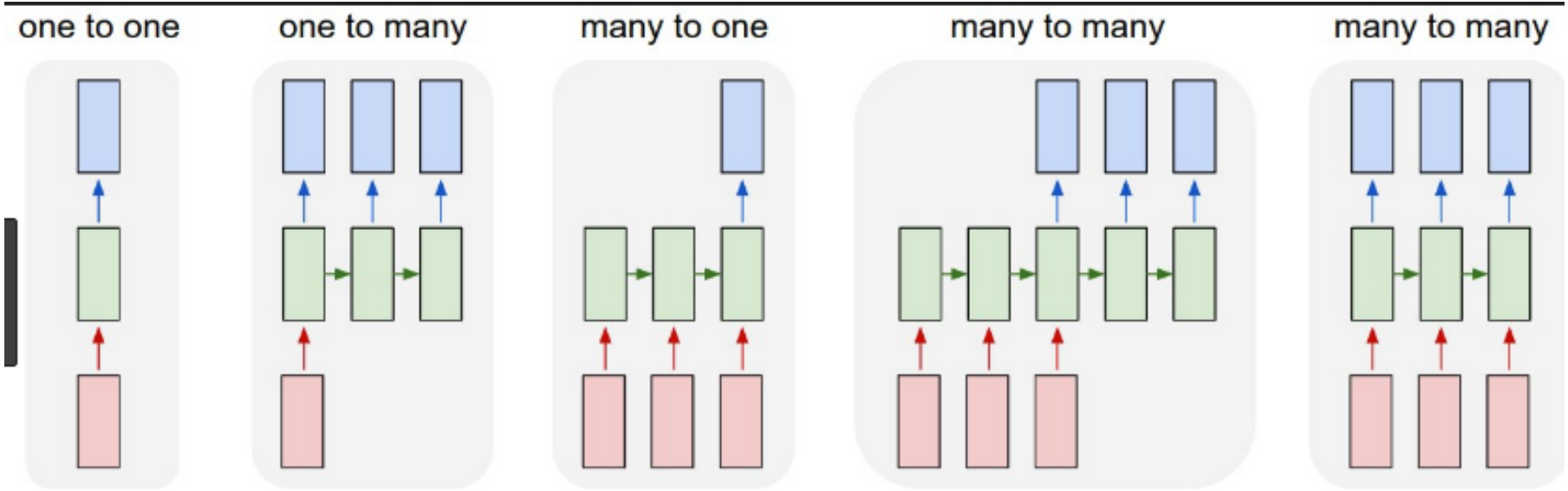
# Attention!

Mikhail Yurushkin

Brouton  Lab

# Recap



# SMT (Statistical Machine Translation)

- Requires skills in linguistics, morphological analyzes
- Context capturing issues
- Many hand written rulles - difficult to implement

# BLEU (2002) - bilingual evaluation understudy

Algorithm for evaluating the quality of text which has been machine-translated from one natural language to another

**Example of poor machine translation output with high precision**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Candidate** | the | the | the | the | the | the | the |
| **Reference 1** | the | cat | is | on | the | mat | |
| **Reference 2** | there | is | a | cat | on | the | mat |

$$P = \frac{m}{w_t} = \frac{7}{7} = 1$$

m - is number of words from the candidate that are found in the reference

Wt - is the total number of words in the candidate

- It makes sense to consider “the” no more than twice.
- N-grams usage (N=4 often) is preferable
- Penalizing for short sentences
- Not ideal, but you’re welcome to suggest something better!

# Sequence 2 sequence

- Can we just use simple LSTM ?!

