

Semantically Equivalent Adversarial Rules for Debugging NLP Models

Sameer Singh (UC Irvine)



Carlos Guestrin



NLP / ML models are getting smarter: VQA

What type of road sign is shown?

> STOP.

Visual7A [Zhu et al 2016]



NLP / ML models are getting smarter: MC (SQuAD)

The biggest city of [redacted] is Cologne, Germany with a population of more than 1,050,000 people. It is [redacted] in Central and Western Europe (after the [redacted] at about 1,230 km (760 mi))

How long is the Rhine?

>1230km

BiDAF [Seo et al 2017]



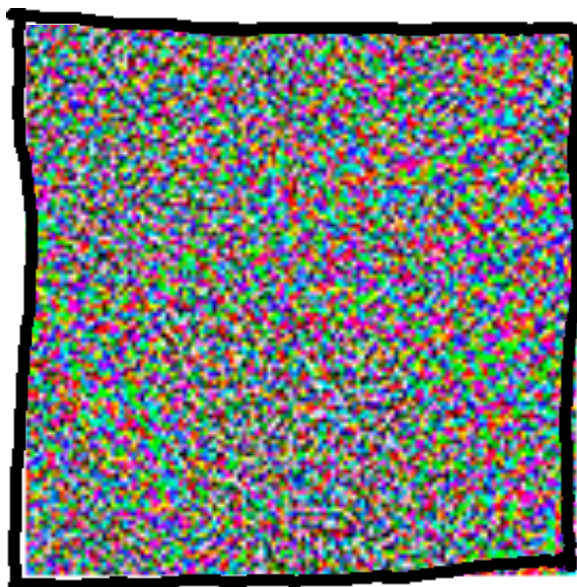
Oversensitivity in images



“panda”

57.7% confidence

+ ϵ



=



“gibbon”

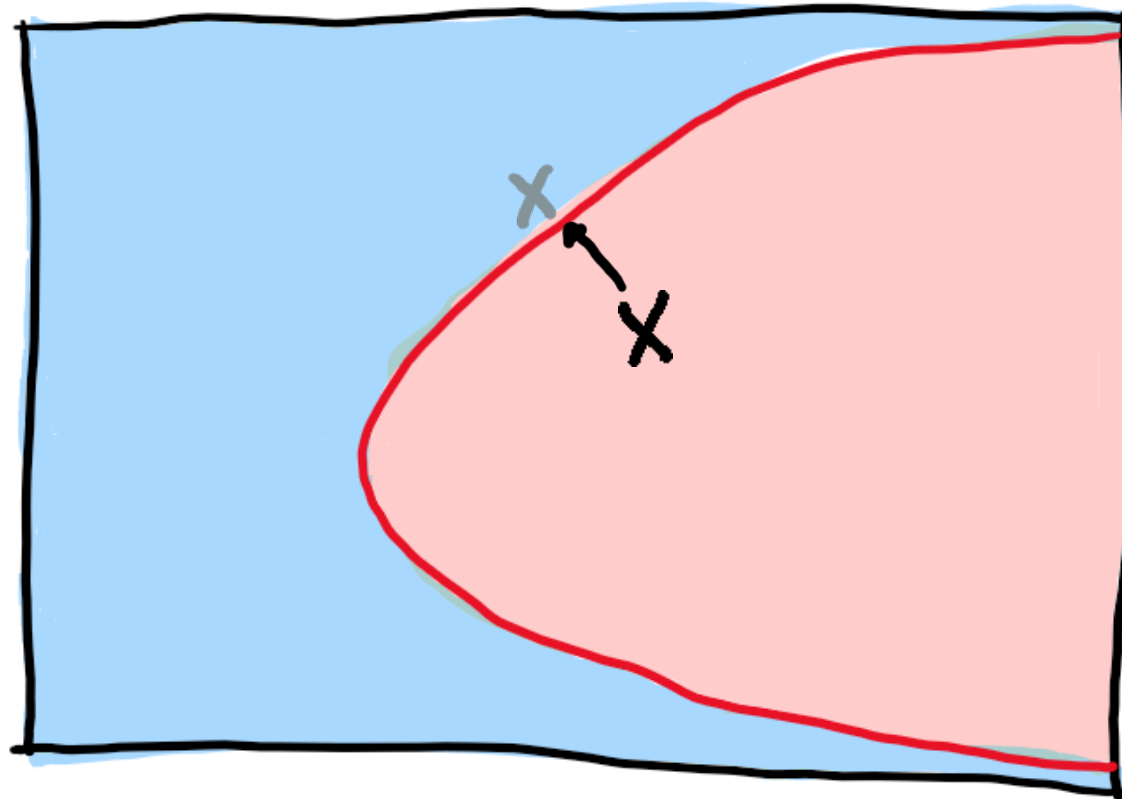
99.3% confidence

But



(links)

Adversarial examples



Fi

on

What about text?



What type of road sign is shown?

> STOP.

What ~~z~~ type of ~~one~~ ^{door} sign is shown?



What about text?



What type of road sign is shown?

> STOP.

What type of road sign ~~is~~ shown?



Semantics matter

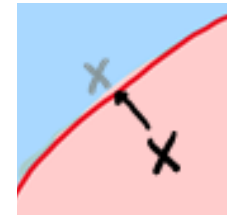


What type of road sign is shown?

> STOP.

Which
~~What~~ type of road sign is shown?

> Do not Enter.



Semantics matter

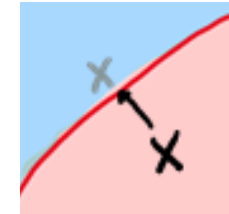
The biggest city of [redacted] is Cologne, Germany with a population of more than 1,050,000 people. It is [redacted] river in Central and Western Europe (after the Danube [redacted] at about 1,230 km (760 mi))

How long is the Rhine?

> 1230km

How long is the Rhine?

