

Deep Learning
Prof. Mitesh M. Khapra
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Lecture – 01
Chapter 8: The Madness (2013-)

So, this was all happening where deep learning now started showing a lot of promise in a lot of fields N L P, vision speech and again this deep reinforcement learning and so on, which led to this complete madness starting from 2013.

(Refer Slide Time: 00:28)

He sat on a chair.

Language Modeling


- Mikolov et al. (2010)^[26]
- Kiros et al. (2015)^[27]
- Kim et al. (2015)^[28]

Module 8

Well almost for every application the traditional methods were then overwritten or kind of beaten by deep neural network based system. So, something like language modelling, which has been around since probably 1950s or so.

Now the reigning algorithm or the better algorithm for language modelling is now something which is based on deep neural networks.

(Refer Slide Time: 00:50)



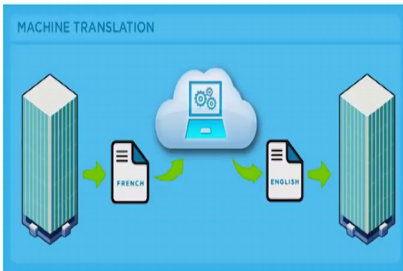
Speech Recognition

- Hinton et al. (2012)^[29]
- Graves et al. (2013)^[30]
- Chorowski et al. (2015)^[31]
- Sak et al. (2015)^[32]

Module 8

Then similarly for speech recognition, lot of work, a lot of probabilistic, lot of work based on probabilistic models was done in this or in the speech area or the speech literature for the past 30 40 years, and now all of that has been overcome by deep neural network based solutions.

(Refer Slide Time: 01:08)



Machine Translation

- Kalchbrenner et al. (2013)^[33]
- Cho et al. (2014)^[34]
- Bahdanau et al. (2015)^[35]
- Jean et al. (2015)^[36]
- Gulcehre et al. (2015)^[37]
- Sutskever et al. (2014)^[38]
- Luong et al. (2015)^[39]
- Zheng et al. (2017)^[40]
- Cheng et al. (2016)^[41]
- Chen et al. (2017)^[42]
- Firat et al. (2016)^[43]

Module 8

Same for machine translation, a lot of interest in this field, a lot of companies now have their machine translation systems based on deep neural networks as opposed to the

earlier phrase based statistical machine translations or the probabilistic models, which were used earlier.


(Refer Slide Time: 01:23)

Time	User	Utterance
03:44	Old	I dont run graphical ubuntu, I run ubuntu server.
03:45	kuja	Taru: Haha sucker.
03:45	Taru	Kuja: ?
03:45	bur[n]er	Old: you can use "ps ax" and "kill (PID#)"
03:45	kuja	Taru: Anyways, you made the changes right?
03:45	Taru	Kuja: Yes.
03:45	LiveCD	or killall speedlink
03:45	kuja	Taru: Then from the terminal type: sudo apt-get update
03:46	_pm	if i install the beta version, how can i update it when the final version comes out?
03:46	Taru	Kuja: I did.

Sender	Recipient	Utterance
Old		I dont run graphical ubuntu, I run ubuntu server.
bur[n]er	Old	you can use "ps ax" and "kill (PID#)"

Conversation Modeling

- Shang et al. (2015)^[44]
- Vinyals et al. (2015)^[45]
- Lowe et al. (2015)^[46]
- Dodge et al. (2015)^[47]
- Weston et al. (2016)^[48]
- Serban et al. (2016)^[49]
- Bordes et al. (2017)^[50]
- Serban et al. (2017)^[51]



Module 8

Similarly, for conversation modelling dialogue, a lot of new work started in dialogue post a deep learning era, where people now realize that if you have a lot of sequences of conversations, you could actually try to train a deep neural network to learn from this sequence and have conversations with humans. Of course, you are nowhere close to human level conversations, we are very very far off from them, but in limited domains these bots are showing some success now.

