

Evaluating Agents

CS6960 MultiModal LLM Agents

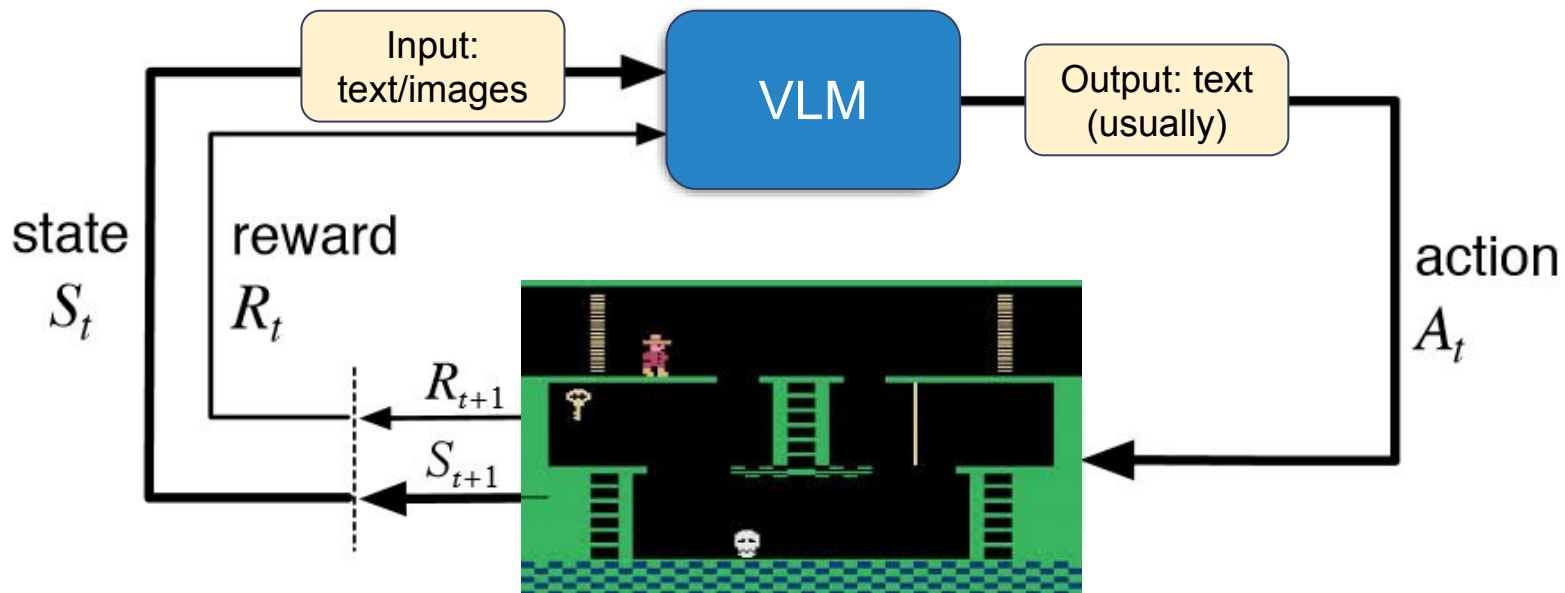
Kenneth Marino

Announcements

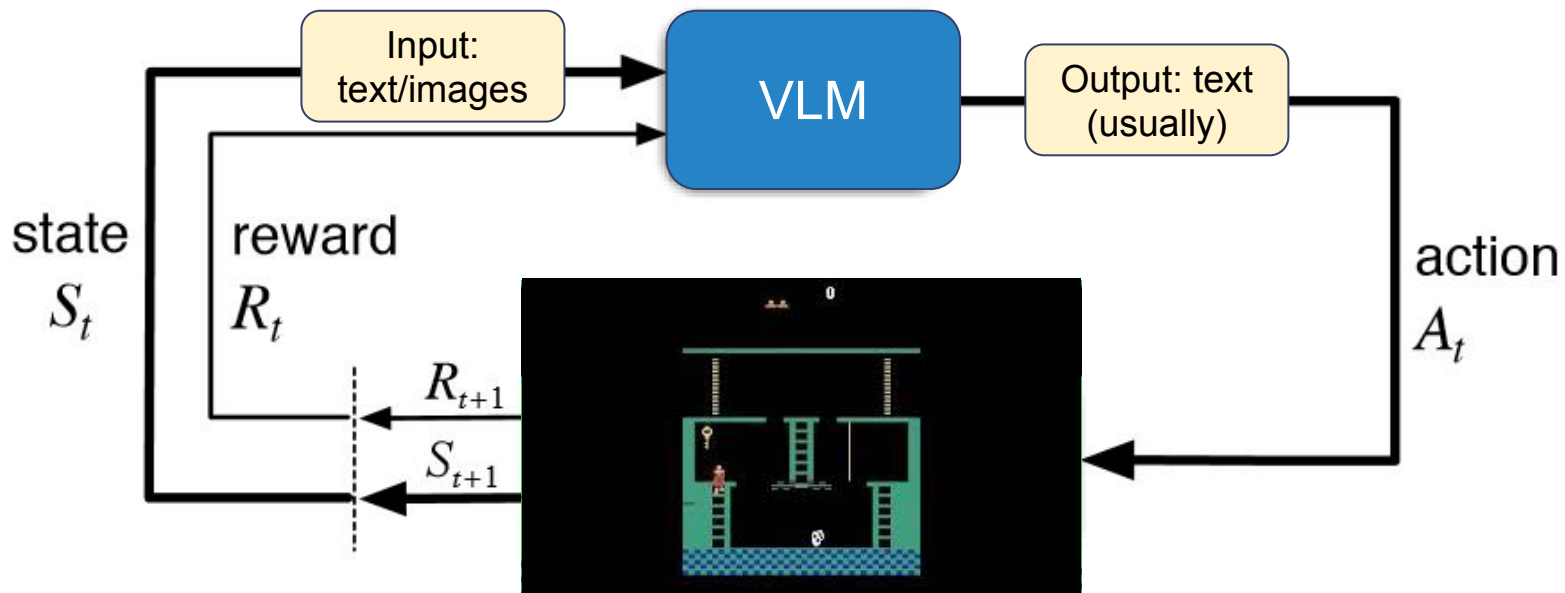
- HW2 due today
 - You have a few more late days to complete
- Projects
 - Due Friday
 - If you don't have an assigned group on Canvas to submit, let us know
 - All info now in the Canvas Assignment
 - In-class activity Thursday to help you double-check before you submit