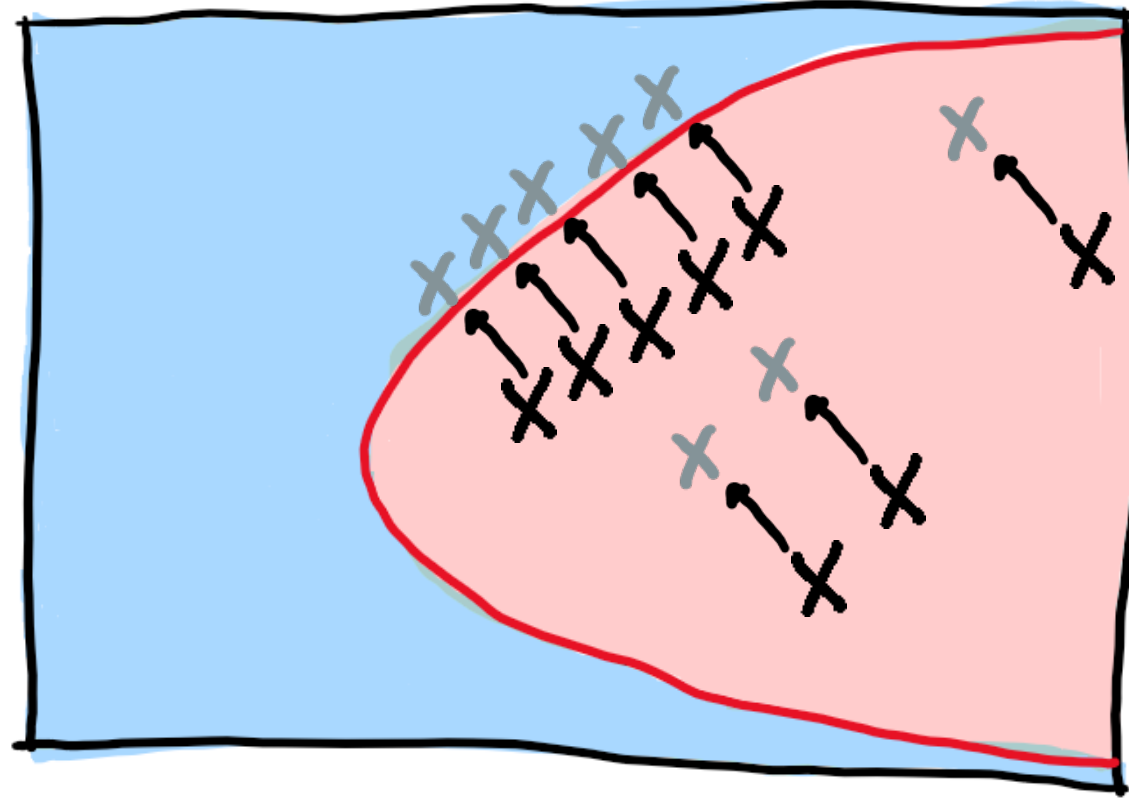
> More than 1,050,000



No

er

Adversarial Rules



Generalizing adversaries

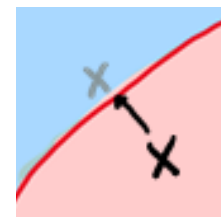


Which **NOUN** of road sign is shown?

> STOP.

Which **NOUN** of road sign is shown?

> Do not Enter.



Semantics matter

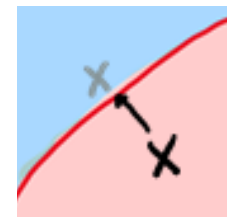


^{NOUN}
[redacted] is the sky?

> Blue.

^{Which NOUN}
[redacted] is the sky?

> Gray.



Semantics matter

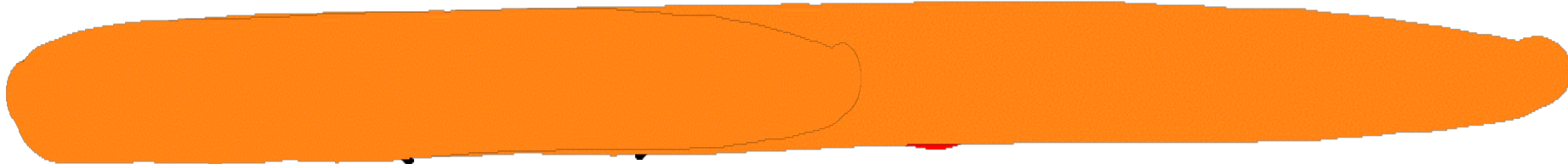
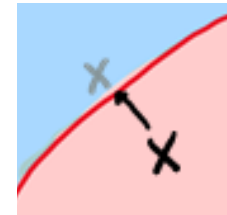
The biggest city of **Rhine** is Cologne, Germany with a population of more than 1,050,000 people. It is the **longest** river in Central and Western Europe (after the Danube) at about 1,230 km (760 mi)

How long is the Rhine?



> 1230km

How long is the Rhine?

> More than 1,050,000



Semantics matter

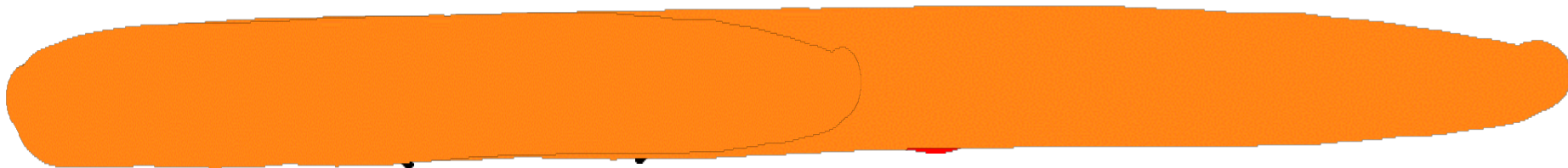
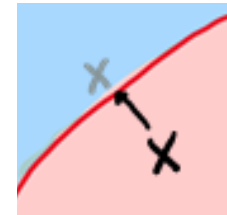
Detailed investigation of   chus keta, showed that these fish digest ctenophores 20 times as fast as an equal weight of shrimps.

What is the oncorhynchus also called?

