



Deep Learning for Computer Vision
Professor Vineeth N Balasubramanian
Department of Computer Science and Engineering
Indian Institute of Technology, Hyderabad
Lecture - 58
Going Beyond Captioning: Visual QA, Visual Dialog

(Refer Slide Time: 00:14)





Deep Learning for Computer Vision

Going beyond Captioning: Visual QA, Visual Dialog

Vineeth N Balasubramanian



Department of Computer Science and Engineering
Indian Institute of Technology, Hyderabad



Vineeth N B. (IIT-H) §9.3 Visual QA and Dialog 1 / 29


The relevance of RNN models and Attention models in Computer Vision becomes pronounced when you look at problems at the combination of vision, and language. This setting results in many sequence learning problems. We saw one in the last lecture, Image Captioning. And we will now go and go even further and talk about tasks such as Visual Question Answering, and Visual Dialogue.

(Refer Slide Time: 00:51)



Review: Question



Can we do the opposite (caption-to-image) of what we learned in the previous lecture? How?
Yes, we will see deep generative models soon that can do this



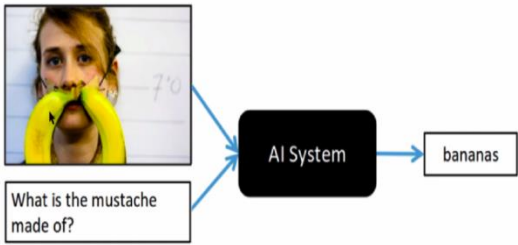
Vineeth N B (IIT-H) §9.3 Visual QA and Dialog 2 / 29

One question that we left behind was, instead of image captioning, can we do the opposite, go from caption to image, using any models that we have seen so far? We can, but not using the models that we have seen so far. We will see generative models quite soon. And see how we can use them to do caption to image generation.

(Refer Slide Time: 01:19)




Visual Question Answering (VQA): Task Overview¹



Credit: Aishwarya Agrawal, Devi Parikh, Georgia Tech

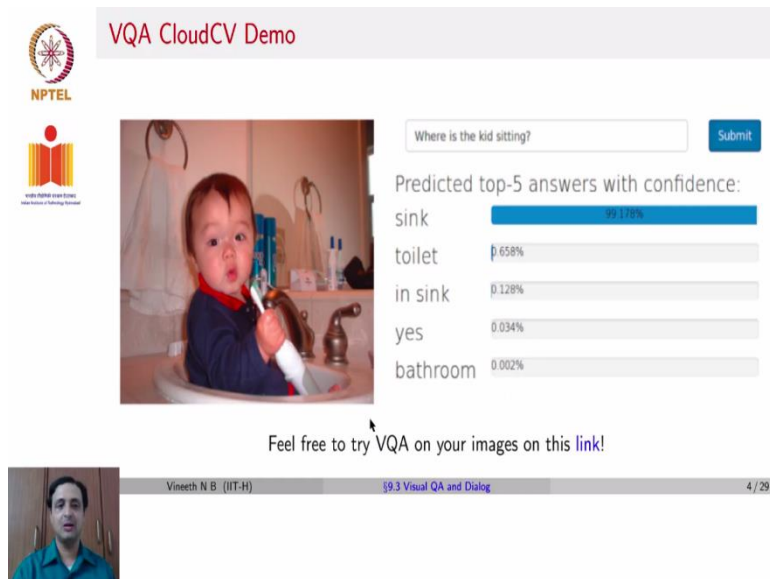
¹Agrawal et al, VQA: Visual Question Answering, IJCV 2015



Vineeth N B (IIT-H) §9.3 Visual QA and Dialog 3 / 29

In visual question answering, which you can think of as an extension of the image captioning task. The overview of the task is given an image such as what you see here, and given a question, such as, what is the mustache made off? The goal is for our system, a deep learning model to give the answer bananas. So you need an understanding of the question. You also need an understanding of the image and the relevance of that question to a specific part of an image and then be able to give a reasonable answer.

(Refer Slide Time: 02:05)



VQA CloudCV Demo

Where is the kid sitting?

Predicted top-5 answers with confidence:

sink	99.126%
toilet	0.658%
in sink	0.128%
yes	0.034%
bathroom	0.002%

Feel free to try VQA on your images on this [link!](#)

Vineeth N B (IIT-H) 99.3 Visual QA and Dialog 4 / 29

Here is a demo of how this VQA task works. Given an image and the question, where is the kid sitting? If you had a good train model for this particular task, the kind of answer you would get is the sink gets 99 percent confidence, which is true. If you look at the image. This is a demo from a website known as cloud CV, developed by Dhruv Batra and Devi Parekh a Georgia Tech. You are welcome to go to this link and try out visual question answering on your images.

(Refer Slide Time: 02:49)

VQA Dataset²

- Open-ended answers and Multiple-choice answers
- 250K images (MS COCO + 50K abstract images)
- 750K questions, 10M answers
- Each question is answered by 10 human annotators

²Agrawal et al, VQA: Visual Question Answering, IJCV 2015
Vineeth N.B. (IIT-H) §9.3 Visual QA and Dialog 5 / 29

Before we go forward and talk about what models you can use to solve VQA problems. We will be looking at combinations of CNNs RNNs and Attention. When we say RNNs, we also subsume LSTM's and GRUs in that same term. Before we discuss those models, let us discuss the kind of datasets that one needs to solve VQA problems. So far, we have needed Image Classification datasets, Image Detection data sets, Image Segmentation datasets and to some extent, what we saw in the last lecture, Image Captioning datasets.

In each of those, the expected data set to train the model is self-evident. For this task, VQA, things are a bit more complex. So let us see a few datasets that have been developed to address this problem. So this was the first dataset known as the VQA dataset developed in 2015. This dataset has images and questions such as what you see on this image here. Given an image, there are two questions what color are her eyes? What is the mustache made of?

Similarly, for the second image, how many slices of pizza are there? Is this a vegetarian pizza? The third image is this person expecting company? What is just under a tree under tree? And the last one Does it appear to be rainy? Does this person have a 2020 vision? As you can see, for each of these questions, the model needs to understand the question as well as the image and the relationship between the image, and the question.

So in this VQA data set, you can see that the answers are open-ended. And the answers are also multiple choice answers. So you are given a set of options. And the model has to choose one of those options as the output. Why is this important, which means the task for the model becomes classification? Again, given a set of options, the model has to output a softmax over or a probability vector over the set of options.