Ad hoc approaches:

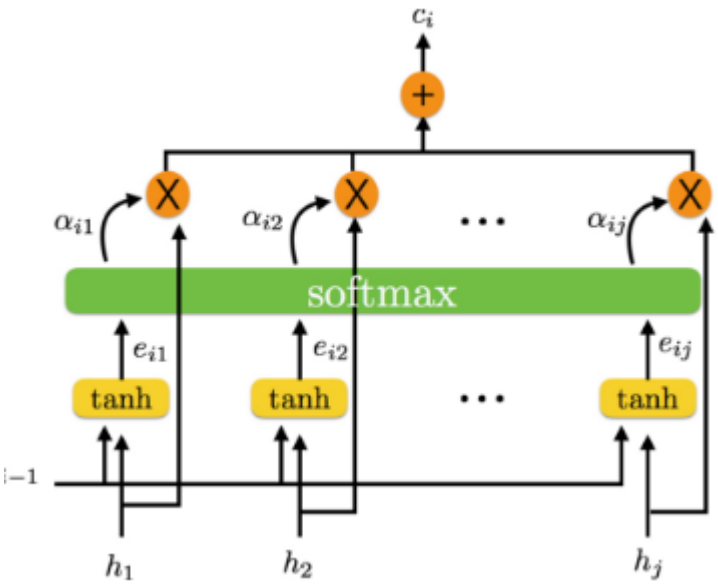
- Splitting matrices (encoder/decoder architecture)
- Repeating hidden state to all decoder cells
- Stacking of LSTM cells
- Feeding previous generated words in decoder
- Reversing of input sentences

**Sequence to Sequence Learning with Neural Networks’14**

https://arxiv.org/abs/1409.3215

# Attention mechanism

**Neural Machine Translation by Jointly Learning to Align and Translate'14**



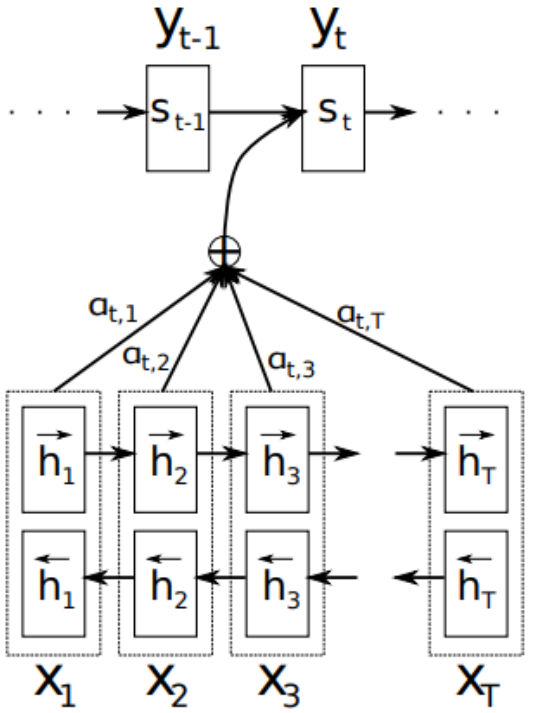
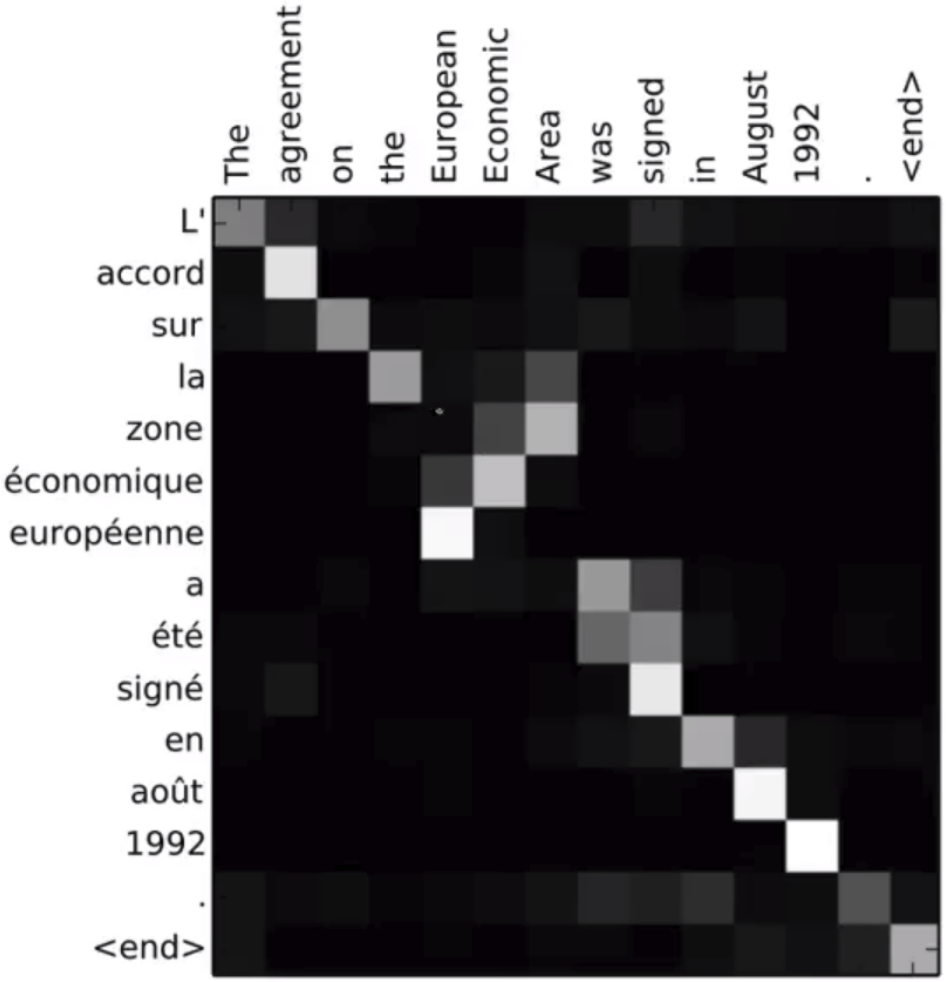