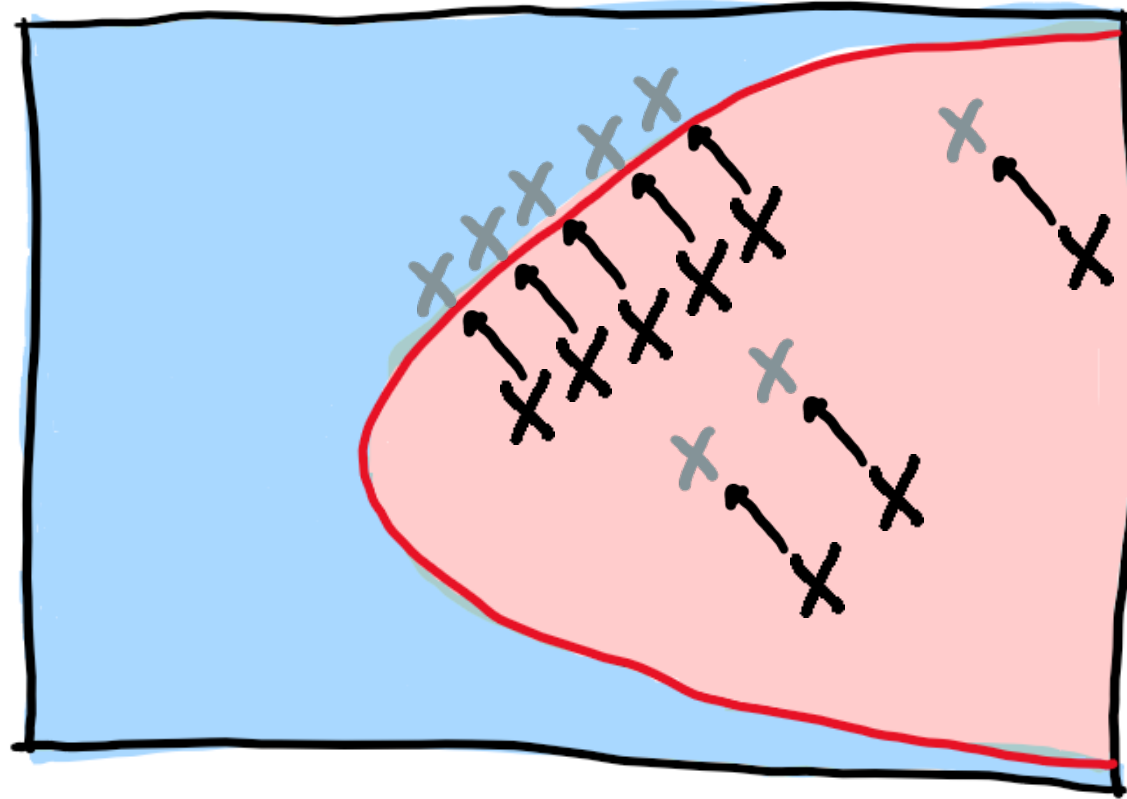
> chum salmon

What is the oncorhynchus also called?

> Oncorhynchus keta

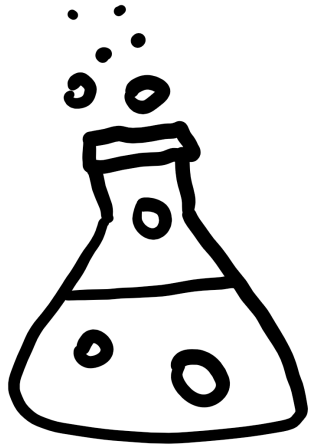


Adversarial Rules



Semantically Equivalent Adversary (SEA)