(Refer Slide Time: 01:50)

The slide displays four tasks for question answering:

- Task 1: Single Supporting Fact**
Mary went to the bathroom.
John moved to the hallway.
Mary travelled to the office.
Where is Mary? A: office
- Task 2: Two Supporting Facts**
John is in the playground.
John picked up the football.
Bob went to the kitchen.
Where is the football? A: playground
- Task 3: Three Supporting Facts**
John picked up the apple.
John went to the office.
John went to the kitchen.
John dropped the apple.
Where was the apple before the kitchen? A: office
- Task 4: Two Argument Relations**
The office is north of the bedroom.
The bedroom is north of the bathroom.
The kitchen is west of the garden.
What is north of the bedroom? A: office
What is the bedroom north of? A: bathroom

Question Answering

- Hermann et al. (2015)^[52]
- Chen et al. (2016)^[53]
- Xiong et al. (2016)^[54]
- Seo et al. (2016)^[55]
- Dhingra et al. (2017)^[56]
- Wang et al. (2017)^[57]
- Hu et al. (2017)^[58]

Module 8

Same for question answering where you are given a question and you want to answer it, either from a knowledge graph or from a document or from a image and so on.

(Refer Slide Time: 01:57)

The slide shows an image of a person sitting on a chair with a dog on the floor. Bounding boxes are drawn around the person, dog, and chair. Labels 'person', 'dog', and 'chair' are placed next to their respective boxes.

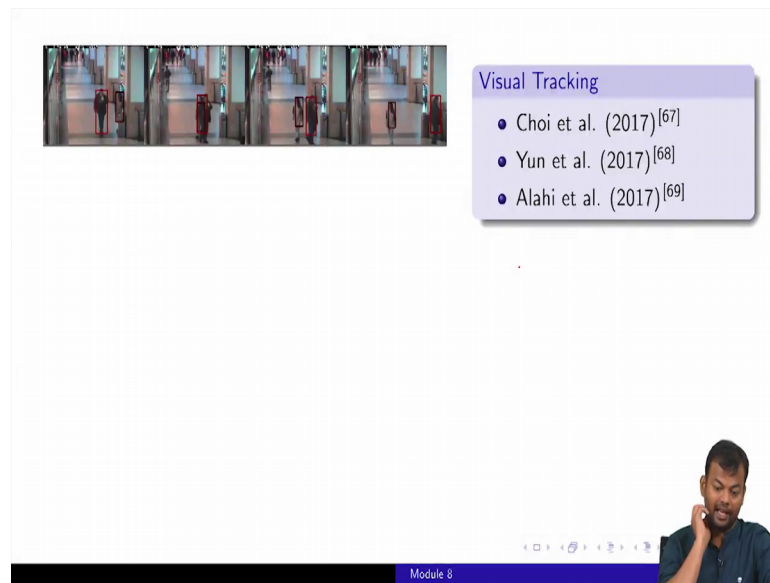
Object Detection/Recognition

- Semantic Segmentation (Long et al, 2015)^[59]
- Recurrent CNNs (Liang et al., 2015)^[60]
- Faster RCNN (Ren et al., 2015)^[61]
- Inside-Outside Net (Bell et al., 2015)^[62]
- YOLO9000 (Redmon et al., 2016)^[63]
- R-FCN (Dai et al., 2016)^[64]
- Mask R-CNN (He et al., 2017)^[65]
- Video Object segmentation (Caelles et al., 2017)^[66]

Module 8

And in the field of computer vision things like object detection, most of the raining systems or the best performing systems, nowadays are deep neural network based systems, a lot of advances are being made on these systems over in the last few years.

(Refer Slide Time: 02:17)



Visual Tracking

- Choi et al. (2017)^[67]
- Yun et al. (2017)^[68]
- Alahi et al. (2017)^[69]


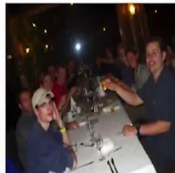
Module 8

Same for visual tracking where you want to track the same person in a video or image captioning, where you want to generate captions for images. For example, people upload a lot of images on Facebook.