Figure 1: The graphical illustration of the proposed model trying to generate the $t$-th target word $y_t$ given a source sentence $(x_1, x_2, \ldots, x_T)$.

# Attention visualization



# Attention helps to process long sentences

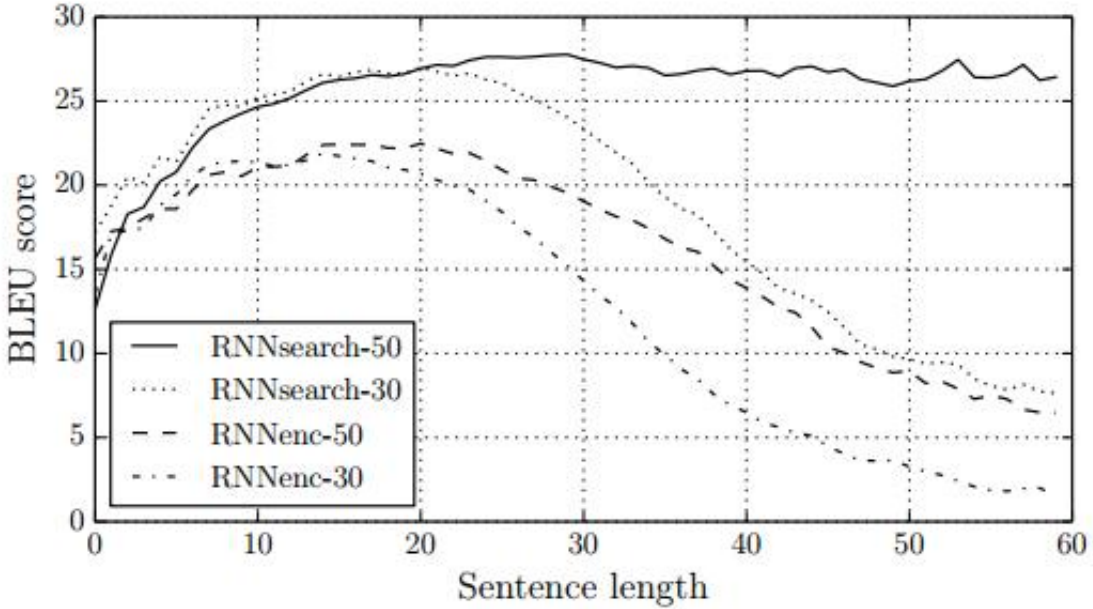



Figure 2: The BLEU scores of the generated translations on the test set with respect to the lengths of the sentences. The results are on the full test set which includes sentences having unknown words to the models.