Ingredients



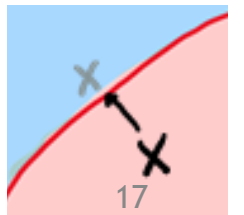
① Semantic score function $S(x, x')$

② A black box model $f(x)$ 

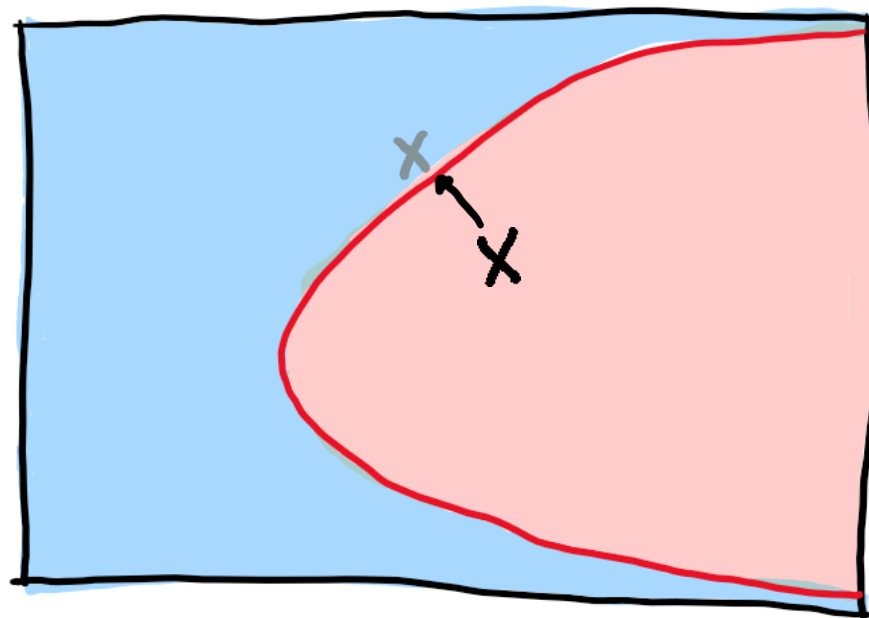
AND

Semantically
Equivalent

Different
prediction



Revisiting adversaries



Fi

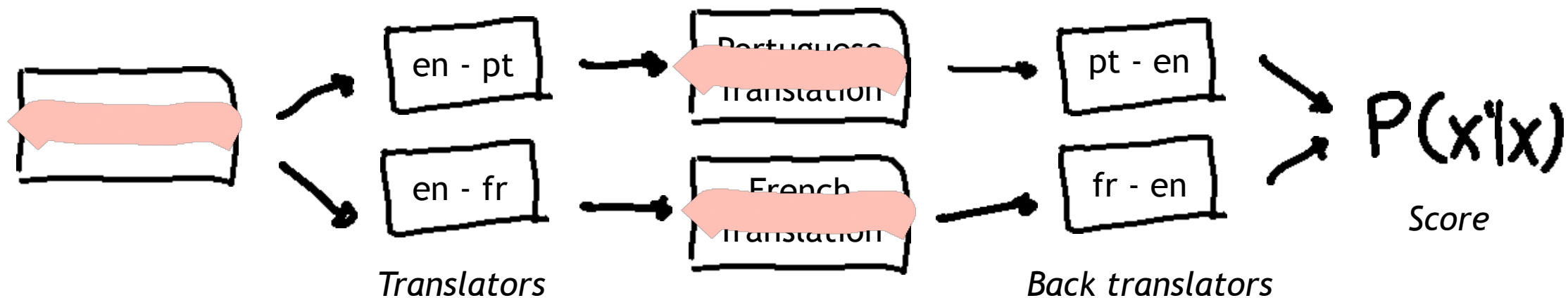
SEA

on

$$\max_{x, x'} S(x, x') > \gamma \text{ s.t. } SEA(x, x') =$$

Semantic Similarity: Paraphrasing

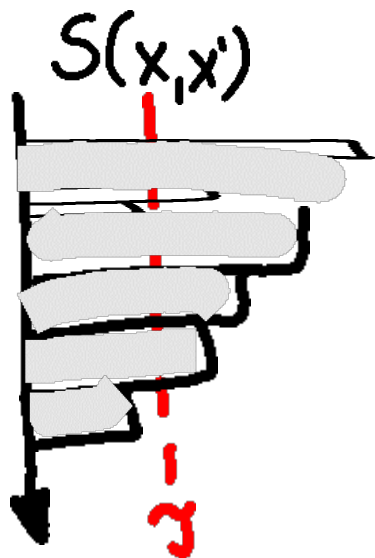
[Mallinson et al, 2017]



comes for free

	$P(x' x)$
Good movie	
Good film	
Great movie	
...	
Movie good	
0.35	
0.34	
...	
...	

Finding an adversary

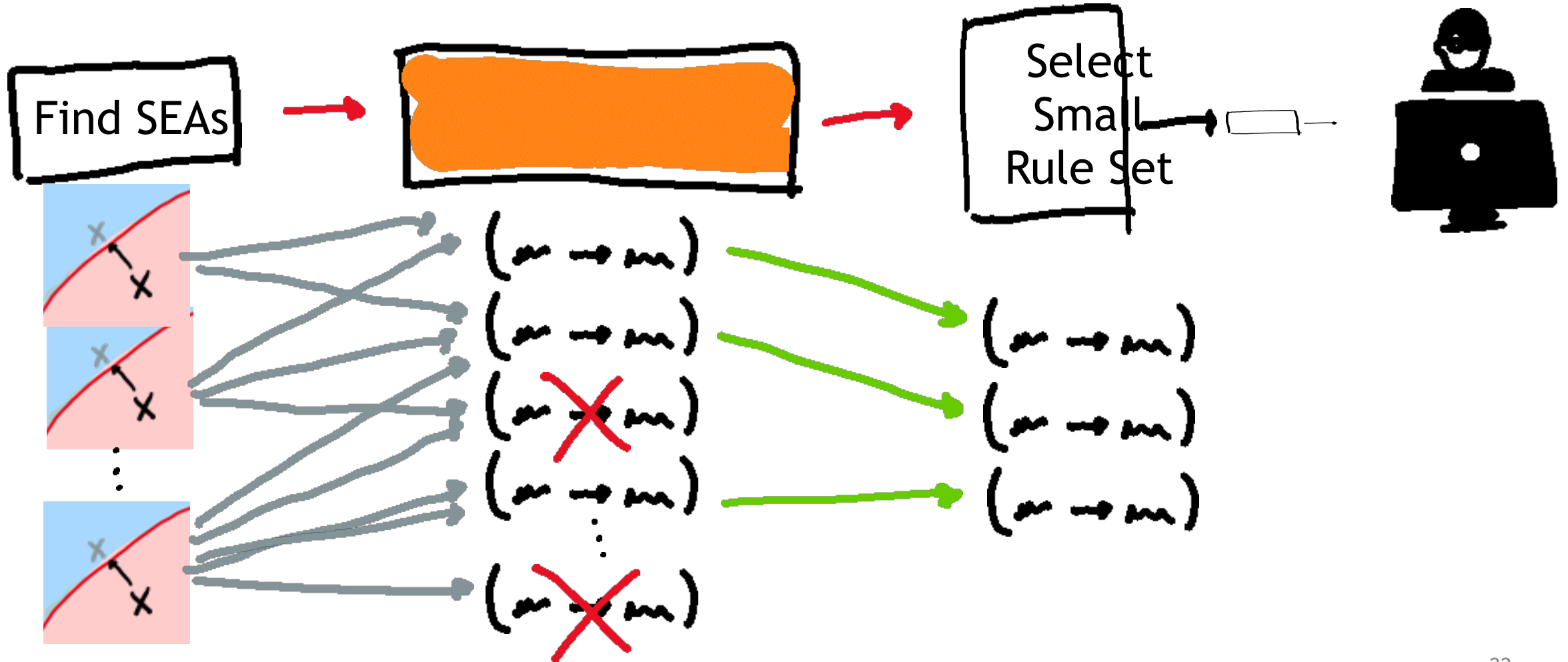


$x =$ What color is the tray? Pink

x'	$f(x')$
What color is the tray?	Green
What color is the tray?	Green
What color is it?	Green
What color is the tray?	Green
What color is the tray?	Green

Semantically Equivalent Adversarial Rules (SEARs)

From SEAs to Rules



Proposing Candidate Rules

What type of road sign is shown?



~~What~~ **Which** type of road sign is shown?

Candidate

Rules:



(What type → **Which type**)

(What NOUN → **Which NOUN**)

(WP type → **Which type**)

(WP NOUN → **Which NOUN**)

...



