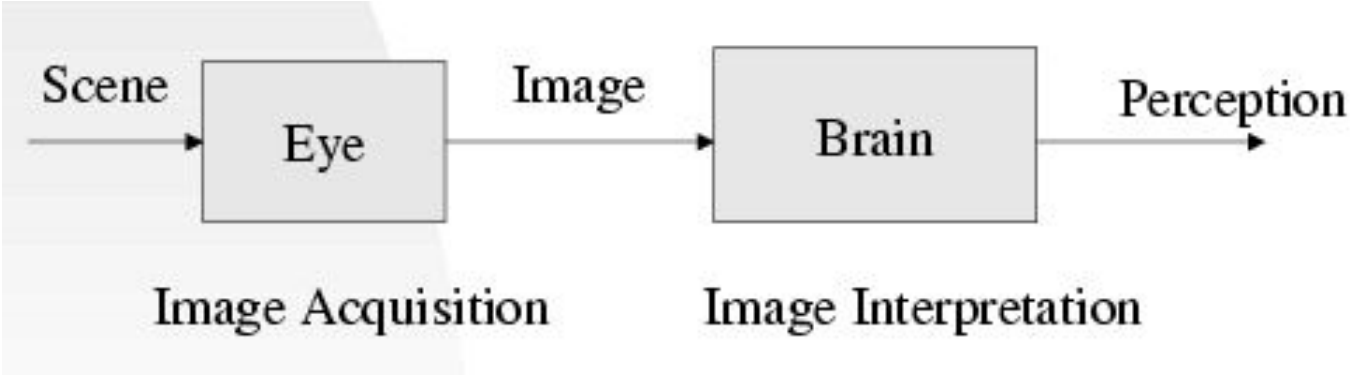
Computer Vision Applications

Presented by:

Veronica Naosekpam
PhD Scholar
Computer Science & Engineering
Indian Institute of Information Technology Guwahati, Assam, India

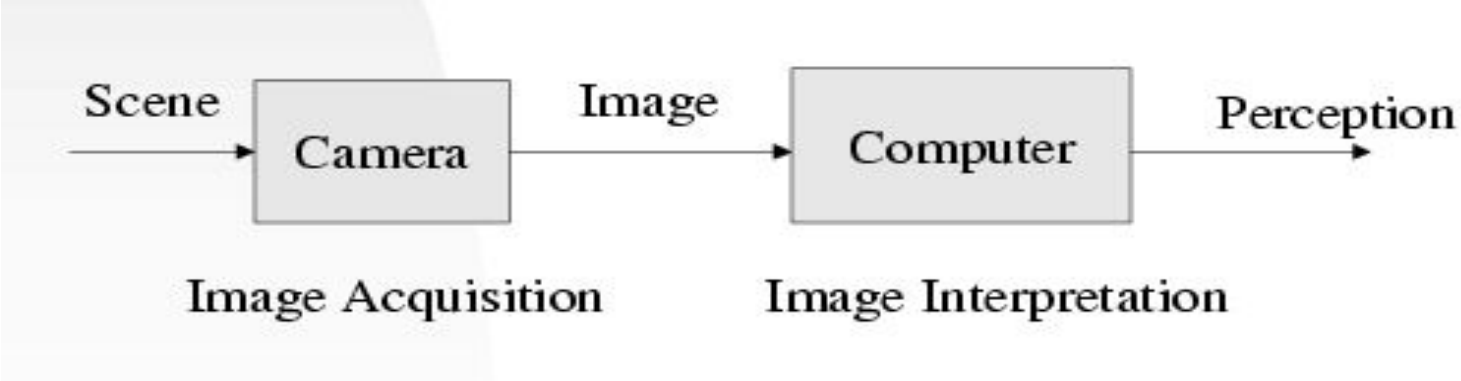
Vision

- Vision is the process of discovering what is present in the world and where it is by looking.



Computer Vision

- Computer Vision is the study of analysis of pictures and videos in order to achieve results similar to those as by people.



What is Computer Vision?

- Make computers understand images and videos.

- Given an image, interpret or understand it using computer extract properties of the 3D world.