Any Questions

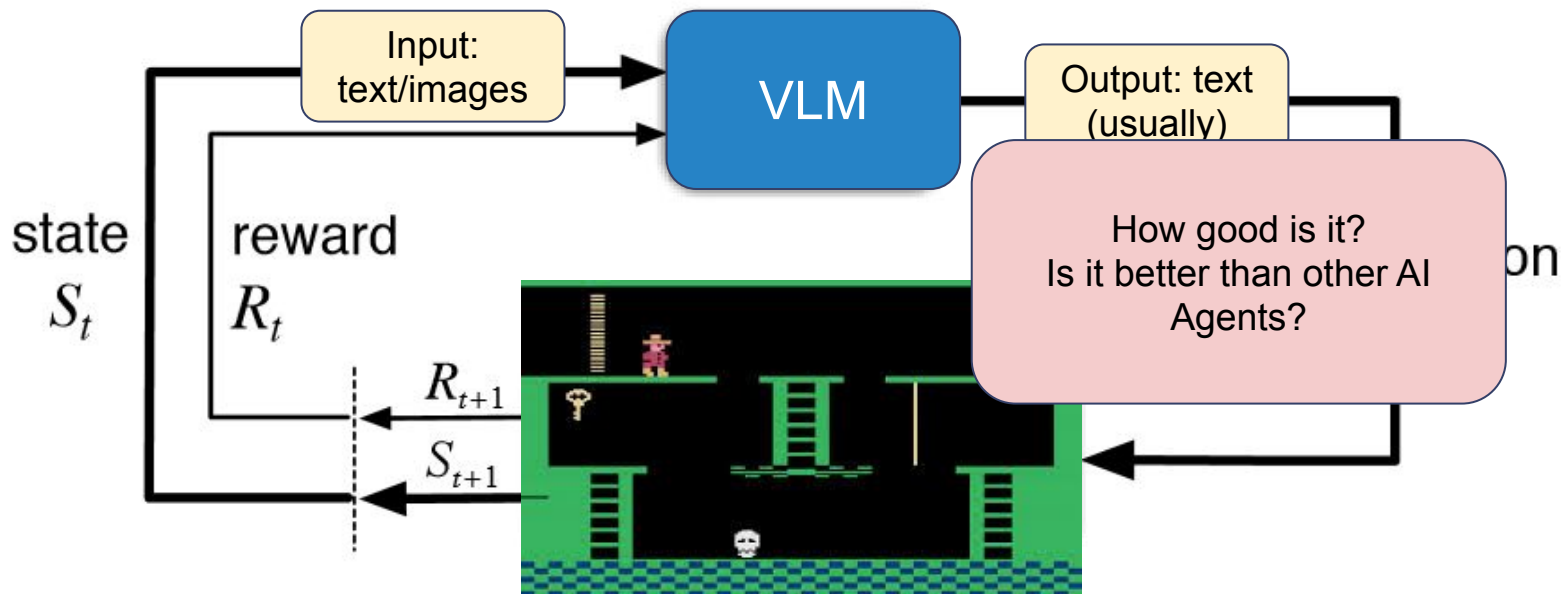
We have a VLM Agent



We have a VLM Agent



We have a VLM Agent



COMS 6998 Datasets in Machine Learning

Kenneth Marino

My first special topics class

Datasets/Evaluations often overlooked

- Datasets are often overshadowed by models and methods
- Often the most influential driver of progress in Machine Learning, Computer Vision and NLP.
- Overview the important topics in datasets
 - Dataset analysis
 - Collection
 - Evaluation
 - Dataset bias
 - Ethics
- Cover the most influential datasets



What does an evaluation do?

- Let's us compare different models
- Measure/drive progress in a field
- Understand behaviors/limitations

Good case study: ImageNet

IM GENET

Where have we been? Where are we going?

Li Fei-Fei & Jia Deng



https://image-net.org/static_files/files/imagenet_ilsrv2017_v1.0.pdf

While Others Targeted Detail...



LabelMe

Per-Object Regions and Labels

Russell et al, 2005



Lotus Hill

Hand-Traced Parse Trees

Yao et al, 2007

...They Targeted Scale

SUN, 131K

[Xiao et al. '10]

LabelMe, 37K

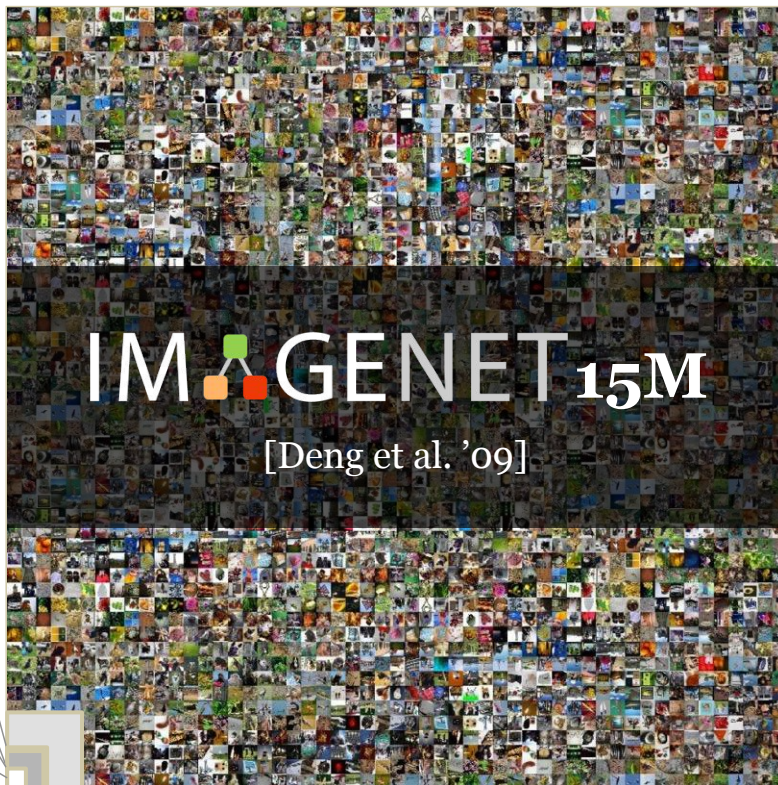
[Russell et al. '07]