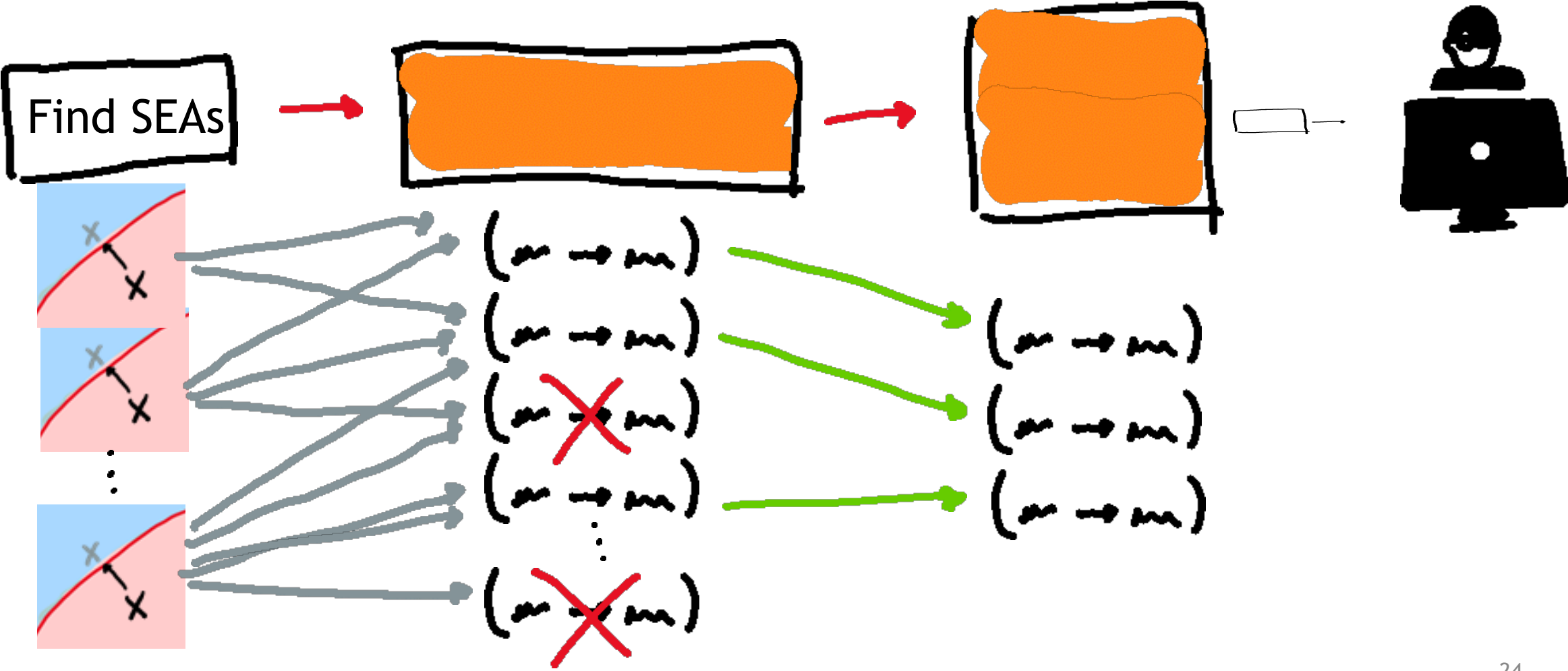
✓ What **Which** type of road sign is shown?

✗ What **Which** is the person looking at?

✗ What **Which** was I thinking?

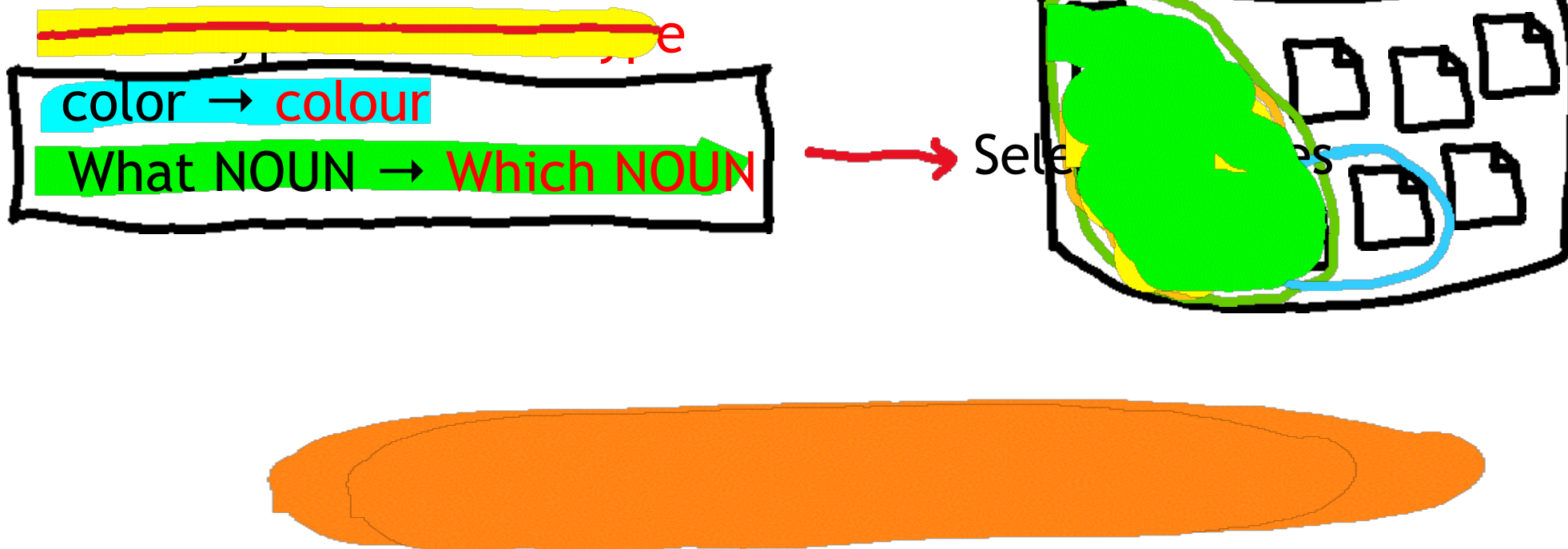


From SEAs to Rules



Semantically Equivalent Adversarial Rules (SEARS)