(Refer Slide Time: 02:21)



Image Captioning

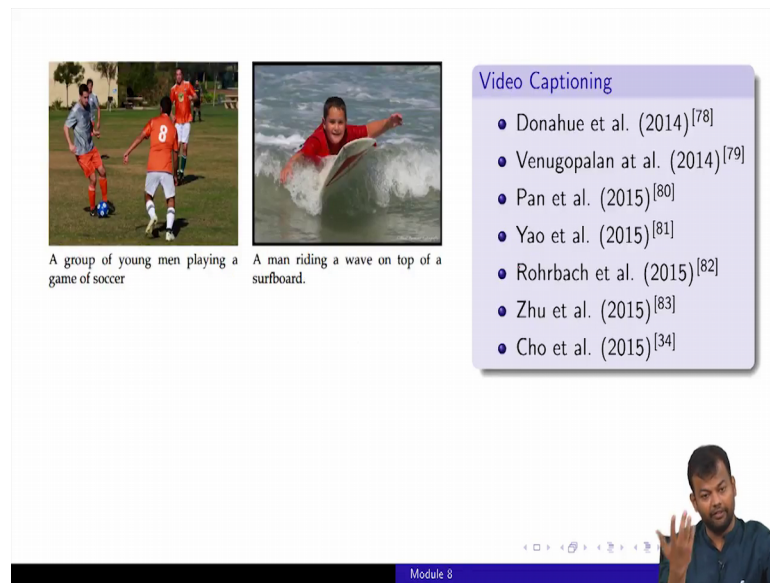
	
Retr. 1. Top view of the lights of a city at night, with a well-illuminated square in front of a church in the foreground; 2. People on the stairs in front of an illuminated cathedral with two towers at night;	1. Tourists are sitting at a long table with beer bottles on it in a rather dark restaurant and are raising their bierglaeser; 2. Tourists are sitting at a long table with a white table-cloth in a somewhat dark restaurant;
Gen. A square with burning street lamps and a street in the foreground;	Tourists are sitting at a long table with a white table cloth and are eating;

- Mao et al. (2014)^[70]
- Mao at al. (2015)^[71]
- Kiros et al. (2015)^[72]
- Donahue et al. (2015)^[73]
- Vinyals et al. (2015)^[74]
- Karpathy et al. (2015)^[75]
- Fang et al. (2015)^[76]
- Chen et al. (2015)^[77]

Module 8

And if you want to automatically caption them or imagine you are on a reselling site right, something like O L X where you upload your furniture, and you do not provide a description from that, but can the machine already automatically generate a description for it. So, it is easier for the human to read what that product is and so on right.

(Refer Slide Time: 02:45)



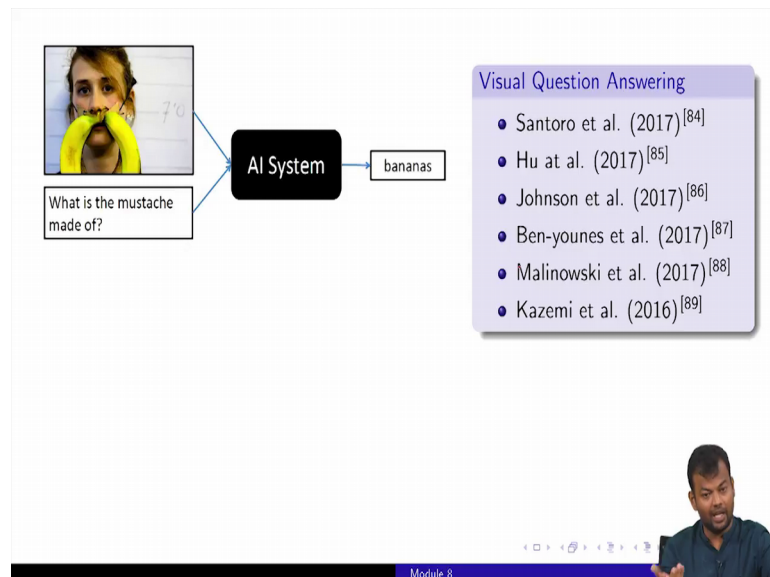
Video Captioning

- Donahue et al. (2014)^[78]
- Venugopalan et al. (2014)^[79]
- Pan et al. (2015)^[80]
- Yao et al. (2015)^[81]
- Rohrbach et al. (2015)^[82]
- Zhu et al. (2015)^[83]
- Cho et al. (2015)^[34]

Module 8

So, similarly video captioning, I given a video anyone to caption the main activity which is happening in that video; all of these problems are being solved using deep learning based solutions, using a combination of something known as feed forward neural networks or convolutional neural networks or recurrent neural networks and so on.