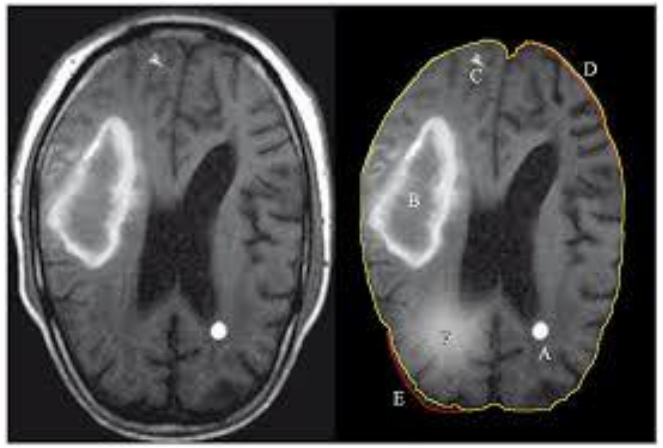
Qs with ref. to image on the right (Fig 1) :

- No. of vehicles.
- Type of vehicles.
- Location of closest obstacle.
- Assessment of congestion.
- Location Traffic scene.

Fig 1: Traffic Scene

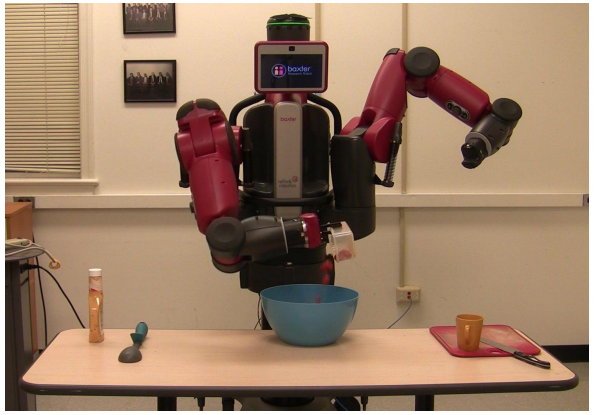
Why computer vision matters?

Safety



Health

Security



Comfort

Fun

Access

Image courtesy : Google

Computer Vision vs Machine Learning vs Artificial Intelligence

- **Artificial Intelligence** refers to the computational tools that are able to mimic and simulate human intelligence.
- AI systems will typically performs : planning, learning, reasoning, knowledge representation, perception, motion, social intelligence, creativity.
- **Machine Learning** is a science of designing and applying algorithms that are able to learn things from past cases. If some behaviour exists in past, then we may predict if or it can happen again.
- The aim of ML is to increase accuracy, but it does not care about success

Computer Vision vs Machine Learning vs Artificial Intelligence

- In Machine Learning, it usually does not care about how to obtain the data or sensors
- In Computer Vision, we care how to obtain the visual data (sensor design, active vision), how to represent the visual data, and others
- **Computer vision** relies on pattern recognition to recognize what's in a picture or video..
- Current trend in CV -> Deep Learning for Computer Vision.

