From Recognition to Cognition: Visual Commonsense Reasoning

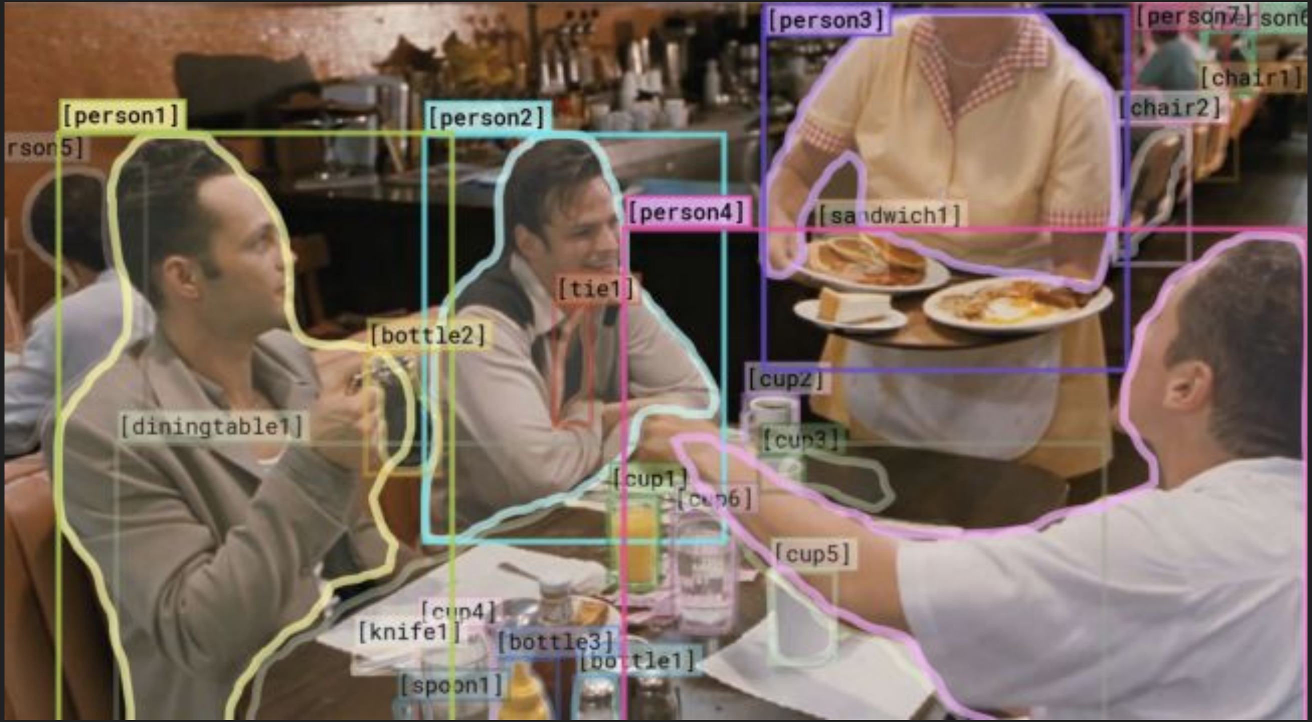
Rowan Zellers, Yonatan Bisk, Ali Farhadi, Yejin Choi

Paul G. Allen School of Computer Science & Engineering, University of Washington
Allen Institute for Artificial Intelligence