**Sequence to Sequence Learning with Neural Networks'14**

https://arxiv.org/abs/1409.3215

# Attention = (Fuzzy) Memory?

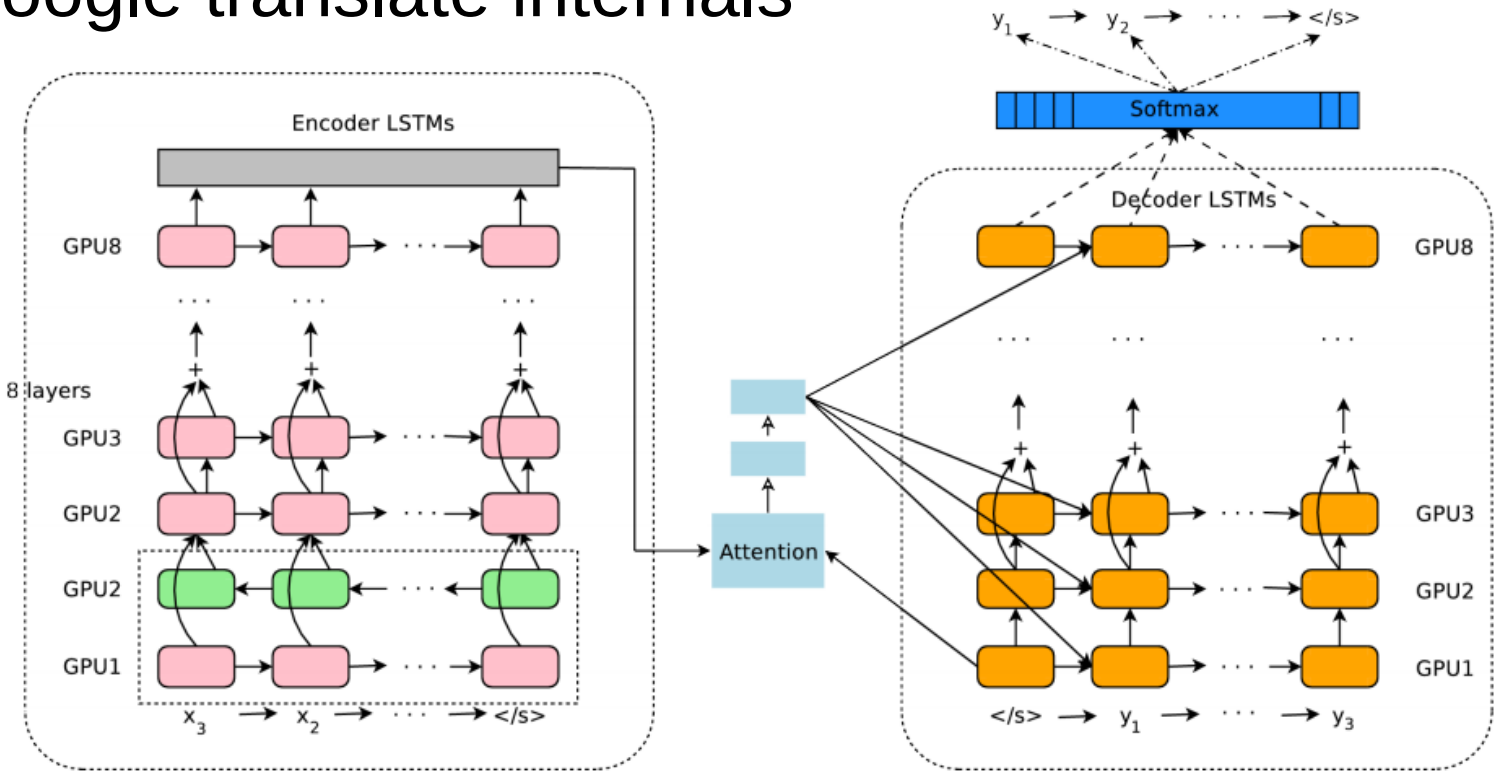
RNN has memory BUT ...