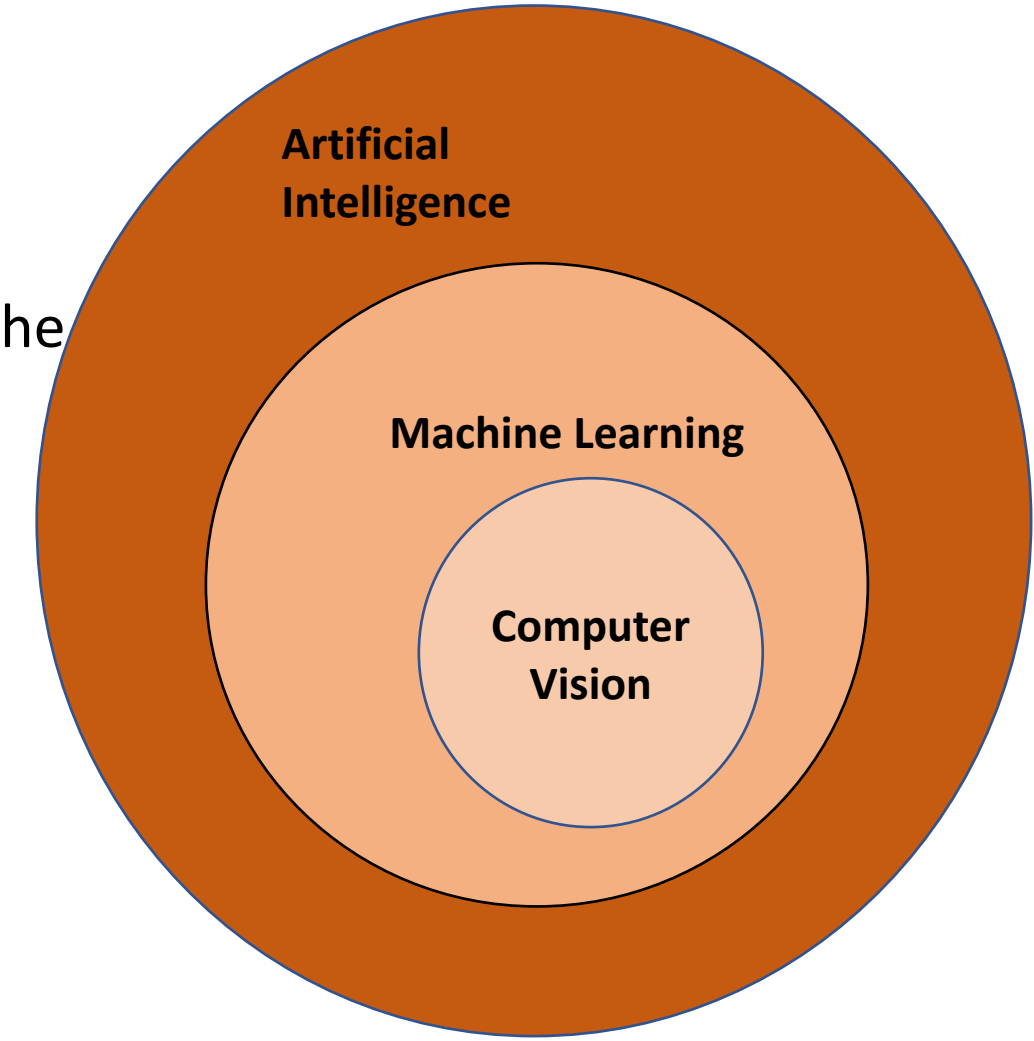
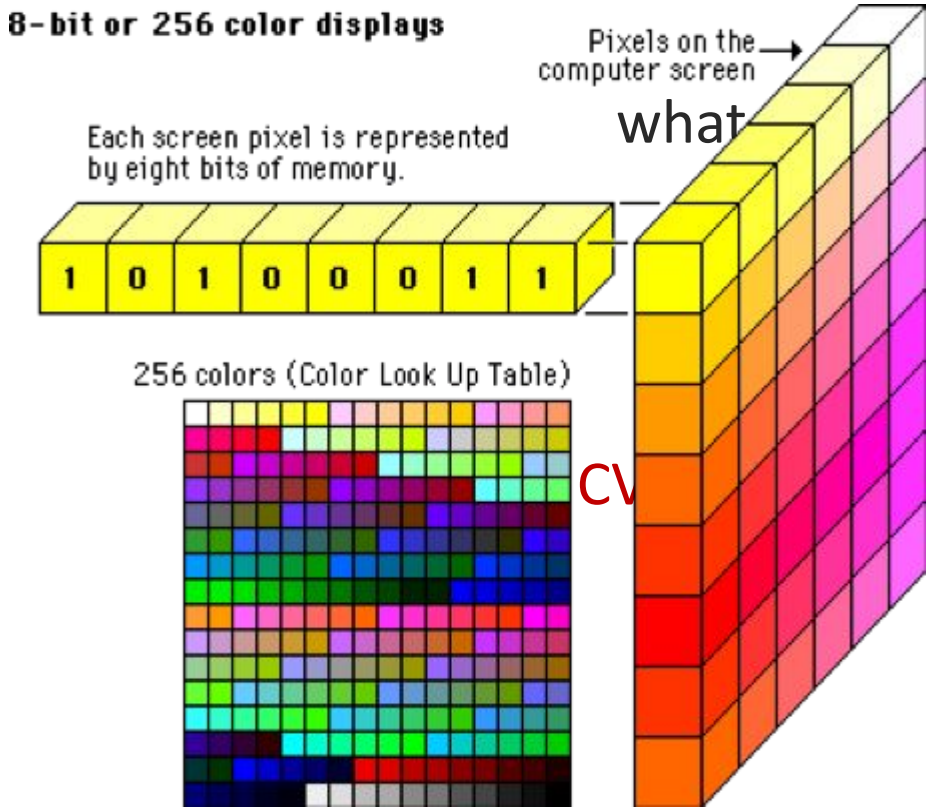