And the option with the highest probability is the predicted output of the model. So this VQA data set has about 250 thousand images, which is obtained from a dataset known as MS COCO. MS COCO stands for Microsoft Common Objects in Context. And this dataset was developed for the image captioning task. That dataset is extended for the VQA dataset plus about 50 k abstract images such as these.

There are a total of about 750,000 questions across these images. As you can see, given a single image, you can have multiple questions. And there are a total of about 10 million answers. Because for each question, you need multiple answers to choose from. And each question is answered by 10 Human annotators. This is the VQA dataset.

(Refer Slide Time: 06:45)

COCO-QA³

COCOQA 5078
How many leftover donuts is the red bicycle holding?
Ground truth: three

COCOQA 1238
What is the color of the tee-shirt?
Ground truth: blue

COCOQA 26088
Where is the gray cat sitting?
Ground truth: window

- Automatically generate QA pairs with MS COCO captions
- 118K QA pairs on 123K images
- 4 types of questions: **What object, How many, What color, Where**

³Ren et al, Exploring Models and Data for Image Question Answering, NeurIPS 2015

Vineeth N.B. (IIT-H) §9.3 Visual QA and Dialog 6 / 29

Another dataset that was developed in 2015 is known as the COCO QA dataset. This again used the Microsoft COCO dataset. As I just mentioned, the Microsoft COCO data set was an image

captioning dataset. So the COCO QA data set used those captions to automatically generate QA pairs question-answer pairs.

So for example, given this image, if there was a caption, the caption is converted now into a question and answer. The question could be how many leftover doughnuts is the red bicycle holding? And the answer is three. What is the color of this T-shirt? The answer is Blue. And where is the gray cat sitting? The answer is Window. You can see that one can obtain question-answer pairs from a caption.

The only constraint here is that the answer is a single word which is true of many VQA datasets. This data set had hundred and 118,000 question-answer pairs on 123,000 images. There were four types of questions, what object, how many, what color, and where were the kinds of questions in the dataset.

(Refer Slide Time: 08:23)

Visual7W⁴

NPTEL
National Programme on Technology Enhanced Learning

<p>Q: What endangered animal is featured on the truck?</p> <p>A: A bald eagle. A: A sparrow. A: A hummingbird. A: A raven.</p>	<p>Q: Where will the driver go if turning right?</p> <p>A: Onto 24th Rd. A: Onto 25th Rd. A: Onto 23th Rd. A: Onto Main Street.</p>	<p>Q: When was the picture taken?</p> <p>A: During a wedding. A: During a bar mitzvah. A: During a funeral. A: During a Sunday church service.</p>
<p>Q: Which pillow is farther from the window?</p>	<p>Q: Which step leads to the tub?</p>	<p>Q: Which is the small computer in the corner?</p>

- 7W stands for **what, where, when, who, why, how** and **which**
- 328K QA pairs on ~ 47K images
- Two types of tasks: **telling** and **pointing**

⁴Zhu et al, Visual7W: Grounded Question Answering in Image, CVPR 2016


Vineeth N B (IIT-H) §9.3 Visual QA and Dialog 7 / 29

In subsequent years, in 2016 was developed another popular dataset known as Visual 7W was. This dataset has images and questions such as what you see here. What end endangered animal is featured on the truck in this image? The answer is a bald eagle and the other options are Sparrow, Hummingbird, and a Raven. Similarly, where will the dry go driver go if turning right? If you look at this particular image here, you could then say onto 24 3 4th Road and then you have other options, which are similar road names that may be relevant.

So 7W in the title of the dataset stands for What, Where, When, Who, Why, How? And finally, this is a different kind of question compared to other datasets. The question here is given an image at this bottom left, the question could be which pillow is farther from the window? And the answer would be what you see in this yellow box here. Which step leads to the tub.


Once again, the answer would be the yellow box here which is the small computer in the corner. Once again, the answer is in the yellow box here. That is the reason it is called Visual 7W. This dataset has about 328,000 question-answer pairs on around 47,000 images. As we just saw, there are two kinds of tasks, the telling task, which is what we saw on the top row, and the pointing task, where we have to point to a particular part of the image to answer the question.

(Refer Slide Time: 10:35)




NPTEL

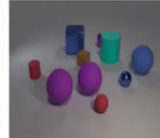
CLEVR⁵



Q: How big is the gray? Q: There is a purple rubber object that is ball that is the same behind the big shiny size as the red cylinder behind the big der; what material is metallic thing that is it?
A: metal
Q-type: query_material
A: small
Q-type: query_size
Size: 17




Q: There is a tiny Q: What is the shape rubber thing that is of the tiny green thing the same color as the that is made of the metal cylinder; what same material as the shape is it? large cylinder?
A: cylinder
Q-type: query_shape
Size: 9



Q: There is a small Q: Is the size of the ball that is made of red rubber sphere the same material as same as the purple the large block; what metal thing? color is it?
A: yes
Q-type: equal_size
Size: 12

- 100K rendered images and 1M automatically generated questions
- Questions are complex and require reasoning skills

⁵Johnson et al, CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning, CVPR 2017



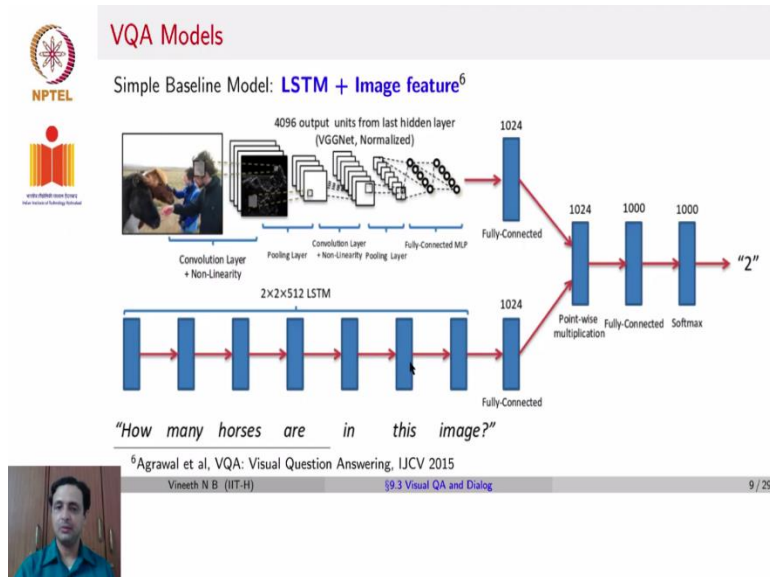
Vineeth N B (IIT-H) §9.3 Visual QA and Dialog 8 / 29

Another dataset more popular in recent times is known as CLEVR was developed in 2017. And this dataset is a semi-synthetic dataset, which helps answer reasoning questions. So the questions can be a bit complex. So given an image as what you see in the top left here, the question is, how big is the gray rubber object that is behind the big, shiny thing behind the big metallic thing? That is on the left side of the purple ball.

