in convolution for network in network. *IEEE transactions on neural networks and learning systems*, *29*(5), 1587-1597.

[6] Mao, Y., He, Z., Ma, Z., Tang, X., & Wang, Z. (2019). Efficient convolution neural networks for object tracking using separable convolution and filter pruning. *IEEE Access*, *7*, 106466-106474.

[7] Khan, R., Islam, M. S., Kanwal, K., Iqbal, M., Hossain, M., & Ye, Z. (2022). A Deep Neural Framework for Image Caption Generation Using GRU-Based Attention Mechanism. *arXiv preprint arXiv:2203.01594*.

[8] Ghosh, S., Vinyals, O., Strope, B., Roy, S., Dean, T., Heck, L., & Contextual, L. S. T. M. (2016). models for large scale NLP tasks. *arXiv preprint arXiv:1602.06291*.

[9] Xu, X., Dinkel, H., Wu, M., & Yu, K. (2020, November). A crnn-gru based reinforcement learning approach to audio captioning. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)* (pp. 225-229).

[10] Shafahi, A., Saadatpanah, P., Zhu, C., Ghiasi, A., Studer, C., Jacobs, D., & Goldstein, T. (2019). Adversarially robust transfer learning. *arXiv preprint arXiv:1905.08232*.

[11] World Health Organization. (2019). World report on vision.

[12] Alali, R. M. (2017). A case study of visually impaired individuals' preferences of the availability of Braille clothing labels in shopping and selection of apparel.