- ① High Adversary Count
- ② Non-Redundancy



Examples: VQA

Visual7a-Telling [Zhu et al 2016]

SEAR	Questions / SEAs	f(x)	Flips
WP VBZ → WP's	What has What's been cut?	Cake Pizza	3.3%
What NOUN → Which NOUN	What Which kind of floor is it?	Wood Marble	3.9%
color → colour	What color colour is the tray?	Pink Green	2.2%
ADV is → ADV's	Where is Where's the jet?	Sky Airport	2.1%

Examples: Machine Comprehension

BiDAF [Seo et al 2017]

SEAR	Questions / SEAs	f(x)	Flips
What VBZ → What's	What is What's the NASUWT?	Trade union Teachers in Wales	2%
What NOUN → Which NOUN	What resource Which resource was mined in the Newcastle area?	coal wool	1%
What VERB → So what VERB	What was So what was Ghandi's work called?	Satyagraha Civil Disobedience	2%
What VBD → And what VBD	What was And what was Kenneth Swezey's job?	journalist sleep	2%

Examples: Movie Review Sentiment Analysis

FastText [Joulin et al 2016]

SEAR	Reviews / SEAs	f(x)	Flips
movie → film	Yeah, the movie film pretty much sucked .	Neg Pos	2%
film → movie	Excellent film movie .	Pos Neg	1%
is → was	Ray Charles is was legendary .	Pos Neg	4%
this → that	Now this that is a movie I really dislike .	Neg Pos	1%