That sounds like a complex question. But if one reason, we are talking about this gray object at the back. This particular one is what we are talking about. And one has to ideally understand several subparts of the question to be able to answer such reasoning questions. So the CLEVR dataset has

about 100,000 rendered images, and 1 million automatically generated questions. The answers to these questions are single-word answers. But to get the answer, one needs to pass the question very carefully. The questions are indeed complex and require reasoning skills.

(Refer Slide Time: 12:07)



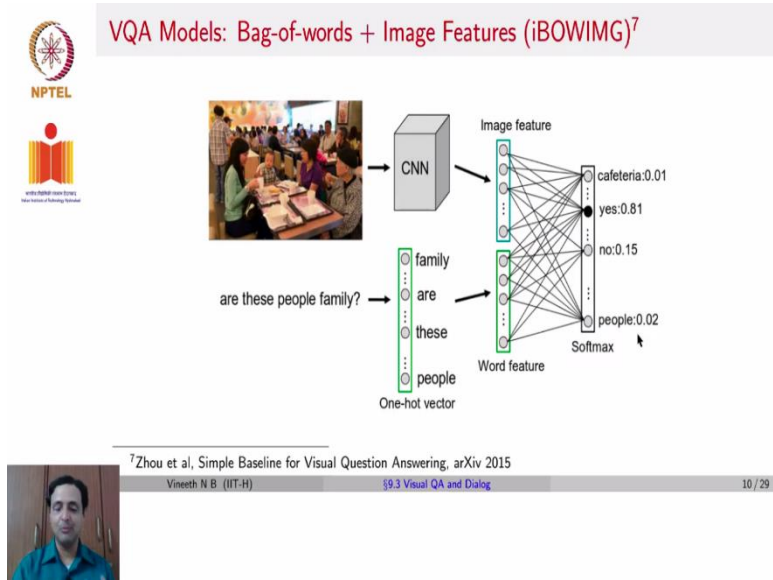
Having seen a few of the datasets that are used for Visual Question Answering, Let us now see a few models that have been developed for addressing the VQA problem. One of the baseline models that you can think of, for using for visual question answering is the combination of an LSTM and image features that you get out of a CNN. So given an image, you have a CNN, in this case, you see a VGG net.

And at the end of the CNN, you get a fully connected layers representation, which is in this case, 1024 dimensional. Similarly, for the question, you pass each word through different time steps of an LSTM. The output of an LSTM is a vector, which is again 1024 dimensions. You concatenate these two and send them through a few fully connected layers to make a final prediction of what the answer should be.

If your answers, were among a set of options, you would have cross-entropy loss. And everything that you see here all the parameters of every part of the network that you see here can be learned by a backpropagation on that cross-entropy loss. So I hope you understand how you can train these architectures.

Although we are adding the obstructions that we have learned so far, CNNs RNNs Attention, we are trying to mix match them to solve these problems, hope you can take away that none of these components violates how you can use backpropagation to learn the weights of the network. Using backpropagation and gradient descent stays constant through all of these kinds of architectures.

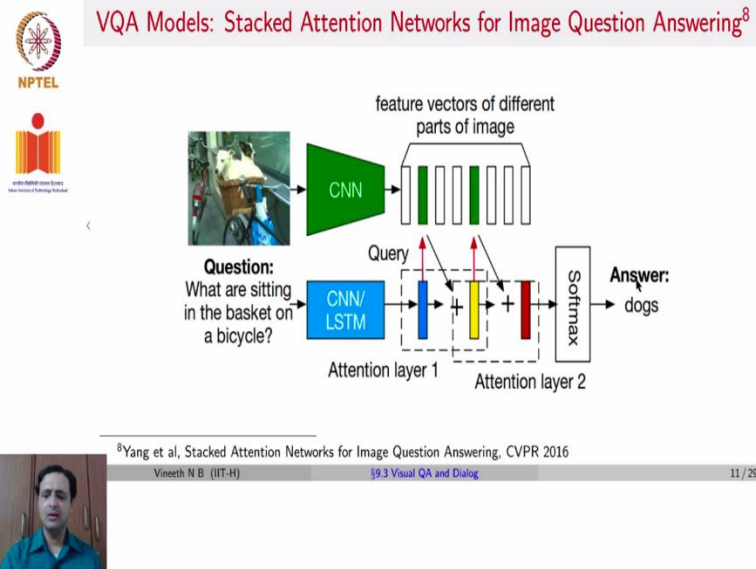
(Refer Slide Time: 14:12)



Another baseline model that was developed in 2015, for visual question answering was a simple method again, where given an image, an image feature was extracted out of one of the layers, one of the outputs of the layers of CNN. And for the question, a simple bag of words frequency histogram was used as the text input.

So these two were concatenated and using a simple neural network, the model predicted what was the right answer among a set of options. Having seen these simple baseline models, let us now try to see if you can improve upon these baseline models by getting more complex models.

(Refer Slide Time: 15:05)




One of the efforts in this direction was in CVPR of 2016, known as Stacked Attention Networks for Image Question Answering, in this particular model, given an image, you get a representation out of a CNN. And very similar to Attention models that we talked about on top of a CNN, you take certain convolution layers to feature map, you would then be able to map that to spatial locations in the original image.

Now, you have feature vectors corresponding to each part of your image. You now take the question and pass it through an LSTM. You can do convolution if you like, we will see this a couple of slides later. And the output of the LSTM gives you a representation of the question. This question is passed to the feature vectors of different parts of the image. And between the feature vectors of the image and the question representation, the model gives us a certain attention map on different parts of the image.

Based on this Attention map, and the representation of the question, there is another level of attention that is performed on the image feature vectors of different parts again. That leads to a different representation, which we expect is the final solution. And that representation is finally taken to the output layer, where a softmax over the options in answers gives you the outcome.