It suffers from vanishing gradients problem :(

Pros of attention:

- Allows the network to refer back to the input sequence
- Unlike typical memory, the differentiable memory is soft. We can train the network end-to-end using back-propagation

# Google translate internals




**Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation'16**

# Google translate (before and after)

It’s supposed that NMT has better language model (in comparison with SMT)

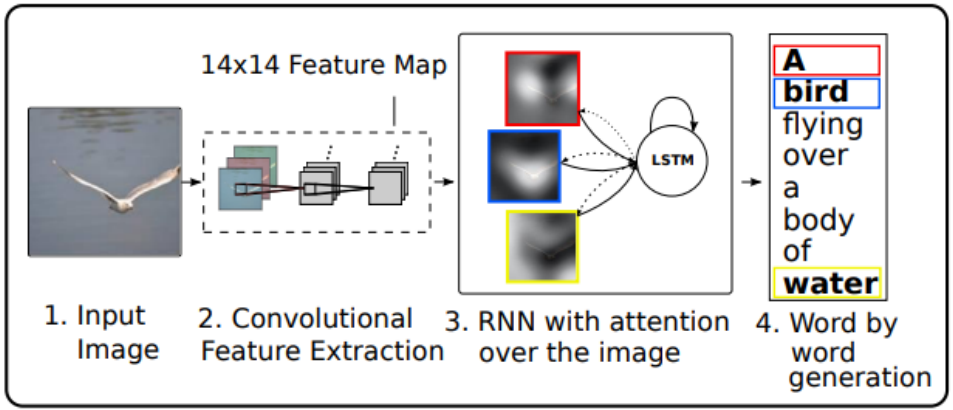
Before:

Kilimanjaro is 19,710 feet of the mountain covered with snow, and it is said that the highest mountain in Africa. Top of the west, “Ngaje Ngai” in the Maasai language, has been referred to as the house of God. The top close to the west, there is a dry, frozen carcass of a leopard. Whether the leopard had what the demand at that altitude, there is no that nobody explained.

After:

Kilimanjaro is a mountain of 19,710 feet covered with snow and is said to be the highest mountain in Africa. The summit of the west is called “Ngaje Ngai” in Masai, the house of God. Near the top of the west there is a dry and frozen dead body of leopard. No one has ever explained what leopard wanted at that altitude.

# Attention for image captioning





**Show, Attend and Tell: Neural Image Caption Generation with Visual Attention'15**

# Examples

*Figure 3.* Examples of attending to the correct object (*white* indicates the attended regions, *underlines* indicated the corresponding word)

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.


A stop sign is on a road with a mountain in the background.

A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.

# Visual Q&A

What vegetable is on the plate?
Neural Net: broccoli
Ground Truth: broccoli

What color are the shoes on the person's feet ?
Neural Net: brown
Ground Truth: brown