(Refer Slide Time: 03:03)



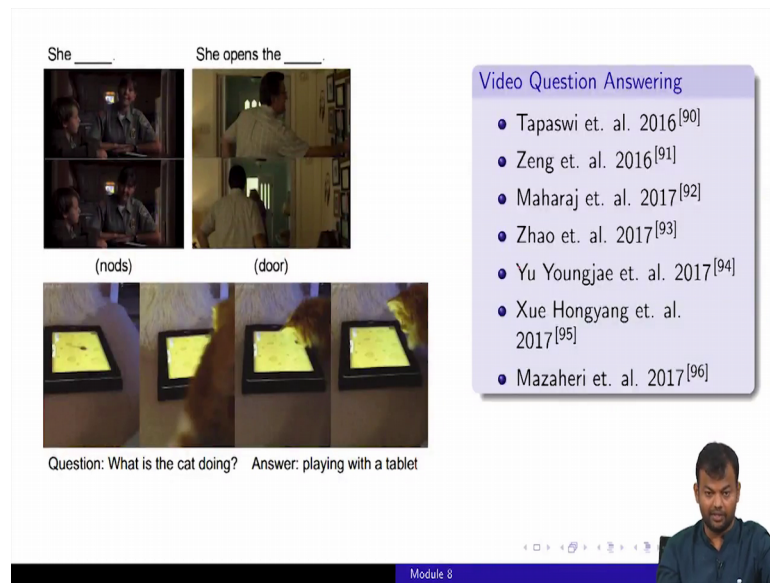
Visual Question Answering

- Santoro et al. (2017)^[84]
- Hu et al. (2017)^[85]
- Johnson et al. (2017)^[86]
- Ben-younes et al. (2017)^[87]
- Malinowski et al. (2017)^[88]
- Kazemi et al. (2016)^[89]

Module 8

Visual question answering, you are given an image and a question and you want to answer that question.

(Refer Slide Time: 03:08)



The slide illustrates video question answering. It features two rows of video frames. The first row shows a person nodding, labeled "(nods)", and a person opening a door, labeled "(door)". The second row shows a cat playing with a tablet, with the text "Question: What is the cat doing? Answer: playing with a tablet" below it. A list of research papers is provided in a purple box on the right, and a presenter's video feed is in the bottom right corner.

She ____ She opens the ____

(nods) (door)

Question: What is the cat doing? Answer: playing with a tablet


Module 8

Video Question Answering

- Tapaswi et. al. 2016^[90]
- Zeng et. al. 2016^[91]
- Maharaj et. al. 2017^[92]
- Zhao et. al. 2017^[93]
- Yu Youngjae et. al. 2017^[94]
- Xue Hongyang et. al. 2017^[95]
- Mazaheri et. al. 2017^[96]

Video question answering; answering questions from videos.

(Refer Slide Time: 03:11)



The slide illustrates video summarization. It shows a grid of "Input video" frames on the left and a "Summary" of three key frames on the right. A list of research papers is provided in a purple box on the right, and a presenter's video feed is in the bottom right corner.

Input video

Summary

Module 8

Video Summarization

- Chheng 2007^[97]
- Ajmal 2012^[98]
- Zhang Ke 2016^[99]
- Zhong Ji 2017^[100]
- Panda 2017^[101]

Video summarizations; if you are given a large video and you want to generate a trailer, a sort of a trailer for that video contains, which kind is the most important frame for that video. Even these systems are based on deep learning.

(Refer Slide Time: 03:22)



The slide displays a 3x3 grid of generated human faces, showing a variety of features and expressions. To the right of the grid is a text box titled "Generating Authentic Photos" containing a bulleted list of generative models and their associated research papers. At the bottom right of the slide, there is a small inset image of a man, likely the presenter, and a navigation bar with the text "Module 8".