(Refer Slide Time: 16:56)

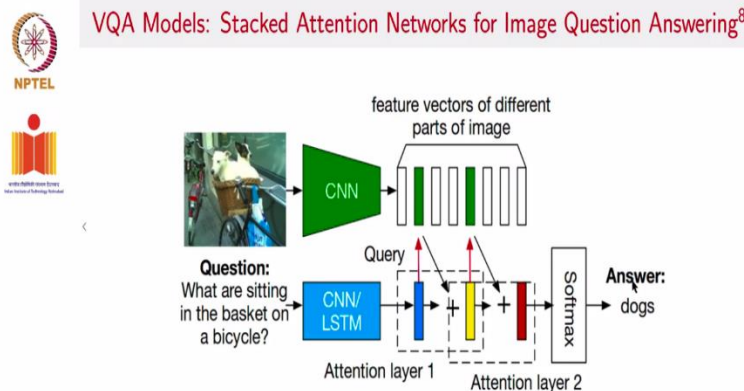

VQA Models: Stacked Attention Networks for Image Question Answering⁸




Original Image First Attention Layer Second Attention Layer

The stacked attention network **first focuses on all referred concepts**, e.g., *bicycle, basket* and objects in the basket (*dogs*) in the first attention layer

⁸Yang et al, Stacked Attention Networks for Image Question Answering, CVPR 2016
Vineeth N.B. (IIT-H) §9.3 Visual QA and Dialog 11 / 29








⁸Yang et al, Stacked Attention Networks for Image Question Answering, CVPR 2016
Vineeth N.B. (IIT-H) §9.3 Visual QA and Dialog 11 / 29



Let us see a visualization. So in this case, if you have the original image, what the stag attention network does is the first attention layer focuses on all referred concepts in the question. So we saw here in the previous slide, that the first attention layer decides which parts of the image to focus on and which parts does it focus on, wherever there are concepts that are present in the question. So it could be in this case, a bicycle basket, and objects in the basket. In this case, the question was, what is sitting in the basket on a bicycle?

(Refer Slide Time: 17:39)


VQA Models: Stacked Attention Networks for Image Question Answering⁸



Original Image First Attention Layer Second Attention Layer

The stacked attention network **first focuses on all referred concepts**, e.g., *bicycle*, *basket* and objects in the basket (*dogs*) in the first attention layer and **then further narrows down the focus in the second layer** and finds out the answer "**dog**".

⁸Yang et al, Stacked Attention Networks for Image Question Answering, CVPR 2016




Vineeth N B (IIT-H) §9.3 Visual QA and Dialog 11 / 29


The second attention layer, based on the attention that you get in the first attention layer, the second attention layer, chooses to focus only on the part of the image that contains the answer. And that is what helps us finally predict the outcome. So, in the second step, you narrow down the focus of the attention only on the part of the image that contains the answer. And hence, the model can predict the answer among the set of options as a dog.

(Refer Slide Time: 18:19)

VQA Models



Can we do attention on question as well?



Vineeth N B (IIT-H) §9.3 Visual QA and Dialog 12 / 29

This leads us to a follow-up question. So, in the model that we just saw, we performed attention only on the image, we took the question representation as it is, and try to use a relationship between the question representation and different parts of the image based on attention. So, even, in the attention that we just used on the previous slide, it was Soft attention, which weights different parts by certain weights that are learned and that can facilitate backpropagation. But now, we ask the question, can we also introduce attention in the question?

(Refer Slide Time: 19:06)

VQA Models: Hierarchical Co-Attention Model⁹

Given a question, extract its word level, phrase level and question level embeddings
 At each level, apply co-attention on both image and question
 Final answer prediction is based on all co-attended image and question features

⁹Lu et al, Hierarchical Question-Image Co-Attention for Visual Question Answering, NeurIPS 2016

Vineeth N B (IIT-H) 59.3 Visual QA and Dialog 12 / 29

This was done in a model known as the Hierarchical Co-Attention model. This was published in 2016 NeurIPS. In this particular model, the words give it all we have Co-Attention, which means we attend both on parts of the image, as well as parts of the question. And the hierarchical part comes from, we first get an embedding at a word level of the question, then add a phrase level of the question and finally, on the entire question itself.



Let us see this in more detail. Given a question, we extract its word-level embeddings, phrase-level embeddings, and question-level embeddings. At this point, you can assume that this is something that you get by passing through some neural network. So how do you what is the input in these cases? Do you give the text as input? Not really, remember, for all of these problems that involve text, for to a large extent, you have a vocabulary of words.

And each word in your question corresponds to a 1 hot vector on your vocabulary, where you put a one for the index in the vocabulary that this word belongs to, and 0 everywhere else. That would be the input that you give for a particular word. So you extract word level, phrase level, and question-level embeddings. At each level, you apply Co-Attention. So both for a word-level map embedding, you apply attention to the question, as well as the corresponding image.

Similarly, for the phrase level, and the question level, the final answer is based on all the Co-Attended image and question features. So if you see carefully here, when we say what color on the stoplight is lit up, so you can see here, this is the question, and the ideal answer you would expect is green. So you first take word-level embeddings. And you attend to one part of it. In this case, you can see that the attention is focused on stop. Similarly, at the same time, the attention on the image also seems to be focused on somewhere around the traffic light. In the next higher level of the hierarchy, the attention is focused on different phrases in the question. In this case, you can see the maximum weight goes to the phrase called the stoplight.

And once again, you can see here that the image is a bit more confident now and focuses completely on the stoplight. Then at the highest level of the hierarchy for the question, the attention map tells us that light is what we are focusing on. And at that stage, the attention on the image focuses only on the light inside the traffic signal. And that results in the answer green.

(Refer Slide Time: 22:35)

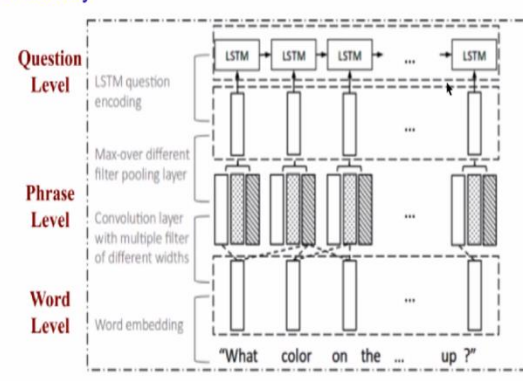
VQA Models: Hierarchical Co-Attention Model

Question hierarchy

Question Level

Phrase Level

Word Level



Vineeth N B (IIT-H)

§9.3 Visual QA and Dialog

13 / 29

So let us understand in some more detail the question hierarchy. So at a word level, you get a word-level embedding, as we just saw, how do we get the phrase level and question-level embeddings. To get the phrase-level embedding, you can take embeddings of each word of your caption, perform convolution across those inputs with multiple filters of different widths.