Presented by - Amish Mittal, Indian Institute of Technology Patna

CVPR 2019, https://visualcommonsense.com/

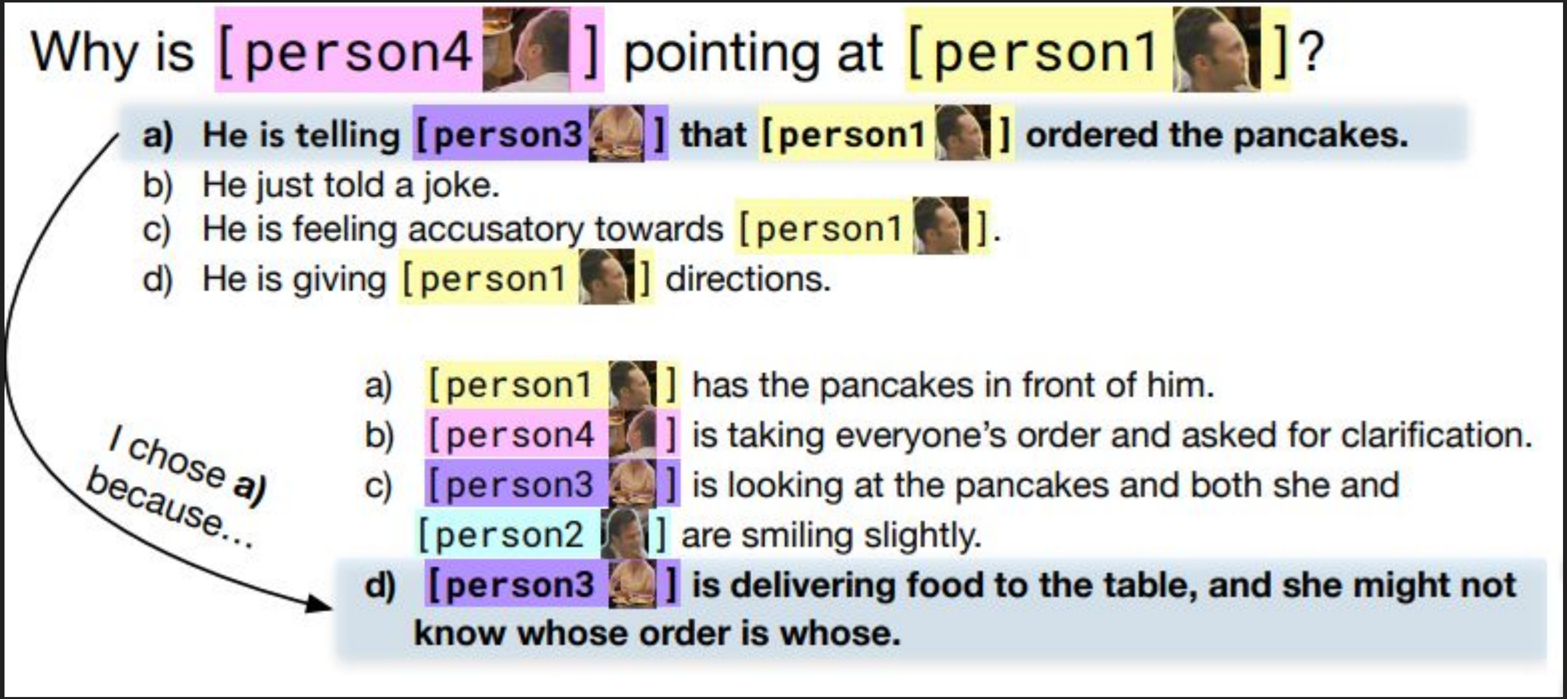
Defining the Problem