Generating Authentic Photos

- Variational Autoencoders (Kingma et. al., 2013)^[102]
- Generative Adversarial Networks (Goodfellow et. al., 2014)^[103]
- Plug & Play generative nets (Nguyen et al., 2016)^[104]
- Progressive Growing of GANs (Karras et al., 2017)^[105]

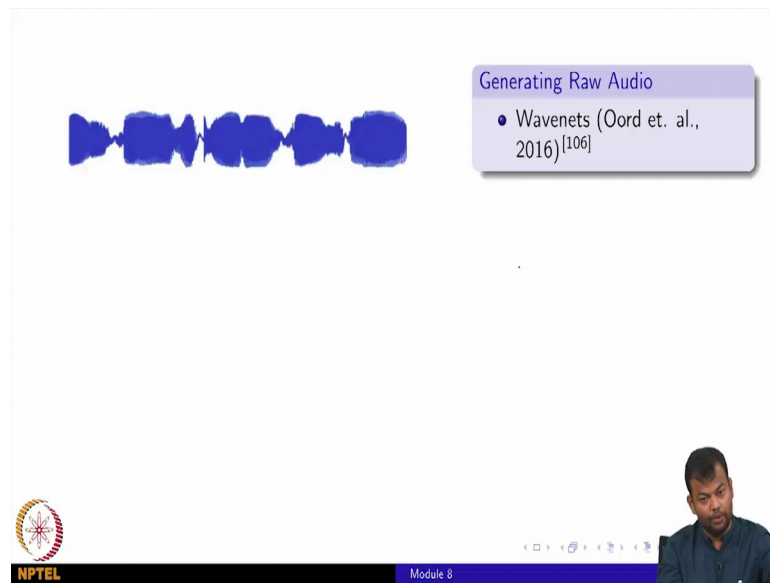
Module 8

Then this was all about classification recognition and so on, but now people started getting more ambitious that can, we humans are very good at creativity. So, can we use machines to be creative right to generate images? So, now, if I have seen a lot of celebrity faces, can I generate new celebrity faces or if I have seen a lot of bedroom images.

And I am if a fireman architect. Now can I generate new bedroom images can i, can we train a machine to generate new bed bedroom images. So, a lot of phenomenal progress or work has happened in this field in the last 4 5 years, starting with things like generative adversarial networks, we reached an auto encoders and so on.

And people are now starting to seriously invest into creativity that how to make machines creative, again we are far off from where the desired output, but there is still significant progress happening in this field generating audio.

(Refer Slide Time: 04:15)



Generating Raw Audio

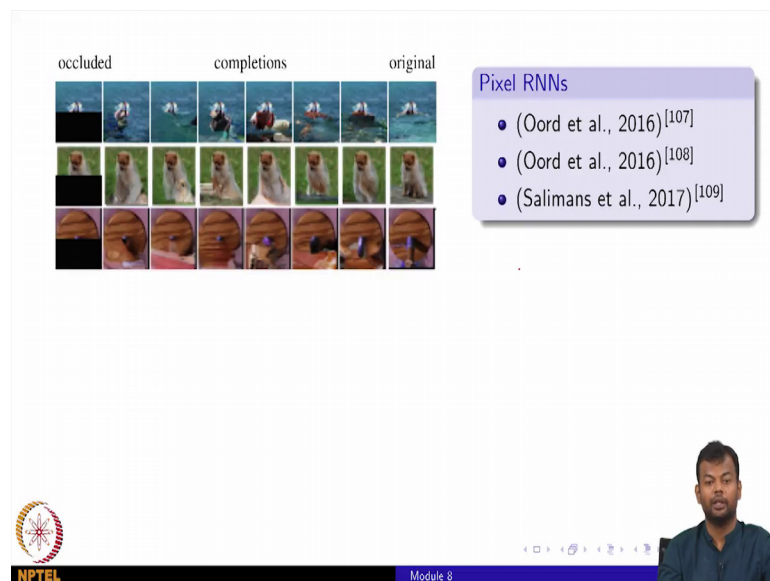
- Wavenets (Oord et. al., 2016)^[106]

NPTEL Module 8

The slide displays a blue waveform representing raw audio. A small inset video of a presenter is visible in the bottom right corner. The NPTEL logo and 'Module 8' are at the bottom.

So, that was about generating images, you can generate music also.

(Refer Slide Time: 04:21)



Pixel RNNs

- (Oord et al., 2016)^[107]
- (Oord et al., 2016)^[108]
- (Salimans et al., 2017)^[109]

NPTEL Module 8

The slide shows a 3x3 grid of image examples. The first column is labeled 'occluded', the second 'completions', and the third 'original'. The images show a boat on water, a monkey, and a person's face. A small inset video of a presenter is visible in the bottom right corner. The NPTEL logo and 'Module 8' are at the bottom.

And this is again about generating images and so on.