And you will get vector representations such as these, which are what we call phrase-level embeddings. Then we perform max pool over the convolution layer. And we then get a question-level embedding through the LSTM, which we give that max pool output. That gives us a representation at the question level.

(Refer Slide Time: 23:31)

Here are some results. So you see here, the image and the question are, what is the man holding a snowboard on top of a snow-covered? Question mark? The answer is Mountain. And the other example is what is the color of the bird? And the answer is White. Here are the word-level co-attention maps. In this case, the question the part of the question that is being focused on is snowboard top snow-covered and the image focus at that point is a bit on the snowboard and a bit on the person.

And in this case, the focus is on the color of the bird. And the model suggests attending to the part of the bird in the image. For phrase-level co-attention in this case, the focus is on holding snowboard, snowboard on top of snow-covered and the focus in the image red being high and blue

being low similar to heat maps that we saw with explanation methods. So you see that the model is focusing on other parts of the image other than the person.

Similarly for what is the color of the bird. The model seems to be focusing on the area just around the bird to be able to isolate the color. And finally, when the model comes to a question-level potential map, what is the man holding a snowboard? On top of a snow-covered? The model looks at everything other than the person to be able to get the answer, mountain. And in the other example, what is the color of the bird, the model is looking exactly at the bird to answer, white.

(Refer Slide Time: 25:24)

Visual Dialog: Task Overview¹⁰

NPTEL

Visual Dialog

A cat drinking water out of a coffee mug

What color is the mug?

White and red

Are there any pictures on it?

No, something is there can't tell what it is

Is the mug and cat on a table?

Yes, they are

Are there other items on the table?

Yes, magazines, books, toaster and basket, and a plate

¹⁰Das et al, Visual Dialog, CVPR 2017

Vineeth N B (IIT-H) §9.3 Visual QA and Dialog 15 / 29

We will now move on to the second task that we planned to discuss in this lecture, which you can consider as an extension of Visual QA, which is Visual Dialogue. So the overview of the task is given in the illustration here. This was first done in 2017. Given this image, and given this caption, which says a cat drinking water out of a coffee mug, Visual dialog is like a chatbot. But a chatbot that asks and answers questions with respect to an image rather than simple free text.

So let us see an example. What color is the mug? The bot answers, White and Red? So you can see the mug there? Are there any pictures on it? No. Something is there? Cannot tell what it is. Is the mug on Cat on a table? Yes, they are. Are there other items on the table? Yes, magazines, books, a toaster, and basket, and a plate. And as you can see there is a toaster in the background, there is a basket. And this is a fairly comprehensive list for a machine to answer Yes.

(Refer Slide Time: 26:53)



The screenshot displays a slide titled "Visual Dialog: CloudCV Demo". On the left, there are logos for NPTEL and IIT Bombay. The main content features an image of a man and a woman holding umbrellas. Below the image is a chat interface with the following messages:

- Bot: A man and a woman are holding umbrellas
- Human: What color is his umbrella?
- Bot: His umbrella is black
- Human: What about hers?
- Bot: Hers is multi-colored
- Human: How many other people are in the image?
- Bot: I think 3. They are occluded

Below the chat, there is a text prompt: "You can try Visual Dialog on this link!". At the bottom of the slide, there is a small video feed of a man, the name "Vineeth N B (IIT-H)", and the slide number "16 / 29".

So here is a demo of how visual dialogue works. Given this image. And the bot says a man and a woman are holding umbrellas. The question that a human can ask is what color is his umbrella? The bot answers his umbrella is black. What about hers? This is a tricky question. Because the model has to understand the question is about the umbrella. And secondly, the question is about her umbrella.

This means the model needs to know that there are two people, one of them is a male, one of them is a female, and then be able to answer this question. And the bot does say hers is multicolored? How many other people are in the image? I think three they are occluded. Once again, this demo comes from Cloud CV developed by Dhruv Batra and Devi Parekh Gatech, you are welcome to go to this link. Try out visual dialogue, and see for yourself how it works.

(Refer Slide Time: 28:01)

Visual Dialog: Task Description

Given

- Image I
- Human dialog history $(Q_1, A_1), (Q_2, A_2), \dots, (Q_{t-1}, A_{t-1})$
- Follow-up question Q_t

Task

- Produce free-form natural language answer A_t

Credit: Abhishek Das, Georgia Tech



Vineeth N B (IIT-H) 59.3 Visual QA and Dialog 17/29

Q: How many people on wheelchairs?
A: Two
Q: What gender are the people in the wheelchairs?
A: One is female, one is male
Q: Which one is holding the racket?
A: The female
Q: Is the other one holding anything?
A: He is not

Let us now formally define how you would go about solving this task. This is not like a traditional classification problem does not even seem like a traditional sequence learning problem where you give out a caption for an image, then how would you go about formalizing it and solving it. Given an image, I have a dialogue history of question-answer couples $Q_1 A_1, Q_2 A_2$ so on and so forth. Till $Q_{t-1} A_{t-1}$.

Those are your previous question. Answer topples. And you have a follow-up question that is current at this time. In this case, the question is, is the other one holding anything? And the task here is to produce a free form natural language answer A_t that answers this current question, Q_t .


(Refer Slide Time: 29:00)



Visual Dialog: Evaluation

- A fundamental challenge in dialog systems is **evaluation**¹¹
- Existing **word-overlap based metrics** such as BLEU, METEOR, ROUGE are known to **correlate poorly with human judgement**
- **Human Turing test**
 - expensive
 - subjective

¹¹Liu et al, How NOT To Evaluate Your Dialogue System, EMNLP 2016
Vineeth N B (IIT-H) 39.3 Visual QA and Dialog 18 / 29



So one question you could ask here is, how do you evaluate the answer? The evaluation here can be tricky and typical scores used in natural language processing such as BLEU, which we saw earlier METEOR, ROUGE. We would not get into them now. But it is known that such scores often correlate poorly with human judgment. For example, you could have multiple answers set in different ways that could answer a question, but not all of them may correlate with a score of one single ground truth answer.

So how do you study, study this, you could probably have what is known as Human Turing tests, where you give out a generated sentence, and then ask a bunch of humans to rate the quality of the sentence. Now, this could work, because based on how a human looks at it, it is called the Human Turing test to check whether the human can make out whether the answer is generated from a machine, or whether the answer is generated by a human.