PASCAL VOC, 30K

[Everingham et al. '06-'12]

Caltech101, 9K

[Fei-Fei, Fergus, Perona, '03]

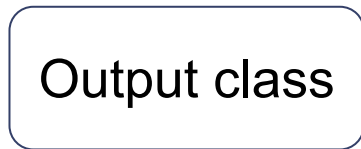


An Emphasis on Community and Achievement

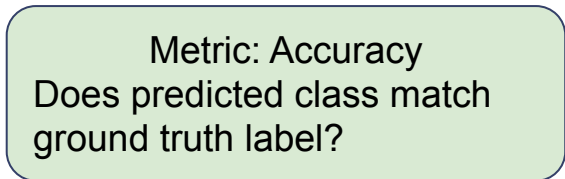
IM  GENET

**Large Scale Visual Recognition
Challenge (ILSVRC 2010-2017)**

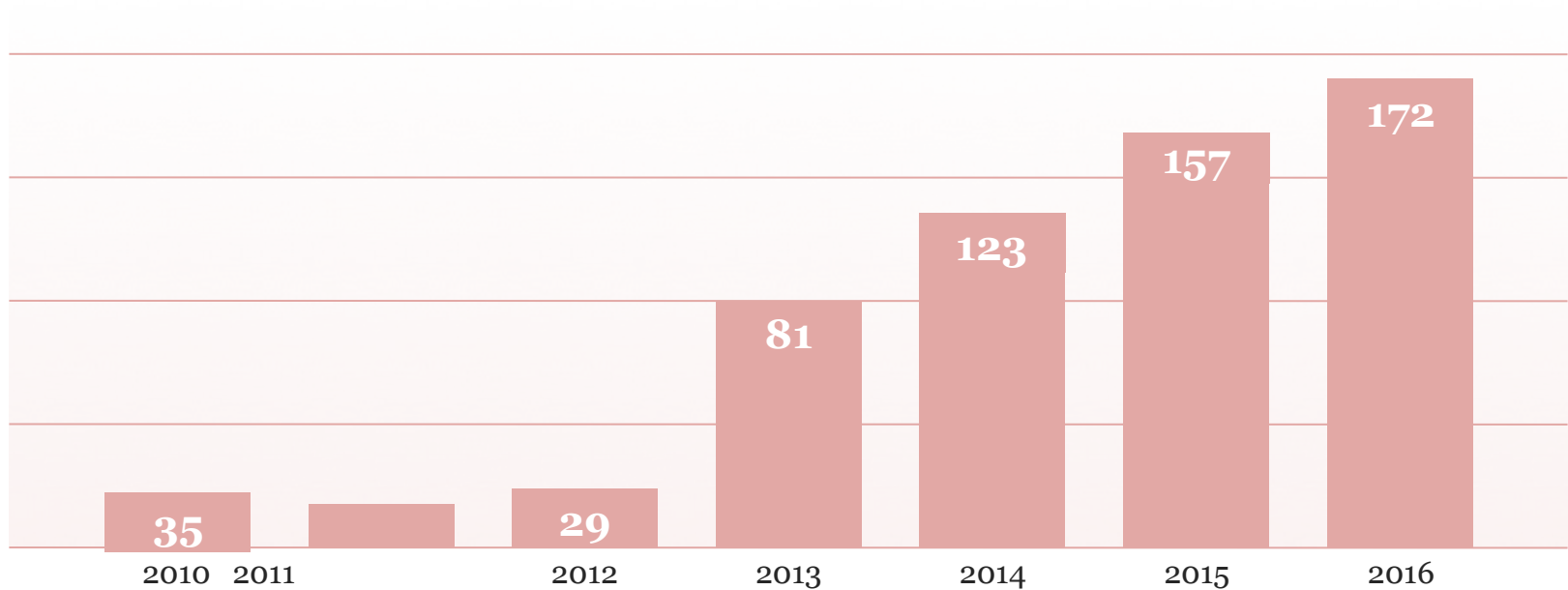
ImageNet Evaluation



"Car"