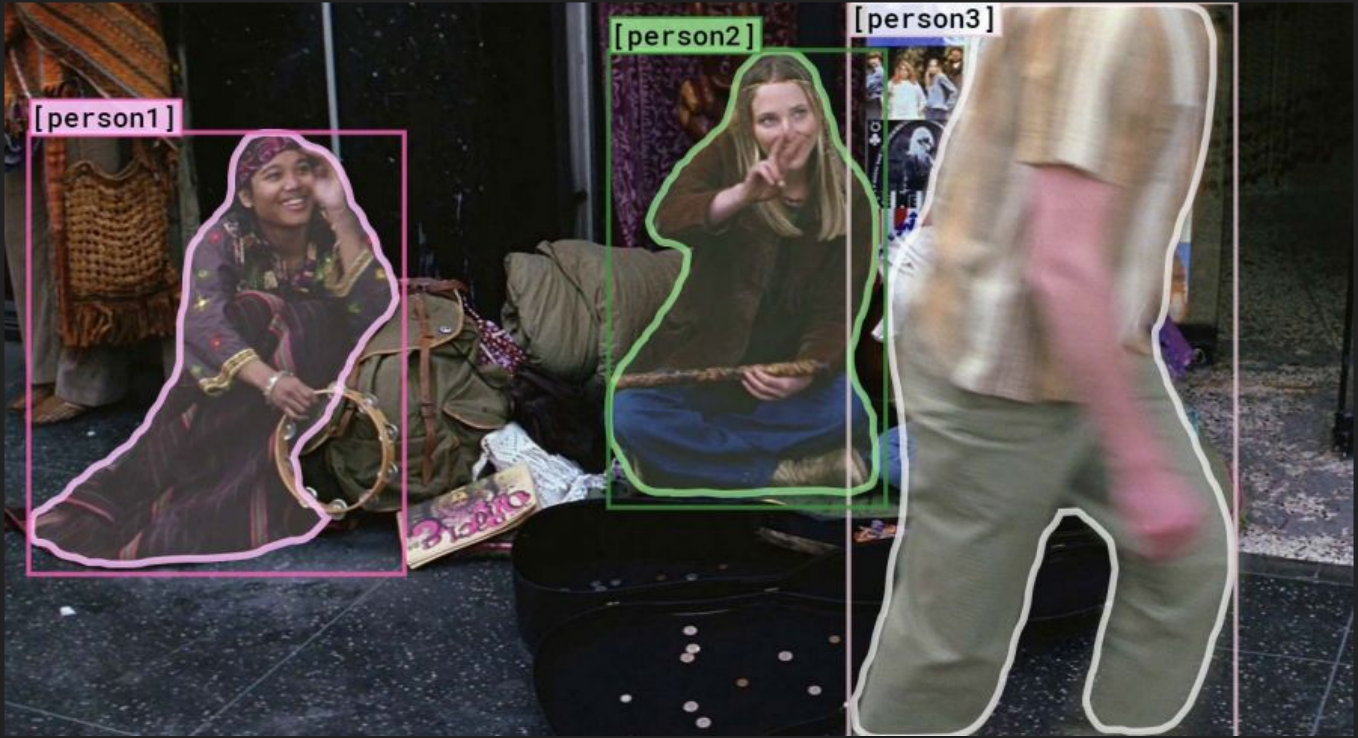
What is happening? Why is it happening?

Why is person 4 pointing to person 1 while looking at person 3?

How to make a machine answer these questions?