So you could introduce that as one of the questions that you ask the human in addition to the rating, the quality of the answer. The challenge here is that this kind of evaluation could become expensive, because you may have to compensate humans to participate in the evaluation. And it could also be subjective, depending on what human is evaluating a system. So different humans could evaluate different answers differently. So how do we go about evaluating such a system?

(Refer Slide Time: 30:50)

Visual Dialog: Evaluation Protocol

- Given**
 - Image I
 - Human dialog history $(Q_1, A_1), (Q_2, A_2), \dots, (Q_{t-1}, A_{t-1})$
 - Follow-up question Q_t
 - 100 answer options
 - Answers to 50 most similar questions
 - 30 popular answers
 - 20 random answers
- Evaluation Task**
 - Rank the list of 100 options
- Accuracy/Error**
 - Mean rank w.r.t. ground truth, Mean reciprocal rank

Credit: Dhruv Batra, Georgia Tech

Vineeth N.B. (MIT-H) §9.3 Visual QA and Dialog 19 / 29

Visual QA Dialog Example:

Q: How many people on wheelchairs?
A: Two

Q: What gender are the people in the wheelchairs?
A: One is female, one is male

Q: Which one is holding the racket?
A: The female

Q: Is the other one holding anything?
A: He is not

Handwritten Calculations:

$$\frac{1 + \frac{1}{2} + \frac{1}{3}}{3} = 1.83$$
$$\frac{1}{3} = 0.61$$

So let us see that now. So given an image, the dialogue history, and the current question, what datasets for visual and dialogue do is give about 100 answer options? These answer options could be answers to 50 most similar questions 30, popular answers, and maybe 20, random answers. So you pick these 100 answers and ask your model to rank all of these answers as the outcome.

So the model's job is, given these 100 answers, you could once again have a softmax, which gives you a probability vector across these 100 options. But using that probability vector, you can now rank order your options, the highest probability will be the first rank, the second-highest probability will be the second rank, so on and so forth. But how do you judge the goodness of this ranking?

You can do a couple of things to measure the performance, you could check the mean rank with respect to the ground truth. So you expect that the model will give the first answer as the ground truth. But that may not always happen. In one case, you may get the third rank answer to be the ground truth. In one case, you could get the seventh rank answer to be the ground truth. And in one case, the 41st answer that you rank could be the ground truth, which would be a poor outcome.

So the mean rank with respect to the ground truth across all your test samples could be one performance metric that you could use. Another that is another metric that is also used is known as the mean, reciprocal rank. So the mean, reciprocal rank could be for one particular test question

if you predicted a set of ranks for the options, and the third position of your predictions was the correct answer.

So which means your correct rank is three? Let us assume that for another question. Your second-ranked option was the correct answer. And for another question, your first ranked option was the correct answer, then your mean reciprocal rank would be given by $1 + \frac{1}{2} + \frac{1}{3}$, divided by 3 because those are the total number of options that you have, which you would get as 1.5, you would get that to be 1.5×3 , which gives you 0.61.

This would be the mean reciprocal rank? When would the mean reciprocal rank be 1, if your first rank option always is the ground truth? And when would it be 0, if your last rank option was your ground. So, mean reciprocal rank is another performance metric that you can use to be able to study the performance.

(Refer Slide Time: 33:58)

Visual Dialog: Models

Encoder-Decoder frameworks

- Encoders
 - Late Fusion Encoder
 - Hierarchical Recurrent Encoder
 - Memory Network Encoder
- Decoders
 - Generative
 - Discriminative

Credit: Abhishek Das, Georgia Tech

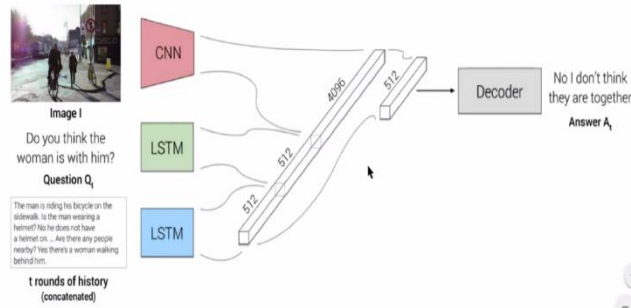
Vineeth N.B. (IIT-H) 39.3 Visual QA and Dialog 20 / 29

Let us now come to the kind of models you can use to perform visual dialogue tasks. So it is once again encoder-decoder frameworks, where the encoder could be a late fusion encoder, a hierarchical recurrent encoder, or a memory network encoder. And the decoder can be generative, or discriminative. We will see some of these examples over the next few slides.

(Refer Slide Time: 34:33)



Visual Dialog: Late Fusion Encoder



Vineeth N B (IIT-H)

§9.3 Visual QA and Dialog

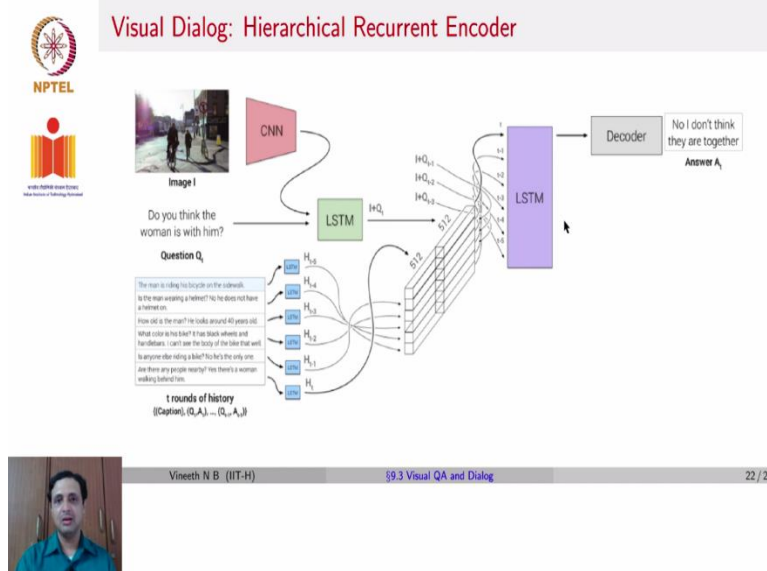
21 / 29

Here is an example of a Late Fusion Encoder. In this particular case, the image is the one that you see here. You have a previous history of questions and answers so the man is riding his bicycle on the sidewalk. Is the man wearing a helmet? No, he does not have a helmet on. Are there any people nearby? Yes, a woman is walking behind him. That is the dialogue history. And the current question Q_t is given by Do you think the woman is with him?