Fig 2: Relationship between AI, ML and CV

Breakdown of Computer Vision

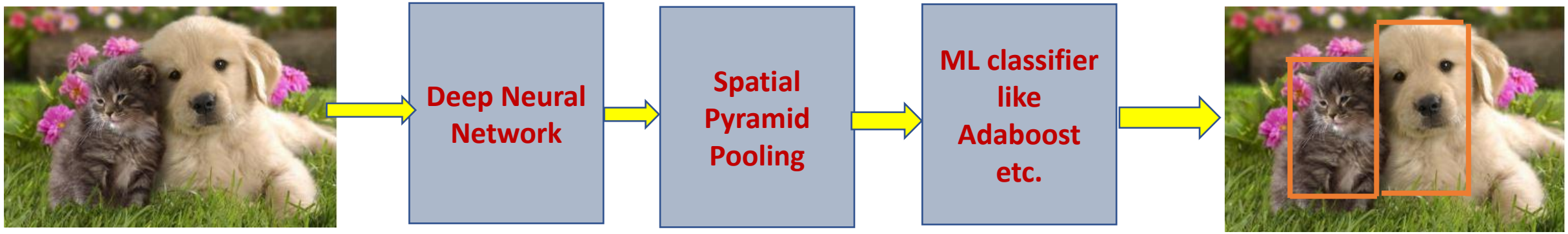
- CV is not just about converting a picture into pixels.
- It is about understanding how to extract information from those pixels and interpret they represent.
- Given an image, interpret or understand it using computer extract properties of the 3D world.
- Neural Network and Deep Learning are more capable of replicating human vision system.



(Image source : https://towardsdatascience.com/an-overview-of-computer-vision-1f75c2ab1b66)

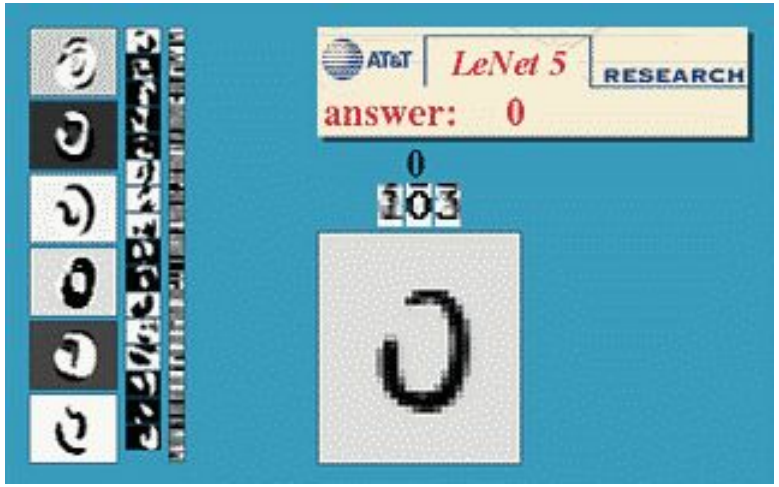
State-of-the-art Computer Vision Applications

- The trend : Deep learning + Computer Vision + Machine Learning.
  - Use deep learning for features extraction.
  - Classical computer vision for region of interest detection.
  - Use machine learning algorithm for classification (Eg: object recognition).



Optical character recognition (OCR)

- Conversion of images consisting of digits or hand-written text images to machine readable format.
- Challenges : Structured and unstructured text.
- Datasets : SVHN dataset, MNIST dataset, SVT dataset.
- Typical The OCR sub-processes are as follows:
  - Pre-processing of the image.
  - detecting and identifying a bounding box for text areas in the image and within each text area, individual text characters
  - identifying the characters.
  - Post processing.



(Source : Digit recognition, AT&T labs http://www.research.att.com/~yann/)

Deep learning based OCR models in the next slide….

- Convolutional Recurrent Neural Network (CRNN) consists of:
  - A standard CNN that creates the “feature columns”, fed into bi-LSTM which provides a sequence, identifying the relationship between the characters.
  - LSTM cell output is fed into a transcription layer, which takes the character sequence, and uses a probabilistic approach to clean the output.
- Recurrent Attention Model (RAM):
  - Based on the idea that when the human eye is presented with a new scene, certain parts of the image catch its attention. The eye focuses on those “glimpses” of information first and obtains information from them.
- Attention OCR: CNN with a decoder borrowed from the Seq2Seq machine translation model.