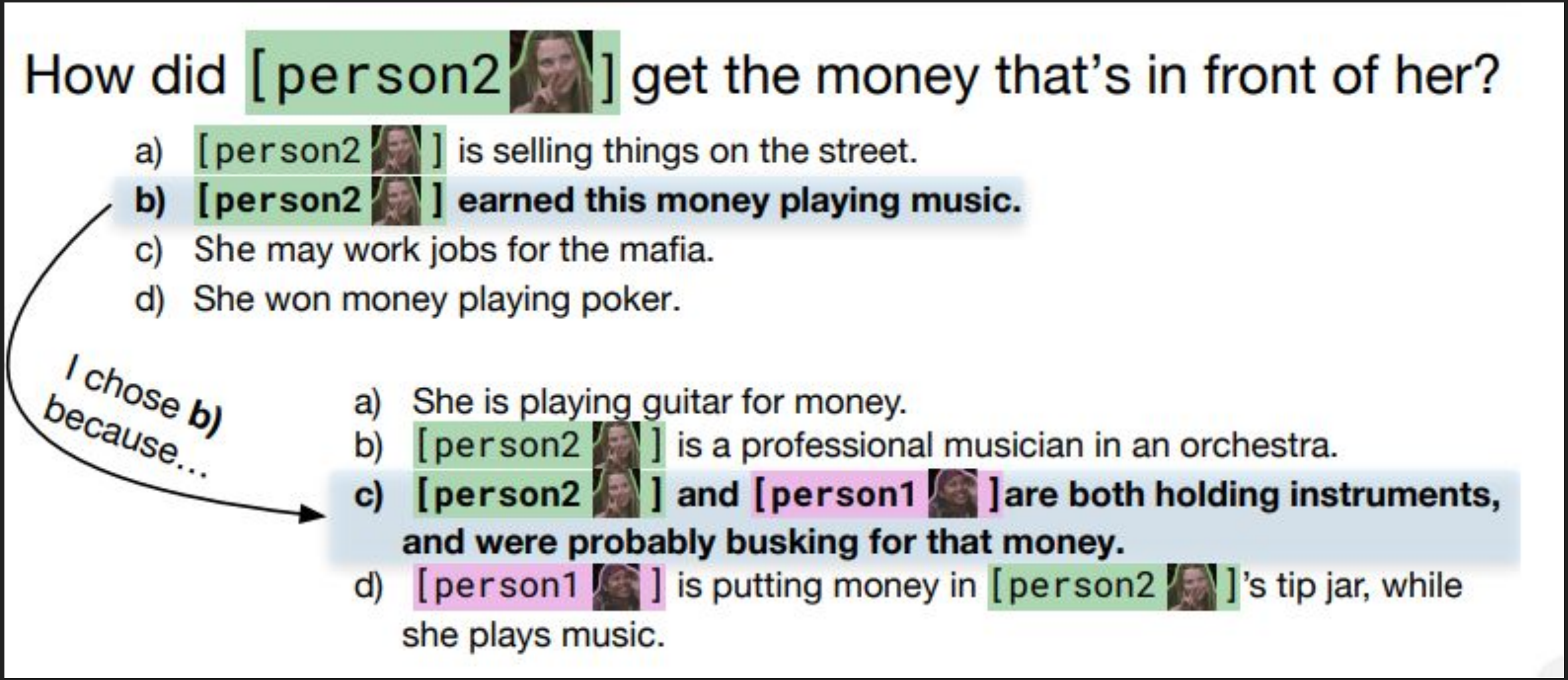
Defining the Problem



What is happening? Why is it happening?

How did person2 get the money that's in front of her?

How to make a machine answer these questions?



Two tasks

Recognition → Cognition

Our questions challenge computer vision systems to go beyond recognition-level understanding, towards a higher-order cognitive and commonsense understanding of the world depicted by the image.

while humans find VCR easy (over 90% accuracy),

state-of-the-art vision models struggle (~45%)

Formal Objective

VCR: Given an image, a list of regions, and a question, a model must answer the question and provide a rationale explaining why its answer is right.

Definition **VCR** *subtask.* A single example of a **VCR** subtask consists of an image $\boldsymbol{I}$, and:

- A sequence $\boldsymbol{o}$ of object detections. Each object detection o_i consists of a *bounding box* $\mathbf{b}$, a segmentation mask $\mathbf{m}$[1], and a class label $\ell_i \in \mathcal{L}$.
- A *query* $\boldsymbol{q}$, posed using a mix of natural language and pointing. Each word $\boldsymbol{q}_i$ in the query is either a word in a vocabulary $\mathcal{V}$, or is a tag referring to an object in $\boldsymbol{o}$.
- A set of N *responses*, where each response $\boldsymbol{r}^{(i)}$ is written in the same manner as the query: with natural language and pointing. Exactly one response is correct.

The model chooses a single (best) response.

Motivation and Importance

Intelligent Image Captioning

Visual Question Answering

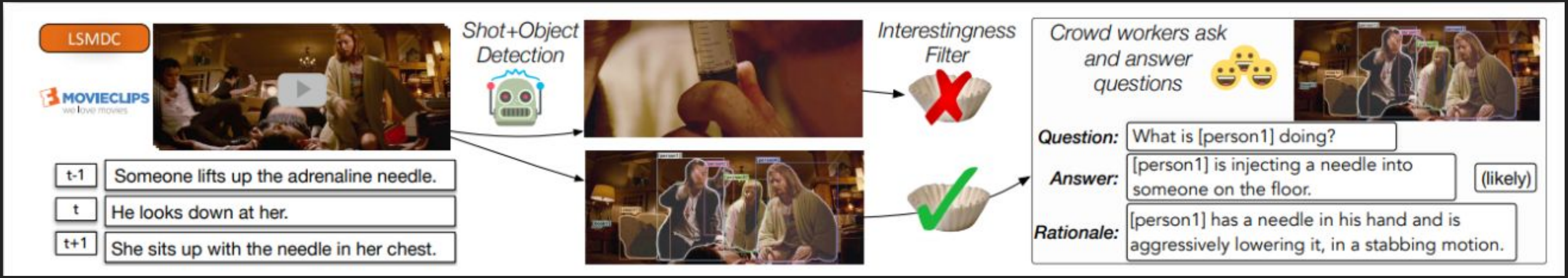
Intelling Vision systems

Dataset Preparation

1. The dataset proposed consists of clips from the Youtube channel MovieClips as well as the Large Scale Movie Description Challenge dataset.