So, we have to answer A_t for this question now. So, you provide the image to a CNN, you provide the current question to an LSTM and you provide the entire history concatenated as single text input to another LSTM, you fuse or concatenate the outputs of each of these models, the CNN, the current question LSTM, and the history LSTM.

And you can now have layers after them to reduce the dimension of that final representation and you provide this to a decoder to get the final answer. We will talk about decoders in a moment, but at this point, we are discussing the encoders of models that are used for visual dialogue. This is one kind of model known as the Late Fusion Encoder.

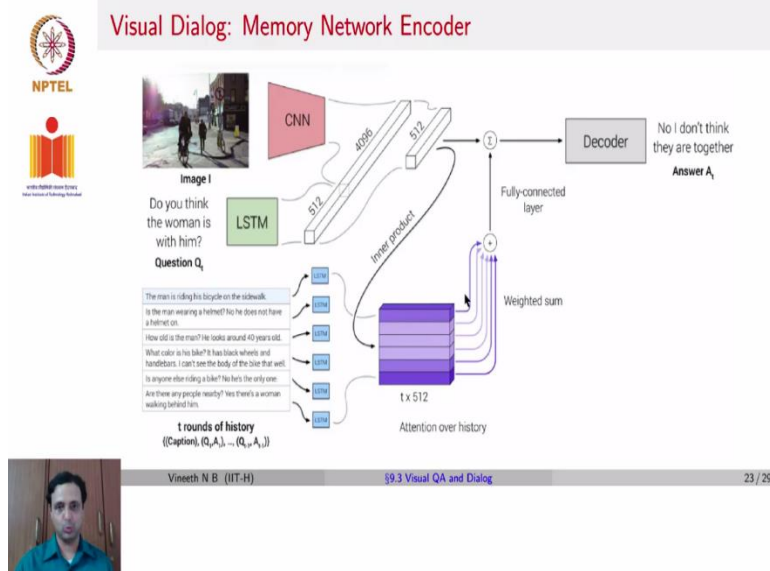
(Refer Slide Time: 36:05)



Another kind of an encoder is a Hierarchical Recurrent Encoder, where you have once again an image goes through a CNN gets a representation, the current question goes to an LSTM. But now, this LSTM also gets the image representation as input, then for each round of the history of the dialogue history, you send it through one instance of the LSTM and get different representations for each step in history.

Now, you have another LSTM that combines these inputs from, from previous dialog history and the current question and the image and the output of that next level of the STM, LSTM is passed on to the decoder to get the output. So, this is called a Hierarchical Recurrent Encoder, because you have an LSTM, to take inputs of the image, current question, and the dialogue. And you have another LSTM that combines these inputs before passing them on to the decoder. This is the Hierarchical Recurrent Encoder.

(Refer Slide Time: 37:23)



The third kind of encoder is a Memory Network Encoder, which uses an attention idea where once again, the image through a CNN question through an LSTM, concatenate representation, compress the representation. That is the first part. Once again, each of the dialogues in history goes through a different instance of the same LSTM and you get different representations for each of these question answers in the dialogue history.

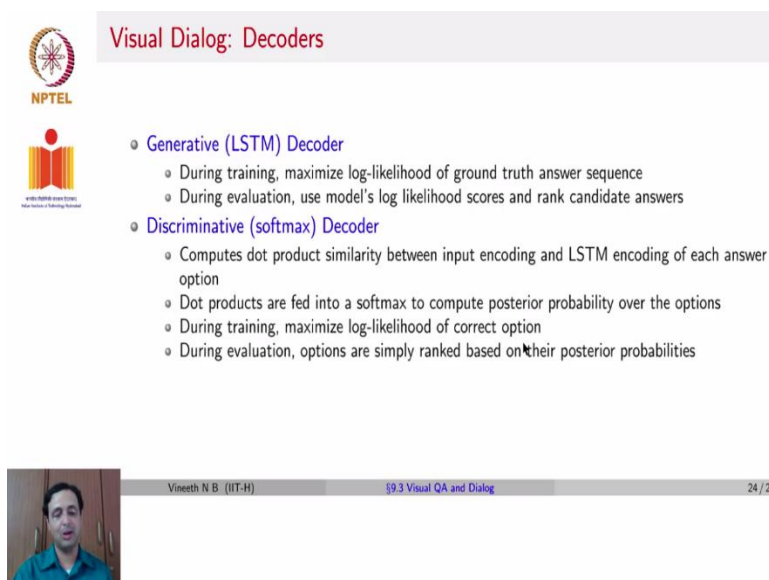
Now, you also bring the current image question representation, along with representations of the history. And you now look at an Attention step here you see attention or history to decide which part of the dialogue is relevant to answer the current question, it remembers for a long dialogue, you may be referring to a sentence that was stated 2-3 conversations before to answer a current question.

And using an attention idea gives us kind of a different approach. So you take an inner product of the current question and image which gives you an idea of the alignment. Remember, we talked about alignment when we talked about attention for encoder-decoder models, where we talked about the attention of the hidden state in a decoder with respect of the alignment of a hidden state in the decoder with respect to a hidden state in the encoder, and we said that one way you could do alignment is through a dot product or inner product.

So, what you see here is attention which checks for alignment between the current image and the question with respect to history, to get an attention map over the history, which is a weighted sum, then you combine this with a current feature vector of the image and current question and then you pass that to the decoder to get the final answer.

Once again to remind you, every component that you see here is differentiable. And you can learn all of these through the back propagation step. Even attention or history is simply a weighted combination of inputs, which can be backpropagated.

(Refer Slide Time: 39:48)



The slide is titled "Visual Dialog: Decoders" and features the NPTEL logo on the left. It contains two main bullet points:

- **Generative (LSTM) Decoder**
 - During training, maximize log-likelihood of ground truth answer sequence
 - During evaluation, use model's log likelihood scores and rank candidate answers
- **Discriminative (softmax) Decoder**
 - Computes dot product similarity between input encoding and LSTM encoding of each answer option
 - Dot products are fed into a softmax to compute posterior probability over the options
 - During training, maximize log-likelihood of correct option
 - During evaluation, options are simply ranked based on their posterior probabilities

At the bottom of the slide, there is a small video feed of a man in a blue shirt, and a footer containing the text "Vineeth N B. (IIT-H)", "9.3 Visual QA and Dialog", and "24 / 29".