- Zhao, Yulei, Wenyuan Xue, and Qingyong Li. "A multi-scale CRNN model for Chinese papery medical document recognition." *2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM)*. IEEE, 2018.
- Lee, Chen-Yu, and Simon Osindero. "Recursive recurrent nets with attention modeling for ocr in the wild." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.
- Brzeski, Adam, et al. "Evaluating performance and accuracy improvements for attention-OCR." *IFIP International Conference on Computer Information Systems and Industrial Management*. Springer, Cham, 2019.

Image Super Resolutions:

- Process of upscaling or improving the details of an image..
- Objective : To improve the low resolution image to be as good (or better) than the target, known as the ground truth.
- To accomplish, a mathematical function (model) takes the low resolution image that lacks details and hallucinates the details and features onto it.
- Result: The function finds detail potentially never recorded by the original camera and upscaled image the model's prediction.
- This mathematical function is known as the and the upscaled image is the model's prediction.

Fig 3: Image Super resolutions

(Image source : https://arxiv.org/pdf/1612.07919.pdf)

- Two approaches for super-resolutions :
  - Reconstruction based : Reconstruct hidden HR(high resolution) pixels out of known linear combinations.
  - Examples based: Use prior knowledge to reconstruct a HR image.
- **Super-resolution methods and techniques :**
  - Pre-Upsampling Super Resolution
  - Post-Upsampling Super Resolution
  - Residual Networks
  - Recursive Networks
  - Progressive Reconstruction Networks
  - Attention-Based Networks
  - Generative Models -> **Very recent!**

Bottleneck : more parameters per layer!

Datasets for Super-resolution : DIV2K, Flickr2k, Waterloo dataset.

- **Super-resolutions Models :**
  - **SRCNN** : simple CNN architecture consisting of three layers: one for patch extraction, non-linear mapping, and reconstruction.
  - **Very Deep Super Resolution (VDSR)**: An improvement on SRCNN based on VGG architectures.

GAN-based architectures for Super resolutions:

- **SRGAN** : uses a GAN-based architecture to generate visually pleasing images. It uses the SRResnet network architecture as a backend, and employs a multi-task loss to refine the results.
- **EnhanceNet** : uses a Fully Convolutional Network with residual learning, which employs an extra term in the loss function to capture finer texture information.
- **ESRGAN** : improves on top of SRGAN by adding a relativistic discriminator. The advantage is that the network is trained not only to tell which image is true or fake, but also to make real images look less real compared to the generated images,

- Sajjadi, Mehdi SM, Bernhard Scholkopf, and Michael Hirsch. "Enhancenet: Single image super-resolution through automated texture synthesis." *Proceedings of the IEEE International Conference on Computer Vision*. 2017.
- Rakotonirina, Nathanaël Carraz, and Andry Rasoanaivo. "ESRGAN+: Further Improving Enhanced Super-Resolution Generative Adversarial Network." *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020.
- Yang, Wenming, et al. "Deep learning for single image super-resolution: A brief review." *IEEE Transactions on Multimedia* 21.12 (2019): 3106-3121.

Image Colorization:

- Technique to add style to a photograph or apply a combination of styles.
- It can add colour to an image that were originally taken in B&W.

Fig 4: Image colorization example

- Recent models on Image Colorization:
  - **ChromaGAN** (Adversarial Picture Colorization with Semantic Class Distribution): It is based on an adversarial strategy that captures geometric, perceptual and semantic information.

- Recent models on Image Colorization contd:
  - Instance-aware Image Colorization: The network architecture leverages an off-the-shelf object detector to obtain cropped object images and uses an instance colorization network to extract object-level features. It use a similar network to extract the full-image features and apply a fusion module to full object-level and image-level features to predict the final colours.
- Dataset: ImageNet, COCO-stuff, Places205.

- Vitoria, Patricia, Lara Raad, and Coloma Ballester. "ChromaGAN: Adversarial Picture Colorization with Semantic Class Distribution." *The IEEE Winter Conference on Applications of Computer Vision*. 2020.
- Su, Jheng-Wei, Hung-Kuo Chu, and Jia-Bin Huang. "Instance-aware Image Colorization." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020..