How many school busses are there?
Neural Net: 2
Ground Truth: 2

What sport is this?
Neural Net: baseball
Ground Truth: baseball

What is on top of the refrigerator?
Neural Net: magnets
Ground Truth: cereal

What uniform is she wearing?
Neural Net: shorts
Ground Truth: girl scout


What is the table number?
Neural Net: 4
Ground Truth:40

What are people sitting under in the back?
Neural Net: bench
Ground Truth: tent

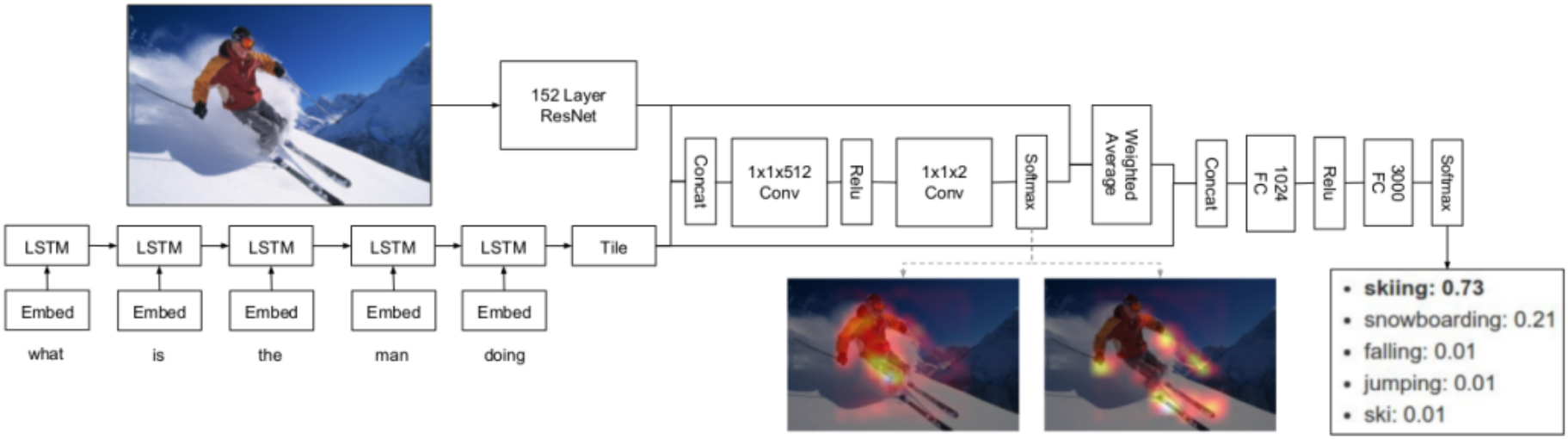



Figure 2. An overview of our model. We use a convolutional neural network based on ResNet [9] to embed the image. The input question is tokenized and embedded and fed to a multi-layer LSTM. The concatenated image features and the final state of LSTMs are then used to compute multiple attention distributions over image features. The concatenated image feature glimpses and the state of the LSTM is fed to two fully connected layers two produce probabilities over answer classes.

**Show, Ask, Attend, and Answer: A Strong Baseline For Visual Question Answering'17**

- **nike: 0.45**
- adidas: 0.29
- polo: 0.15
- reebok: 0.02
- unknown: 0.02

(a) What brand is the shirt?

- **sunset: 0.28**
- evening: 0.24
- dusk: 0.18
- dawn: 0.08
- night: 0.04

(b) What time is it?

- **happy: 0.42**
- sad: 0.24
- serious: 0.07
- tired: 0.05
- neutral: 0.05

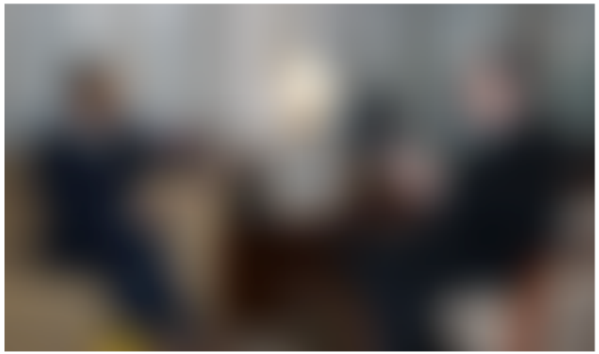
(c) How does the man feel?

- **skateboarding: 0.60**
- skating: 0.09
- playing tennis: 0.06
- skateboard trick: 0.03
- tennis: 0.02

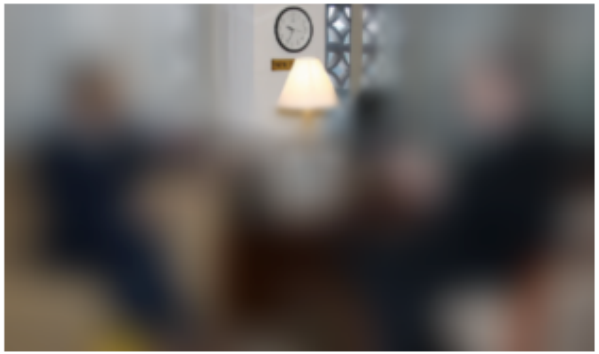
(d) What is the girl doing?

Figure 3. Qualitative results on sample images shows that our model can produce reasonable answers to a range of questions.

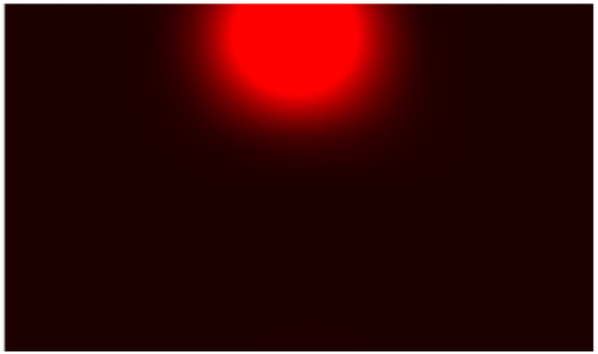
# Is there any correlation between human and model attention?