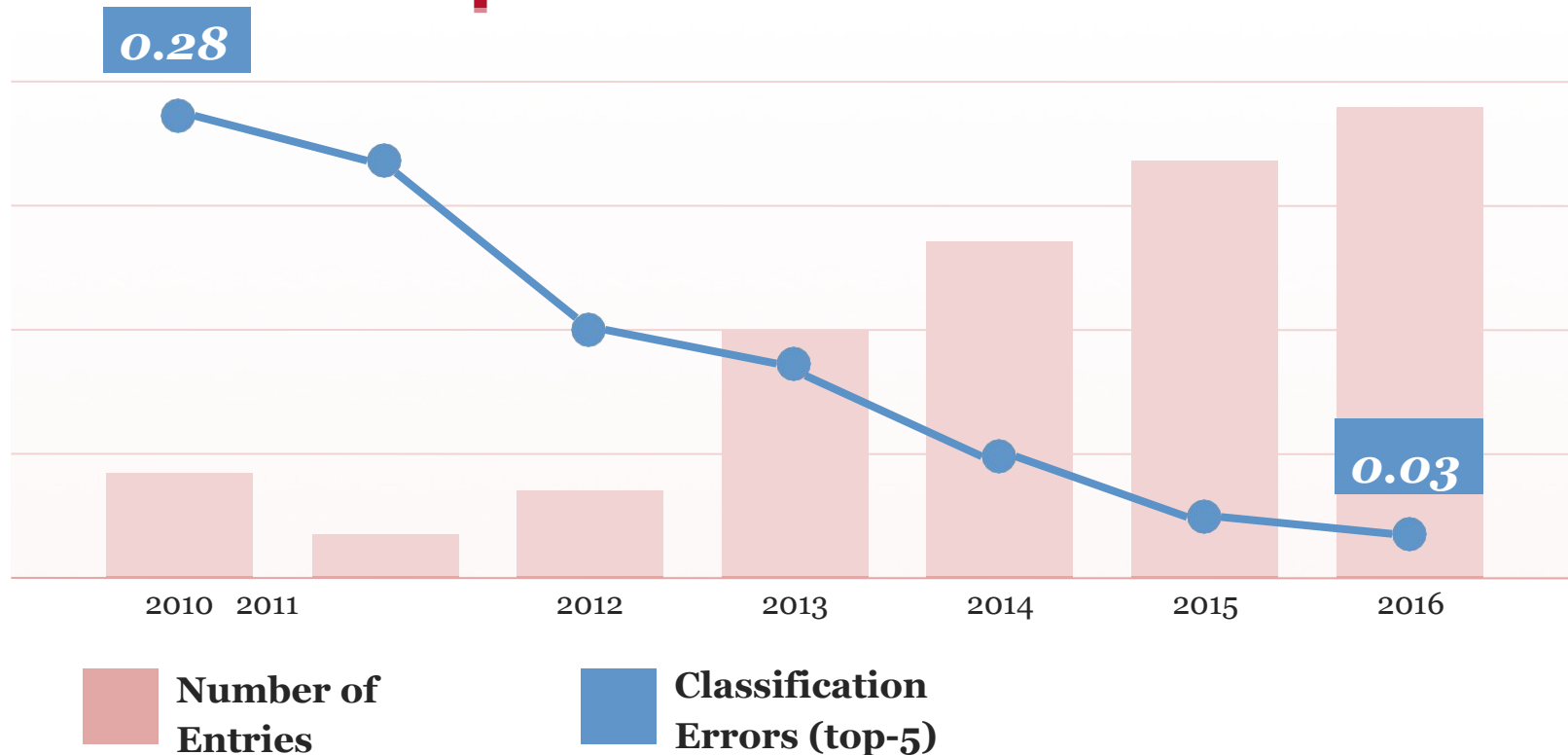


Participation and Performance

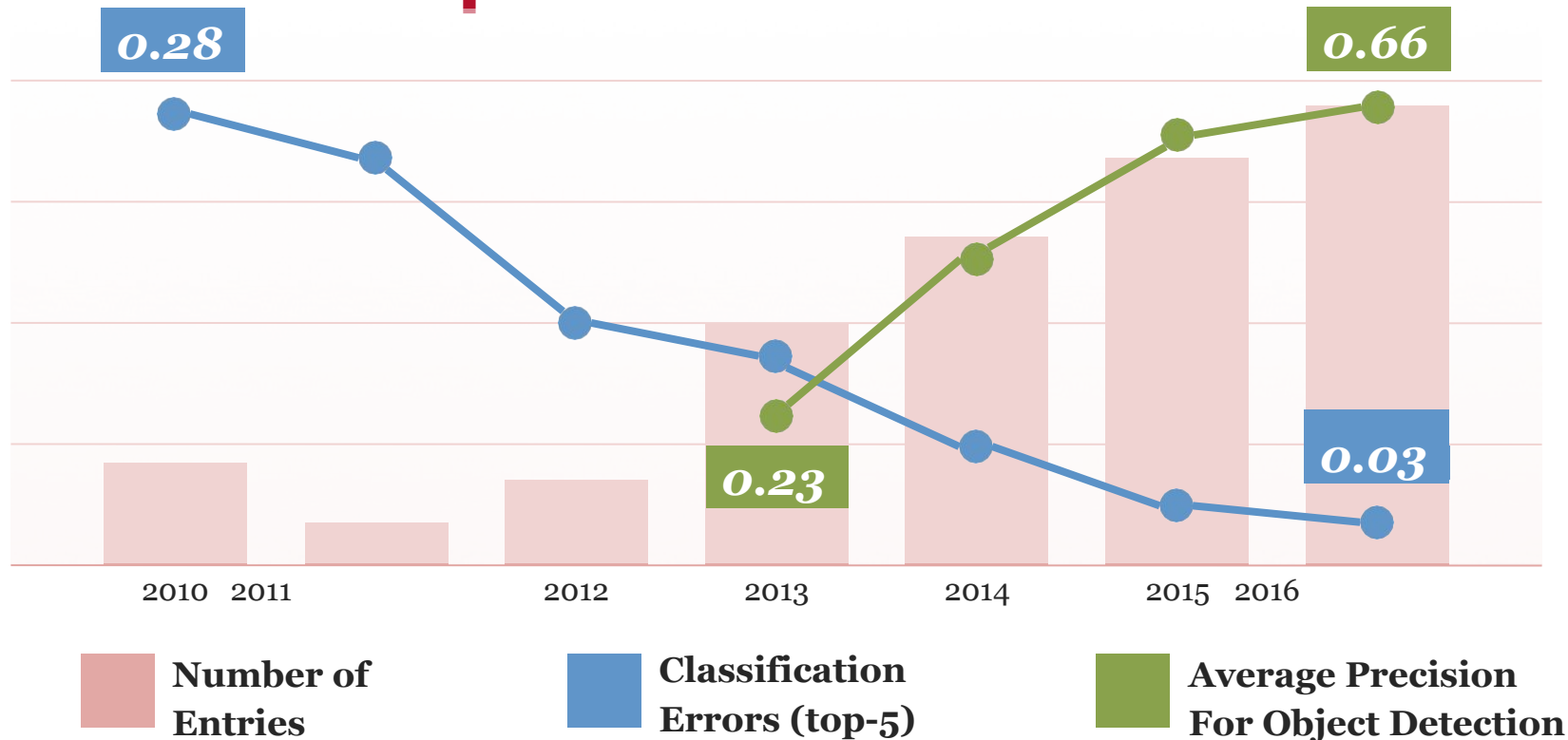


 **Number of
Entries**

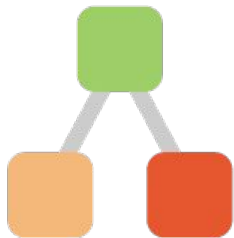
Participation and Performance



Participation and Performance



Outcomes



ImageNet becomes a benchmark



Breakthroughs in object recognition



Machine learning advances and changes dramatically

What if measure of success less clear?