Coming to the decoder part of such models as we said, decoders can be Generative or Discriminative. In a generative decoder as an example, during training, you would maximize the log-likelihood of the ground truth answer sequence; you would want your decoder to generate the correct answer sequence. And during training, you maximize the log-likelihood of generating that ground truth answer sequence.

Once again, when you implement them, these would be implemented as if you had a decoder LSTM, at each step, you would output a word that would give you a cross-entropy or vocabulary. And you would have multiple time steps in your LSTM. And the addition of that cross-entropy would be your loss. Or in this case, the log-likelihood cross-entropy is a way of minimizing cross-entropy is equivalent to minimizing negative log-likelihood, in this case.

During a test time or evaluation, you use the models, log-likelihood scores over your vocabulary, and rank your candidate answers. Whereas in a discriminative decoder, you take the input encoding and compute a dot product similarity between the input encoding and the LSTM encoding of each answer option. You could use any encoding method here, we will not get into that here, you probably need some knowledge of NLP to understand how you can get encodings of text.

But let us assume now that you could get some encoding there are what are known as Word to Vec embeddings, Glove embeddings, so on and so forth. You could use them to get an encoding for the answer options, and the input encoding. Now you compute a dot product, and that dot product is fed into a softmax to compute posterior probability over all your answer options.

And during training, you would maximize the log-likelihood of the correct option. And a test time or evaluation options are simply ranked based on their posterior probabilities, which you would get as the output of a softmax. This is how you could model the decoder of such encoder-decoder models to solve the visual dialogue task.

(Refer Slide Time: 42:27)

Visual Dialog: Results

NPTEL

INDIAN INSTITUTE OF TECHNOLOGY HAWAII

Visual Dialog: Results

Person A (1): how many skiers are there
Person B (1): hundreds

Person A (2): are they getting ready to go downhill
Person B (2): i think so my view is at end of line

Person A (3): is it snowing
Person B (3): no, there is lot of snow though

Person A (4): can you see anybody going downhill
Person B (4): no my view shows people going up small hill on skis i can't see what's going on from there

Person A (5): do you see lift
Person B (5): no

Person A (6): can you tell if they are male or female
Person B (6): skiers closest to me are male

Person A (7): are there any children
Person B (7): i don't see any but there could be it's huge crowd

Person A (8): does anybody have hat on
Person B (8): they all have winter hat of some sort on

Person A (9): is sun shining
Person B (9): yes, all blue sky



Person A (10): do you see any clouds
Person B (10): no clouds

Vineeth N B (IIT-H) §9.3 Visual QA and Dialog 25 / 29


Let us see a couple of examples. So here you have an image. The initial caption is the skiers stood on top of the mountain. Then you have the dialogue. How many skiers are there? 100s? Are they ready? Are they getting ready to go downhill? I think so. My view is at the end of the line. Is it snowing? No. There is a lot of snow though. Can you see anybody going downhill? No, my view

shows people going up a small hill in the skies. I cannot see what is going on from there. And so on, as you can see is a meaningful dialogue.

(Refer Slide Time: 43:14)



Visual Dialog: Results



Caption: A dog with goggles is in a motorcycle side car.

Person A (1): can you tell what kind of dog this is

Person B (1): he looks like beautiful pit bull mix

Person A (2): can you tell if motorcycle is moving or still

Person B (2): it's parked

Person A (3): is dog's tongue lolling out

Person B (3): not really

Person A (4): is picture in color

Person B (4): yes it is

Person A (5): what color is dog

Person B (5): light tan with white patch that runs up to bottom of his chin and he has white paws on 2 front feet

Person A (6): can you see motorcycle

Person B (6): from side, yes

Person A (7): what color is motorcycle

Person B (7): black with white or silver accents, sun is glaring so it's hard to tell

Person A (8): is there anybody sitting on motorcycle


Person B (8): no

Person A (9): is there anybody in picture

Person B (9): in cars on street behind motorcycle

Person A (10): does dog look like he's having fun



Person B (10): yes



Vineeth N B. (IIT-H) §9.3 Visual QA and Dialog 25 / 29

One more example, for you to see at your leisure for Visual Dialog.

(Refer Slide Time: 43:21)




Homework

Questions

- Are there better methods to evaluate Visual Dialog systems? How?

Readings (Optional)

- VQA
 - Deep Compositional Question Answering with Neural Module Networks, CVPR 2016
 - Learning to Reason: End-to-End Module Networks for Visual Question Answering, ICCV 2017
 - MUTAN: Multimodal Tucker Fusion for Visual Question Answering, ICCV 2017
 - Learning Conditioned Graph Structures for Interpretable Visual Question Answering, NeurIPS 2018
 - Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering, CVPR 2018
 - Towards VQA Models That Can Read, CVPR 2019
 - Deep Modular Co-Attention Networks for Visual Question Answering, CVPR 2019



Vineeth N B. (IIT-H) §9.3 Visual QA and Dialog 26 / 29



If you would like to know more, you can read these papers. There are some of them, which are more recent, which we did not get a chance to cover in this lecture. We covered the most, most

basic methods. But if you would like to know more, you can follow some of these recent papers. And we are going to leave one question here. Can you come up with better methods to evaluate visual dialogue systems? Think about it may not be a nontrivial answer may or may not be a trivial answer. Think about it. And we will discuss the next time.

(Refer Slide Time: 43:58)


A few more readings for Visual Dialogue papers, including some very recent papers that could be relevant.

(Refer Slide Time: 44:06)



References I

- Mateusz Malinowski and M. Fritz. "A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input". In: *NIPS*. 2014.
- Aishwarya Agrawal et al. "VQA: Visual Question Answering". In: *International Journal of Computer Vision* 123 (2015), pp. 4–31.
- Mengye Ren, Ryan Kiros, and R. Zemel. "Exploring Models and Data for Image Question Answering". In: *NIPS*. 2015.
- A. Fukui et al. "Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding". In: *ArXiv abs/1606.01847* (2016).
- C. Liu et al. "How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation". In: *ArXiv abs/1603.08023* (2016).
- Jiasen Lu et al. "Hierarchical Question-Image Co-Attention for Visual Question Answering". In: *ArXiv abs/1606.00061* (2016).



Vineeth N B. (IIT-H) §9.3 Visual QA and Dialog 28 / 29

Here are some references.