Object Detection

- Object Detection is a Computer Vision task which deals with identifying and locating object of certain classes in an image.

- Object localization can be done by creating a bounding box around the object.

Fig 5: An example of Object Detection

Approaches of object detection:

Two step object detection

- Identify bounding boxes which may potentially contain objects and classify each bounding separately.
- Eg : R-CNN, Fast R-CNN, Faster R-CNN

One step object detection

- It combines the detection and classification.
- Introduction of the idea of ‘regressing’ the bounding box predictions.
- Eg: YOLO and its variants, SSD, RetinaNet

- Redmon, Joseph, et al. "You only look once: Unified, real-time object detection." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
- Liu, Wei, et al. "Ssd: Single shot multibox detector." *European conference on computer vision*. Springer, Cham, 2016.
- Girshick, Ross. "Fast r-cnn." *Proceedings of the IEEE international conference on computer vision*. 2015.
- Tian, Yunong, et al. "Apple detection during different growth stages in orchards using the improved YOLO-V3 model." Computers and electronics in agriculture 157 (2019): 417-426.
- Yin, Yunhua, Huifang Li, and Wei Fu. "Faster-YOLO: An accurate and faster object detection method." *Digital Signal Processing* (2020): 102756.

Image Captioning

- Image Captioning refers to the process of generating textual description from an image – based on the objects and actions in the image
- Logically divided into two modules :
  - **image based model** (viz encoder) – which extracts the features and nuances out of our image
  - **language based model** (viz decoder)– which translates the features and objects given by our image based model to a natural sentence.

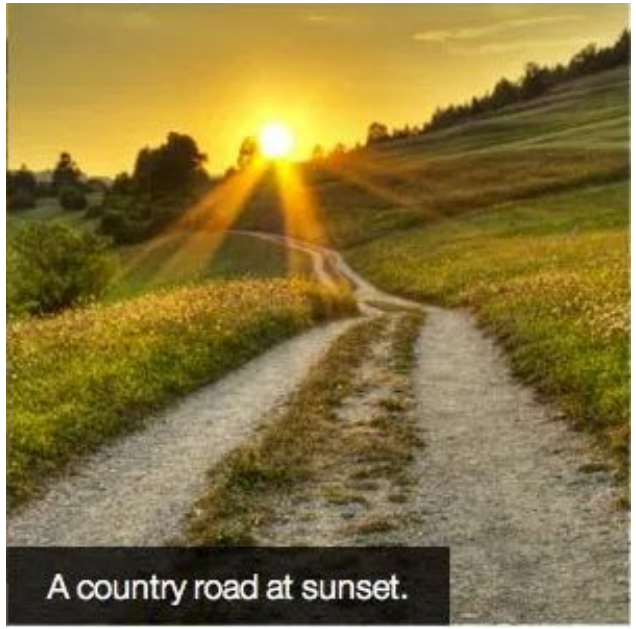


Fig 6: Image captioning

(Image Source: https://css-tricks.com/slide-in-image-captions/)

- A pretrained CNN extracts the features from the input image. The feature vector is linearly transformed to have the same dimension as the input dimension of the RNN/LSTM network. This network is trained as a language model on our feature vector.
- For training the LSTM model, we predefine label and target text are pre-defined.

- You, Quanzeng, et al. "Image captioning with semantic attention." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
- Aneja, Jyoti, Aditya Deshpande, and Alexander G. Schwing. "Convolutional image captioning." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.
- Feng, Yang, et al. "Unsupervised image captioning." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2019.