(a) Initial blurred image

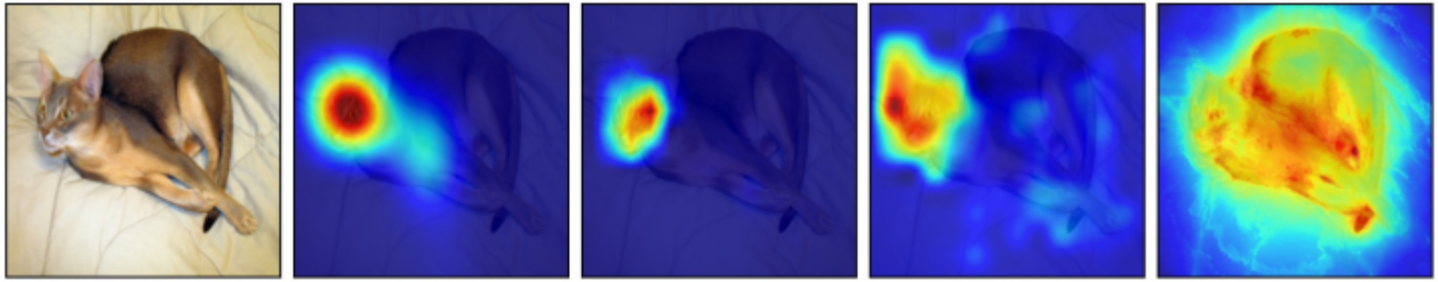


(b) Regions sharpened by subject



(c) Attention map

**Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions?'16**



What color are the animal's eyes? green | Human Attention | SAN-2 (Yang et al.) Correlation: 0.573 | HieCoAtt-Q (Lu et al.) Correlation: 0.527 | Judd et al. Correlation: 0.491

What is covering the windows? blinds | Human Attention | SAN-2 (Yang et al.) Correlation: -0.495 | HieCoAtt-Q (Lu et al.) Correlation: -0.440 | Judd et al. Correlation: 0.078

What is the man doing? playing frisbee | Human Attention | SAN-2 (Yang et al.) Correlation: -0.060 | HieCoAtt-Q (Lu et al.) Correlation: 0.238 | Judd et al. Correlation: 0.423

| Model | Rank-correlation |
|---|---|
| SAN-2 (Yang et al., 2015) | $0.249 \pm 0.004$ |
| HieCoAtt-W (Lu et al., 2016) | $0.246 \pm 0.004$ |
| HieCoAtt-P (Lu et al., 2016) | $0.256 \pm 0.004$ |
| HieCoAtt-Q (Lu et al., 2016) | $0.264 \pm 0.004$ |
| Random | $0.000 \pm 0.001$ |
| Judd et al. (Judd et al., 2009) | $0.497 \pm 0.004$ |
| Human | $0.623 \pm 0.003$ |

Table 2: Mean rank-correlation coefficients (higher is better); error bars show standard error of means. We can see that both SAN-2 and HieCoAtt attention maps are positively correlated with human attention maps, but not as strongly as task-independent Judd saliency maps.

| Model | Rank-correlation |
|---|---|
| SAN-2 (Yang et al., 2015) | $0.038 \pm 0.011$ |
| HieCoAtt-W (Lu et al., 2016) | $0.062 \pm 0.012$ |
| HieCoAtt-P (Lu et al., 2016) | $0.048 \pm 0.010$ |
| HieCoAtt-Q (Lu et al., 2016) | $0.114 \pm 0.012$ |
| Judd et al. (Judd et al., 2009) | $-0.063 \pm 0.009$ |

Table 3: Mean rank-correlation coefficients (higher is better) on the reduced set without center bias; error bars show standard error of means. We can see that correlation goes down significantly for Judd saliency maps since they have a strong center bias. Relative trends among SAN-2 & HieCoAtt are similar to those over the whole validation set (reported in Table 2).