What if measure of success less clear?



Describe everything that is in this image

Case Study: Image Captioning



COCO

Common Objects in Context

160,000 images
5 captions per image



A man checking out a parked black scooter.
A person standing near a small motorcycle on a city street.
A man in a white shirt is looking at a three wheeled motorcycle.
A man looks down at two low riding motor bikes.
A guy staring at a weird looking bike.

Evaluation:

Content overlap metrics

Ref: They walked **to the grocery store** .

Gen: **The woman went to the hardware store** .



- Compute a score that indicates the lexical similarity between *generated* and *gold-standard (human-written) text*
- Fast and efficient and widely used
- *N*-gram overlap metrics (e.g., **BLEU**, ROUGE, METEOR, CIDEr, etc.)

Timeline

August,
2014



120,000 images x 5 captions per image =
600,000 captions

The Great Freak Out

August,
2014

October

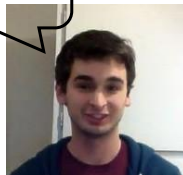


I can't believe it works!!!

Sweet!!!



Yeesss!



This is awesome!!!



It works!!!



The Reckoning

August, 2014

October

April, 2015



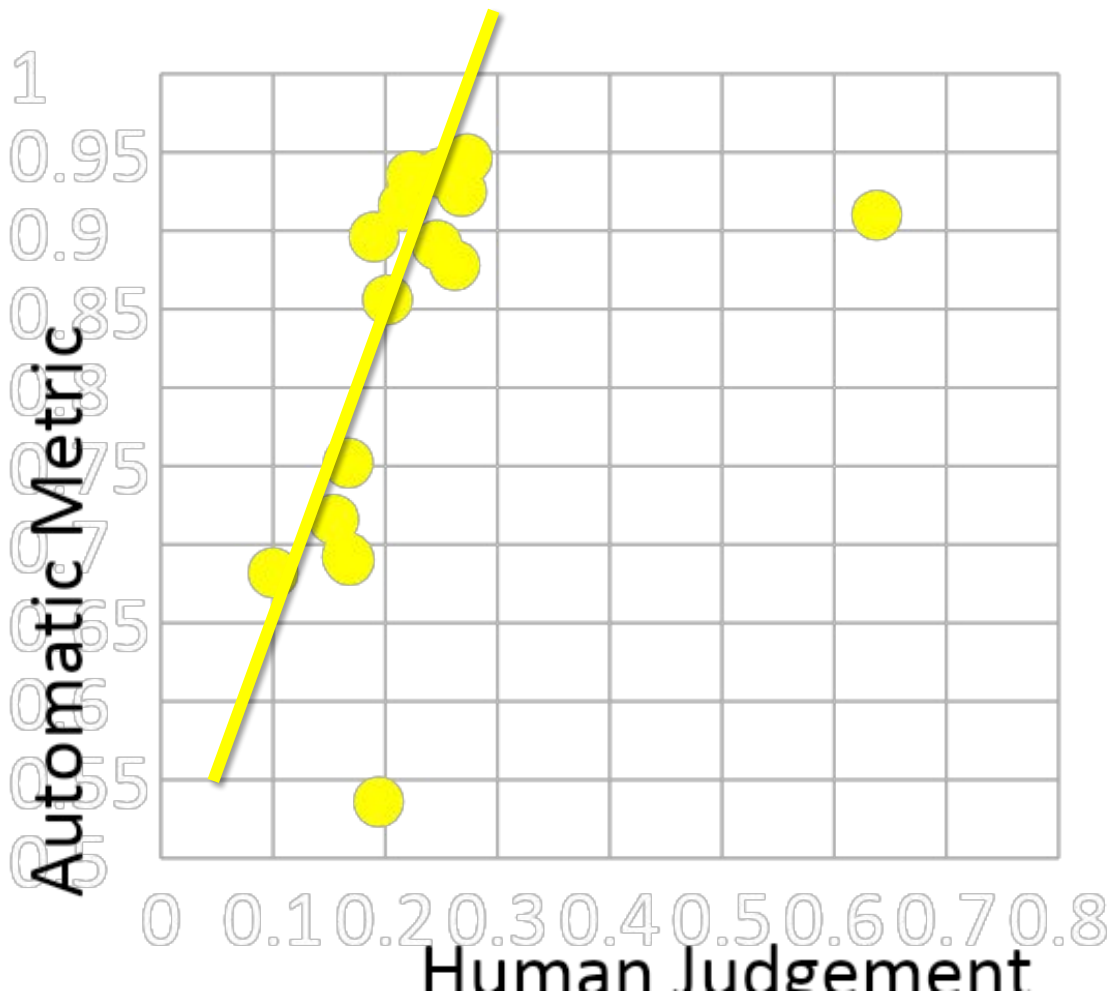
	CIDEr-D	↓ ↑ F	Meteor	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4
MSR ^[5]	0.912		0.247	0.519	0.695	0.526	0.391	0.291
Berkeley LRCN ^[1]	0.869		0.242	0.517	0.702	0.528	0.384	0.277
Human ^[3]	0.854		0.252	0.484	0.663	0.469	0.321	0.217
Google ^[2]	0.834		0.236	0.498	0.673	0.493	0.362	0.272
m-RNN (Baidu/ UCLA) ^[8]	0.819		0.229	0.504	0.685	0.512	0.376	0.279
MLBL ^[4]	0.74		0.219	0.499	0.666	0.498	0.362	0.26
NeuralTalk ^[6]	0.674		0.21	0.475	0.65	0.464	0.321	0.224
Tsinghua Bigeye ^[7]	0.673		0.207	0.49	0.671	0.494	0.35	0.241