Experiments

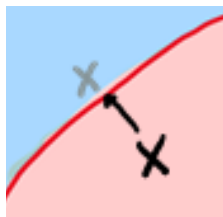
1. SEAs vs Humans

Set up

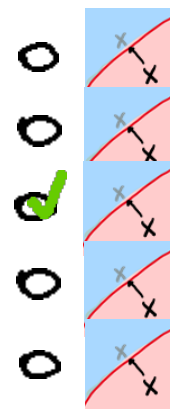
① Humans



② Top scored SEA



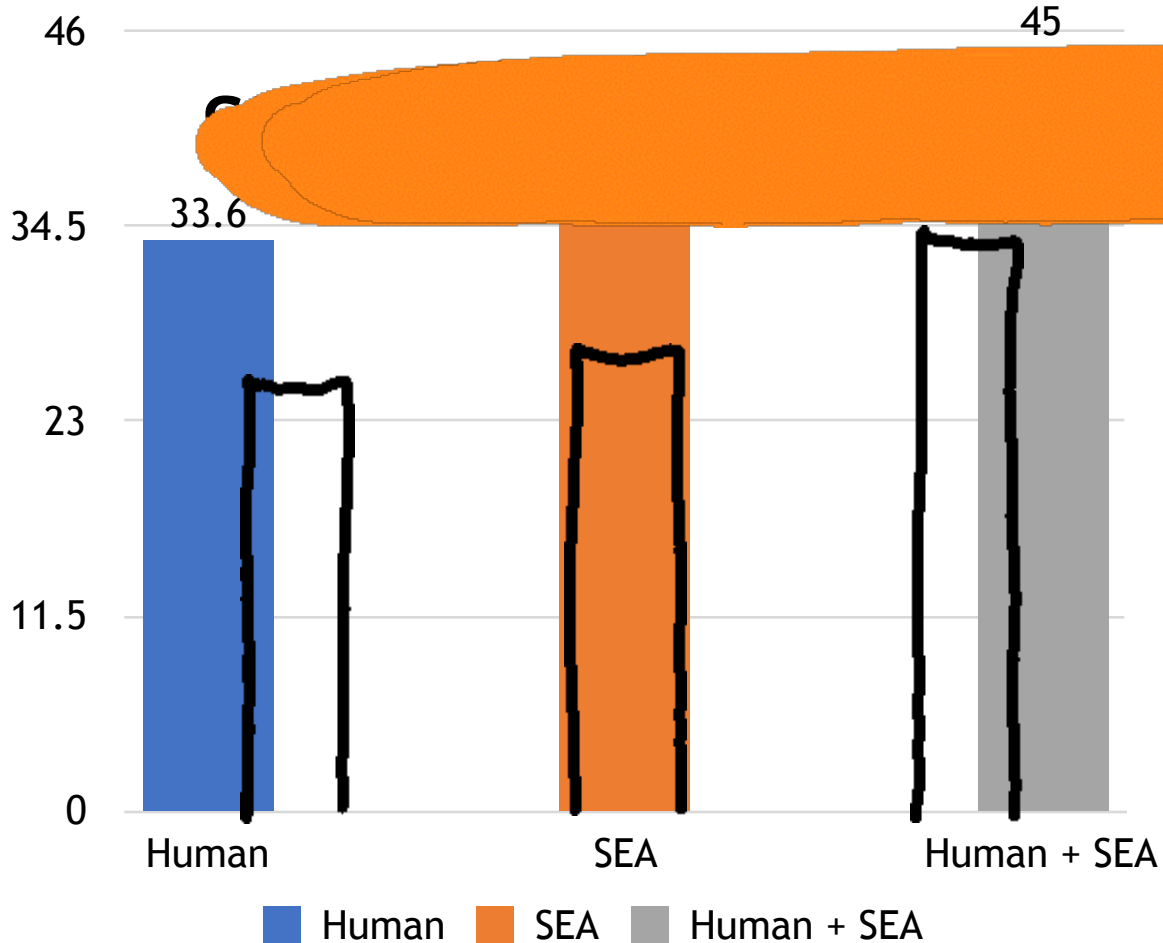
③ SEA (top 5) + Human



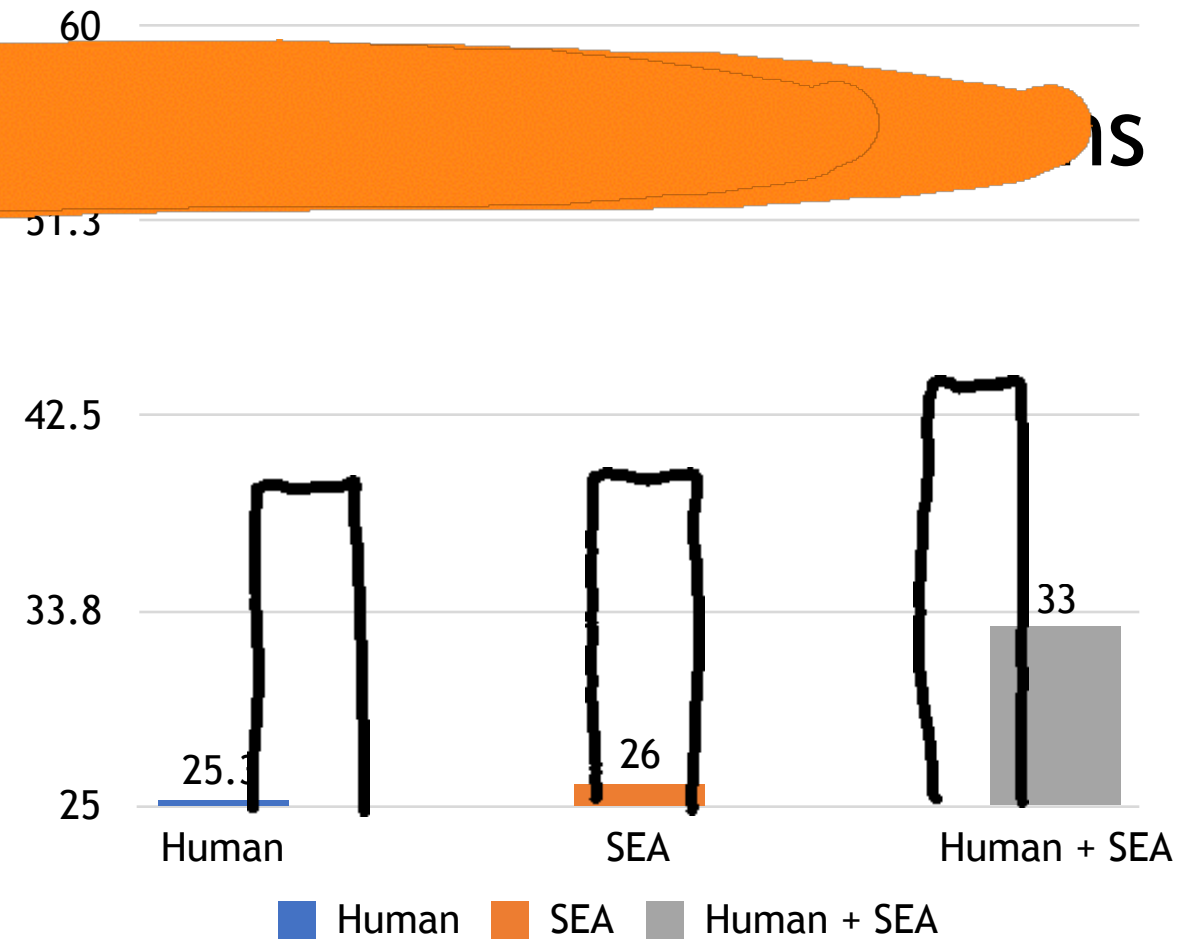
Evaluate adversaries for semantic equivalence

How often can SEAs be produced?

Visual Question Answering



Sentiment Analysis



Humans produce different adversaries:



They ~~are~~^{'re} so easy to love...

What ~~kind~~^{sort} of meat is on the boy's plate?



How many suitcases ~~are~~^{are on the shelf?}

~~Also great directing and photography~~
Photography and directing were on point.

2. SEARs vs Experts

Part 1: experts come up with rules

Individual predictions **Rules**

Try different rules

List of POS tags

Replace first instance of:

What NOUN

With:

Which NOUN

Saved Rules

replace(Who is, Who's) x

replace(color, colour) x



...

Results

replace(What NOUN, Which NOUN)










Mistake examples (click images to see them in more detail)

< **1** 2 3 4 > Compact

Image	Original	After rule
	Q: What color is the lampshade ? Answer: a) A light yellow. b) A bright red. c) A subtle green. d) A vivid orange.	Q: Which color is the lampshade ? Answer: a) A light yellow. b) A bright red. c) A subtle green. d) A vivid orange.
	Q: What food item is above the burger ? Answer: a) Fries. b) Chips. c) Cole slaw. d) Ketchup.	Q: Which food item is above the burger ? Answer: a) Fries. b) Chips. c) Cole slaw. d) Ketchup.