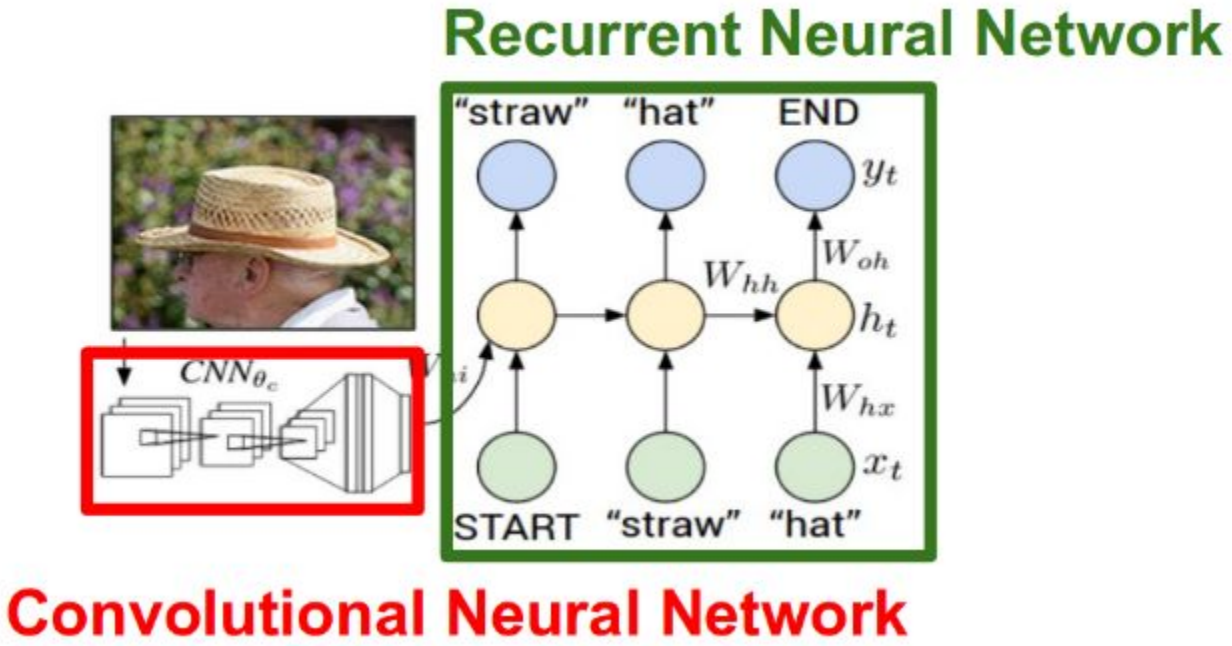


Fig 7: Image captioning steps

(Image Sourcehttps://www.analyticsvidhya.com)

Vision in space

Fig 8: NASA'S Mars Exploration Rover Spirit captured this westward view from atop a low plateau where Spirit spent the closing months of 2007.

Vision systems (JPL) used for several tasks

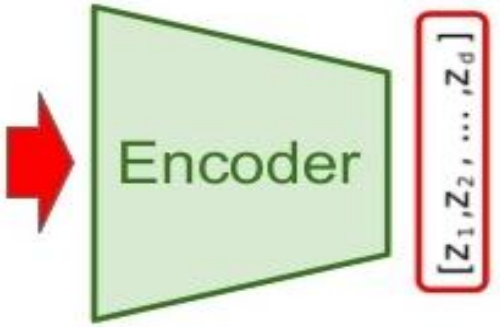
- Panorama stitching
- 3D terrain modeling
- Obstacle detection, position tracking
- For more, read “Computer Vision on Mars” by Matthies et al.

Style Transfer

- It is an optimization technique used to a **style reference** image and the **input** image you want to style — and blend them together such that the input image is transformed to look like the input image, but “painted” in the style of the style image.

Fig 9: Style transfer achieved using Deep Learning

- The Core of Deep learning Techniques for style transfer is to manipulate the encoded features map of the image on which we want to transfer the style according to the feature map of the “Style Image”.
- We use an encoder network for extracting the features from both images.



- After Getting encodings form the encoder , we try to match the encoding of style image and original image , by manipulating the encodings of original image.
- After Manipulation we get the new image with styling of input style image

Complications in the process:

- The content of the original image needs to be preserved.
- Fine tuning of the encoder to get the high quality features.
- While manipulating the features , make sure that the content of the original image does not get lost .
- The content from the style image has to be supressed.

Fig 10: Style transfer example

More Computer vision applications:

- Gesture recognition.
- Scene Image text detection.
- Vision driven surgery.
- Aid to the blind.
- Medical imaging and analysis.
- 3D image reconstruction.
- Autonomous navigations.
- …etc

Active research areas:

- Multi-lingual text detection and recognition.
- Driverless cars with focus on snowy scenes.
- Scene understanding / Scene parsing.
- Deep fake using GAN.
- Shadow removal from an image.
- Video classifications.
- …Etc.

Any Questions…?

THANK YOU