2. Mask-RCNN was then used to detect objects and filter “uninteresting scenes”.

3. They then used Amazon’s Mechanical Turk (crowdsourcing) to write up to three question/answer/rationale triplets per selected frame.



— 290k pairs of questions, answers, and rationales, over 110k unique movie scenes

Adversarial Matching

It was too demanding and error-prone to ask the workers to also write false responses and rationales to the provided questions.

Human-written answers contain unexpected but distinct biases that models can easily exploit.

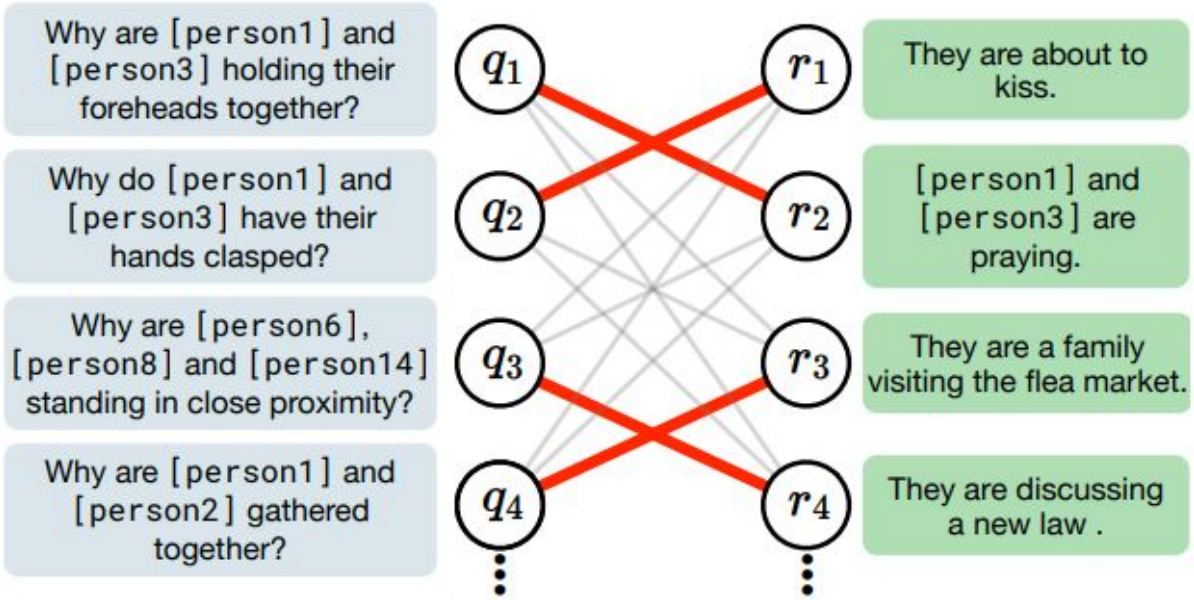


Figure 5: Overview of **Adversarial Matching**. Incorrect choices are obtained via maximum-weight bipartite matching between queries and responses; the weights are scores from state-of-the-art natural language inference models. Assigned responses are highly relevant to the query, while they differ in meaning versus the correct responses.