The Enlightening

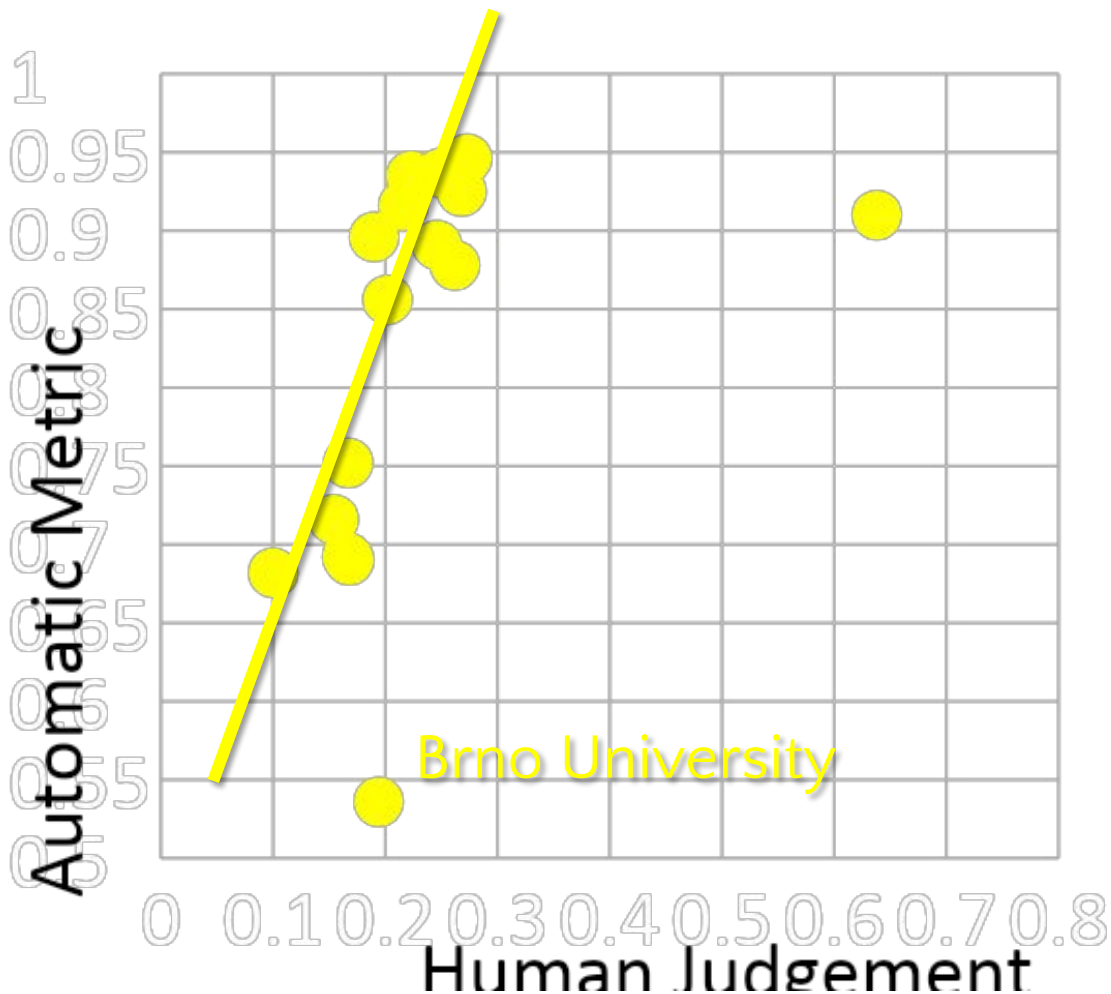


How do humans rate the captions?

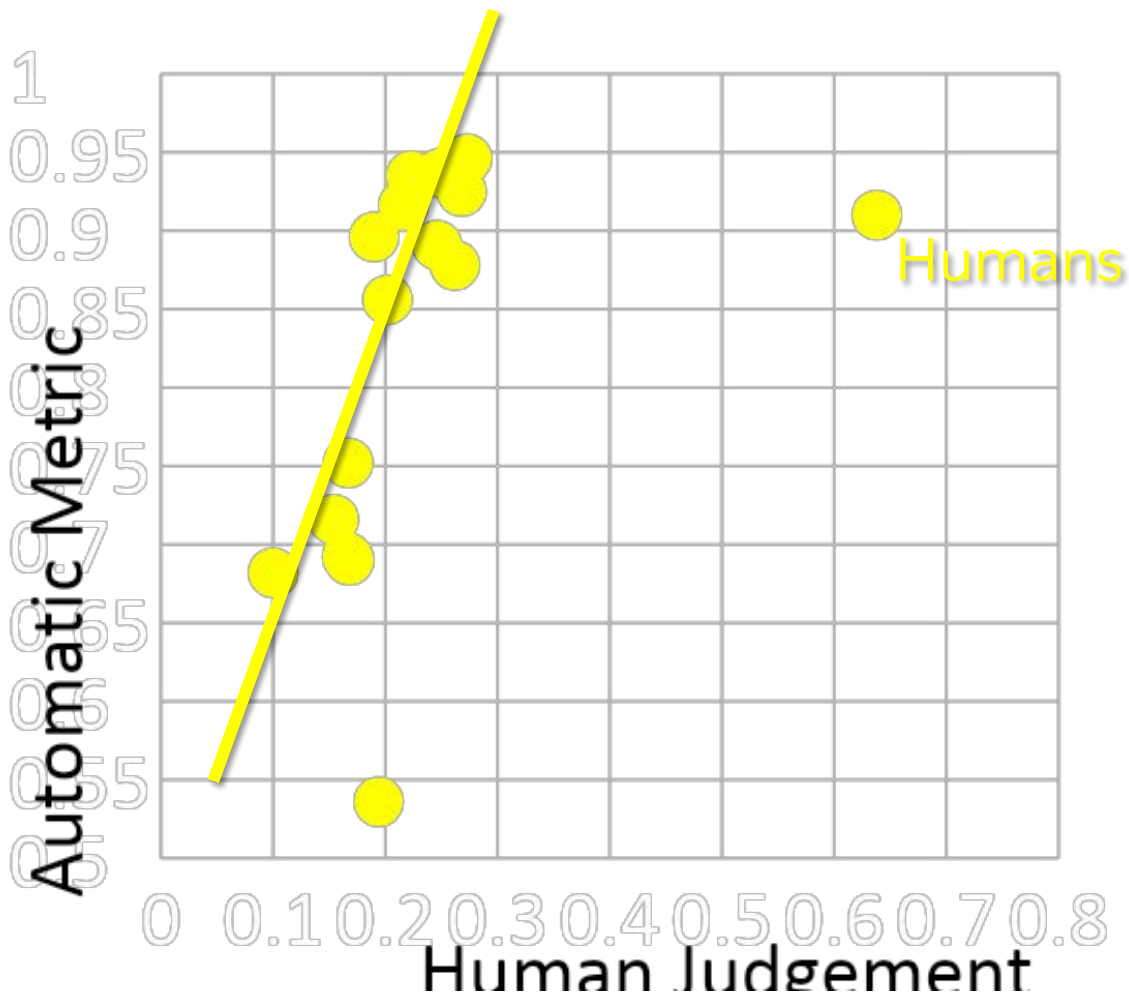
Evaluation



Evaluation



Evaluation



Goodhart's Law

GOODHART'S LAW

WHEN A MEASURE BECOMES A TARGET,
IT CEASES TO BE A GOOD MEASURE

IF YOU
MEASURE
PEOPLE ON...

NUMBER OF
NAILS MADE

WEIGHT OF
NAILS MADE

THEN YOU
MIGHT GET

1000'S OF
TINY NAILS

A FEW GIANT,
HEAVY NAILS



Goodhart's Law



What does an evaluation do?

- Let's us compare different models
- Measure/drive progress in a field
- Understand behaviors/limitations

What we want in an agent evaluation

- It accurately tracks what humans care about
- Measures broad abilities of agent models capabilities
- Automatic - Can be done without additional human effort
- Reproducible - Consistently produces same outcome
- Ranks models - better models do better than worse ones

What we want in an agent evaluation

- It accurately tracks what humans care about
- Measures broad abilities of agent models capabilities
- Automatic - Can be done without additional human effort
- Reproducible - Consistently produce
- Ranks models - better models do better

Doing all of these
is kind of
impossible

Option 1

- It accurately tracks what humans care about
- Measures broad abilities of agent models capabilities
- ~~● Automatic - Can be done without additional human effort~~
- ~~● Reproducible - Consistently produces same outcome~~
- Ranks models - better models do better than worse ones

Just get humans
to evaluate it

Clever ways of getting humans to eval

Agent Arena

Welcome to the LLM Agent Arena. Here, you can pit two agents against each other based on a goal you provide. You can also head to your profile to save prompts for agents and visit the Prompt Hub to see prompts used by other users along with their ratings. Ensure your API keys are configured in your profile for optimal performance.

Upload Search

Search Example Stock Example Financial Data Example Research Example Automation Example Surprise Me

Run Both Agents

Agent 1

Agent 2

A is better B is better Tie Both are bad

Share Result

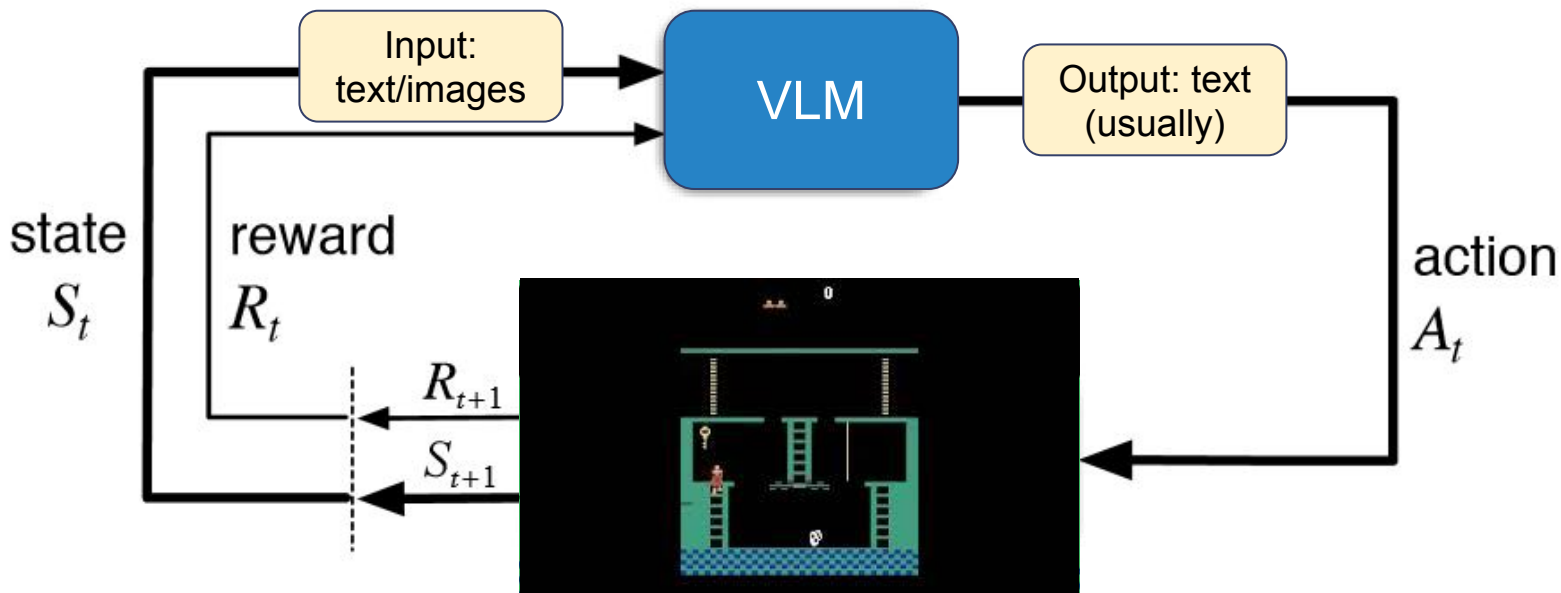
Option 2 - Proxy/Automatic

- It accurately tracks what humans care about
- ~~Measures broad abilities of agent models capabilities~~
- Automatic - Can be done without additional human effort
- Reproducible - Consistently produces same outcome
- Ranks models - better models do better than worse ones

Extremely task specific
May only be doable in some envs
Very difficult to get right

Atari: Use the final reward

Does atari score capture everything we care about?