The Approach - Recognition to Cognition Networks (R2C)

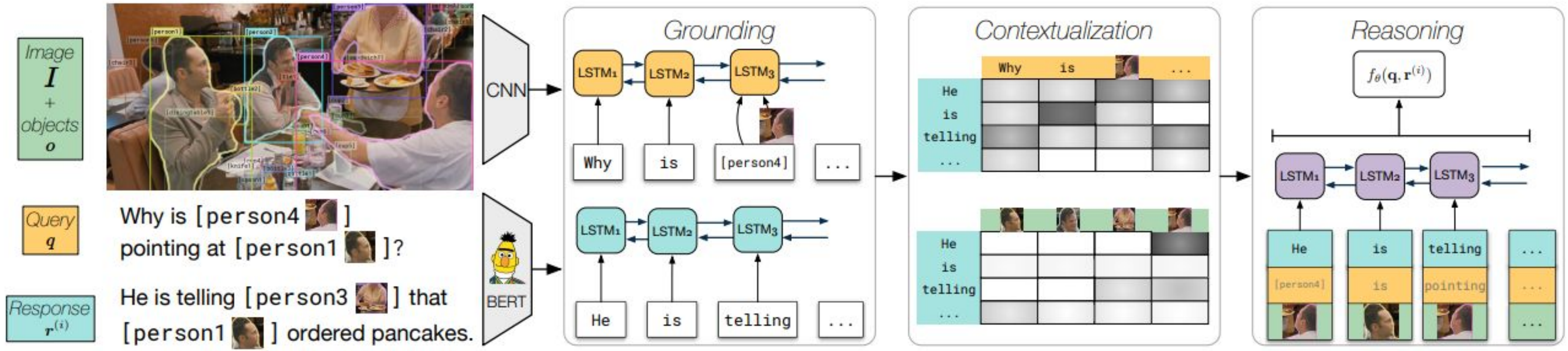


Figure 6: High-level overview of our model, **R2C**. We break the challenge of Visual Commonsense Reasoning into three components: grounding the query and response, contextualizing the response within the context of the query and the entire image, and performing additional reasoning steps on top of this rich representation.

The Approach - Recognition to Cognition Networks (R2C)

The grounding module learns a joint image-language representation for each token of the question and answers.

Contextualization then contextualizes the sentences with each other and the image context.

Reasoning then tries to associate which sentences fit together.

Results

	Model	$Q \rightarrow A$ Val	$Q \rightarrow A$ Test	$QA \rightarrow R$ Val	$QA \rightarrow R$ Test	$Q \rightarrow AR$ Val	$Q \rightarrow AR$ Test
	Chance	25.0	25.0	25.0	25.0	6.2	6.2
Text Only	BERT	53.8	53.9	64.1	64.5	34.8	35.0
Text Only	BERT (response only)	27.6	27.7	26.3	26.2	7.6	7.3
Text Only	ESIM+ELMo	45.8	45.9	55.0	55.1	25.3	25.6
Text Only	LSTM+ELMo	28.1	28.3	28.7	28.5	8.3	8.4
VQA	RevisitedVQA [39]	39.4	40.5	34.0	33.7	13.5	13.8
VQA	BottomUpTopDown[4]	42.8	44.1	25.1	25.1	10.7	11.0
VQA	MLB [43]	45.5	46.2	36.1	36.8	17.0	17.2
VQA	MUTAN [6]	44.4	45.5	32.0	32.2	14.6	14.6
	R2C	**63.8**	**65.1**	**67.2**	**67.3**	**43.1**	**44.0**
	Human		91.0		93.0		85.0

Table 1: Experimental results on VCR. VQA models struggle on both question-answering ($Q \rightarrow A$) as well as answer justification ($Q \rightarrow AR$), possibly due to the complex language and diversity of examples in the dataset. While language-only models perform well, our model **R2C** obtains a significant performance boost. Still, all models underperform human accuracy at this task.

Active Leaderboard on: https://visualcommonsense.com/leaderboard/

Conclusion

The authors presented R2C, a model for this task, but the challenge – of cognition-level visual understanding – is far from solved.

References

1. https://openaccess.thecvf.com/content_CVPR_2019/papers/Zellers_From_Recognition_to_Cognition_Visual_Commonsense_Reasoning_CVPR_2019_paper.pdf

2. https://www.youtube.com/watch?v=Je5LIZlqUt8&t=4776s

3. https://vitalab.github.io/article/2018/11/29/VCR.html

4. https://www.youtube.com/user/movieclips

5. https://sites.google.com/site/describingmovies/lsmdc-2017

6. https://www.mturk.com/