Option 3 - LLM/Agent as judge

- It accurately tracks what humans care about
- Measures broad abilities of agent models capabilities
- Automatic - Can be done without additional human effort
- ~~● Reproducible - Consistently produces same outcome~~
- Ranks models - better models do better than worse ones

LLM as Judge

See student presentation this week

Question: If the FED buys bonds in the secondary market (A) the money supply will increase. (B) the money supply will decrease. (C) the money supply will not be affected. (D) the discount rate would be affected. Output the answer only.

Assistant A:

The Federal Reserve buys bonds in the secondary market to increase the money supply.

Assistant B:

(A) the money supply will increase.

Follow-up Question: How does it affect my daily life? Give 3 examples.

Assistant A:

The Federal Reserve buys bonds in the secondary market to increase the money supply. This affects your daily life by:

1. Increasing the money supply.
2. Increasing the money supply and lowering the discount rate.
3. Increasing the money supply and lowering the discount rate and increasing the money supply.

Assistant B:

When the Federal Reserve buys bonds in the secondary market, it can have several effects on the economy that can impact your daily life. Here are three examples:

1. Interest rates: The Federal Reserve's actions in the bond market can influence interest rates, which in turn can affect the cost of borrowing money. For example, if the Federal Reserve buys bonds and drives up demand for them, interest rates on things like mortgages, car loans, and credit cards may go up, making it more expensive to borrow money.
2. Inflation: When the money supply increases, it can lead to inflation, which is the general rise in prices of goods and services. This can impact your daily life by making it more expensive to buy things like groceries, gas, and other necessities.
3. Employment: The Federal Reserve's actions in the bond market can also have an impact on employment. For example, if the Federal Reserve's actions lead to economic growth and increased demand for goods and services, it can create more job opportunities and boost employment.

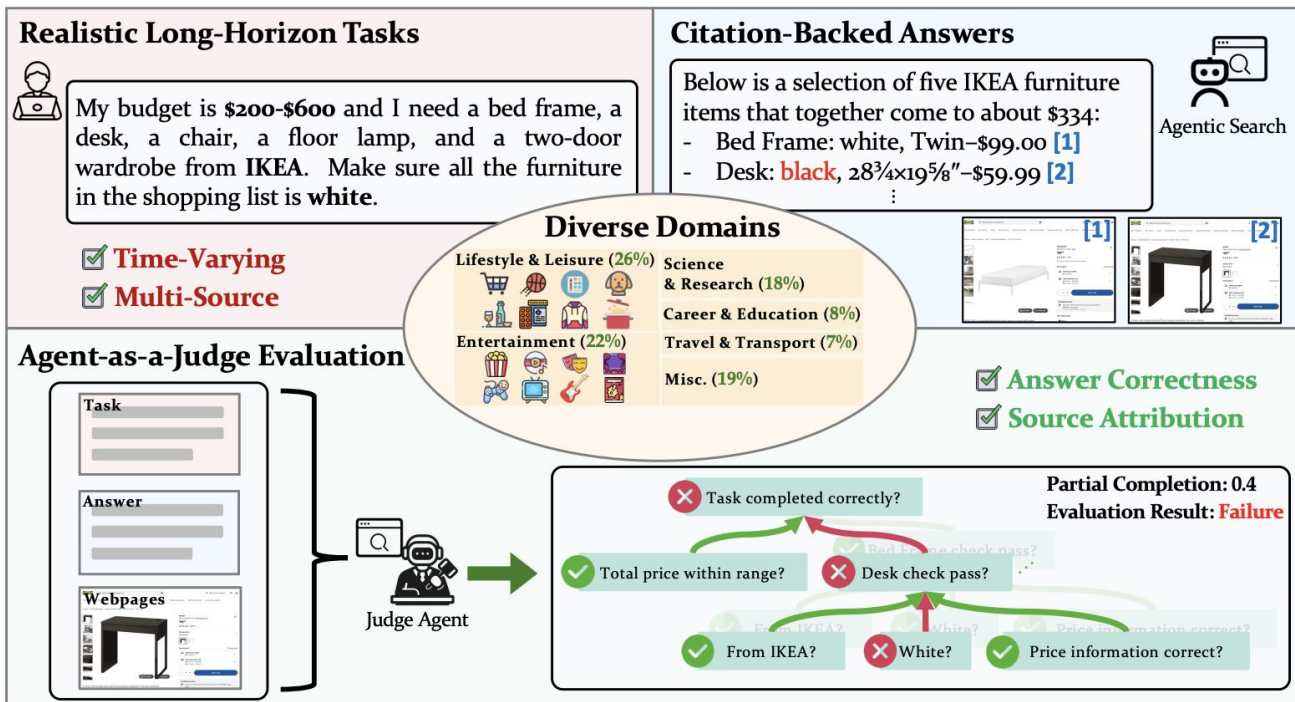
GPT-4 Judgment:

Assistant A provided an incorrect response to the user's question about how the Federal Reserve buying bonds in the secondary market affects daily life. The answer given is repetitive and lacks clear examples of how the action impacts daily life.

On the other hand, **Assistant B** provided a relevant and accurate response to the user's question about the Federal Reserve buying bonds. The answer includes three clear examples of how the action impacts daily life, such as interest rates, inflation, and employment.

Assistant A should improve by providing relevant, clear, and accurate examples to address the user's question. **Assistant B's** response is more helpful, relevant, and detailed.

Extension: Agent as Judge



Other evaluation problems

- Data leakage - are your tasks already in the training set?
- Agent tasks are complicated difficult
- Agent tasks are often ambiguous
- Agents can often find “shortcuts” in evaluations
- Is accuracy the right measure (cost? partial completion?)

As you think about your projects

- Need to think about evaluation
- What is important to measure for your system?
- What evaluation can you do in a short time
- You probably have to compromise somewhere

Any Questions



Questions

Now for the presentations!