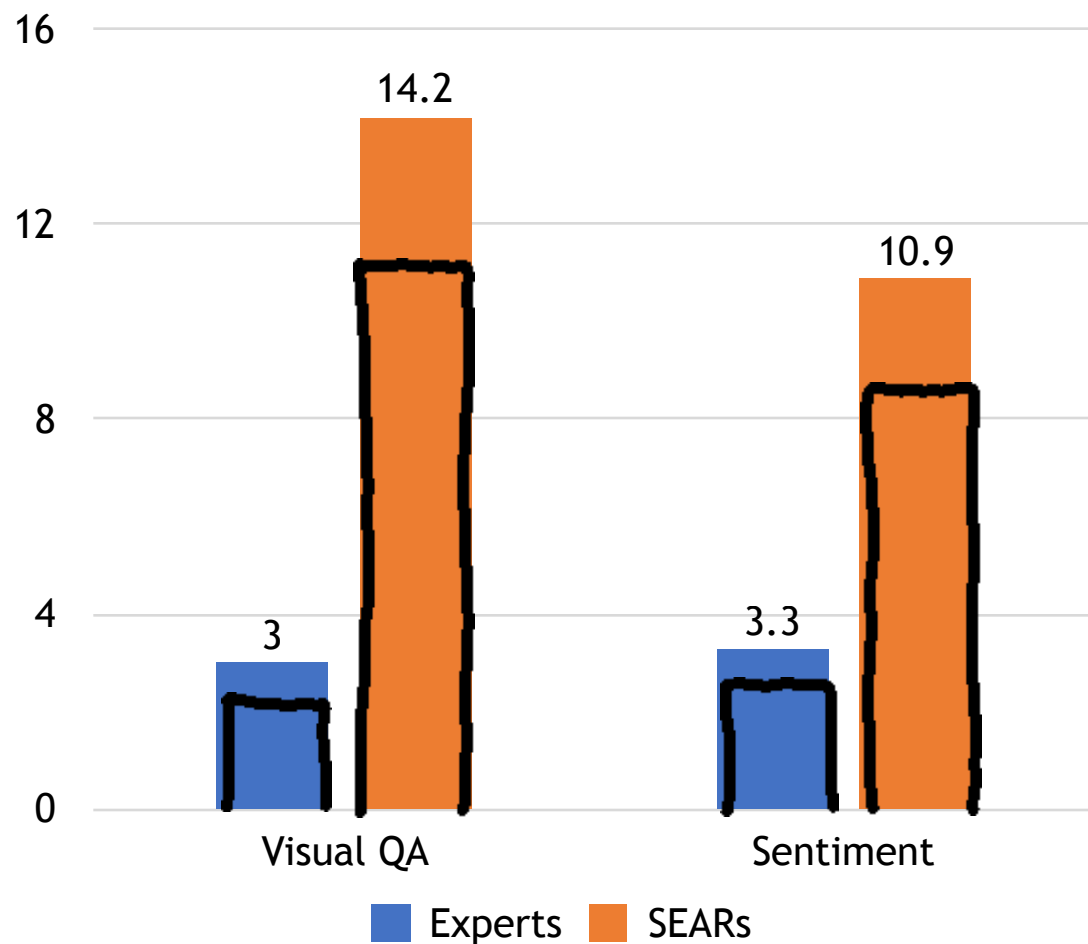
Q: ... S

Part 2: experts evaluate our SEARs

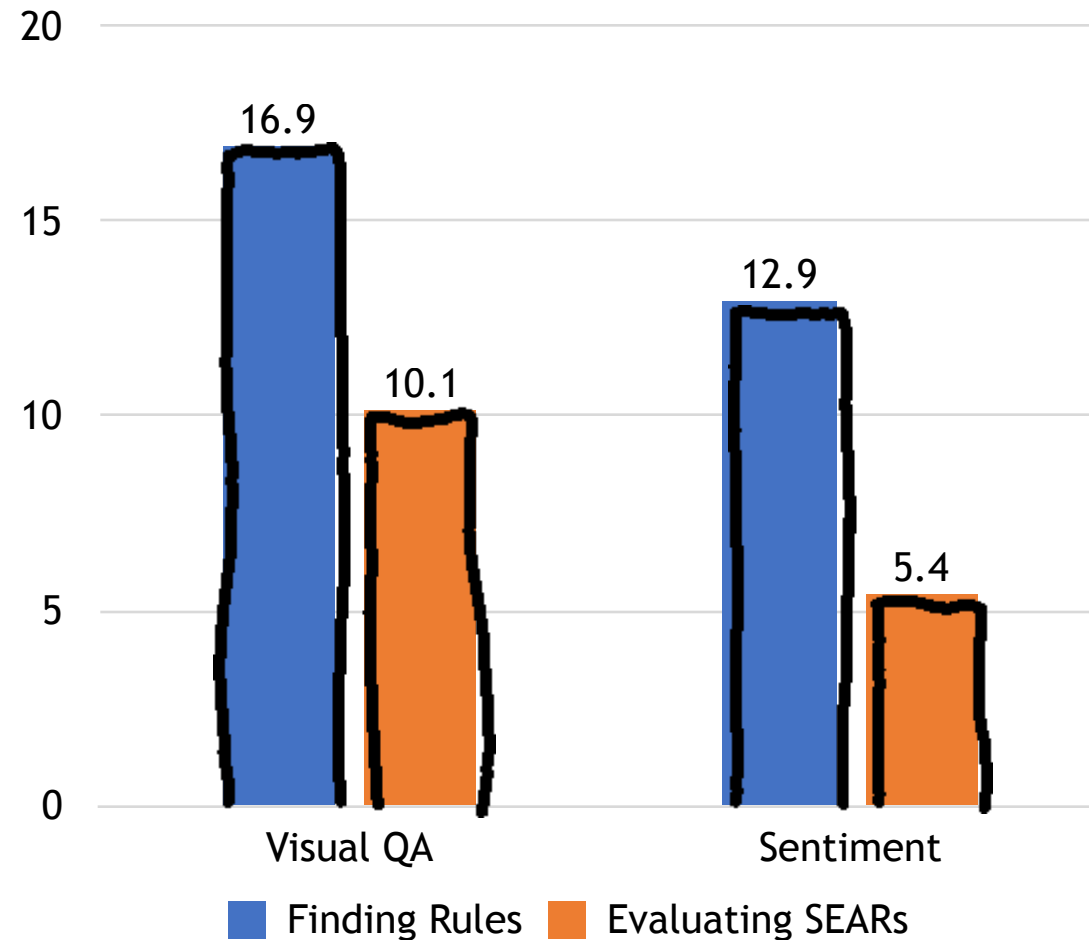
Rules to evaluate	Results												
<p>List of POS tags</p> <p>Please look at the rule results on the right</p> <p>The current rule</p> <p>repla</p> <p>Does the c</p> <p>No Yes</p>	<p>replace(What NOUN, Which NOUN)</p>												
<p>Progress</p> <p>1 of 20.</p>	<table border="1"><thead><tr><th>Image</th><th>Original</th><th>After rule</th></tr></thead><tbody><tr><td></td><td><p>Q: What color are the pots ?</p><p>Answer:</p><ul style="list-style-type: none">a) Silver.b) Black.c) White.d) Gold.</td><td><p>Q: Which color are the pots ?</p><p>Answer:</p><ul style="list-style-type: none">a) Silver.b) Black.c) White.d) Gold.</td></tr><tr><td></td><td><p>Q: What color is the lampshade ?</p><p>Answer:</p><ul style="list-style-type: none">a) A light yellow.b) A bright red.c) A subtle green.d) A vivid orange.</td><td><p>Q: Which color is the lampshade ?</p><p>Answer:</p><ul style="list-style-type: none">a) A light yellow.b) A bright red.c) A subtle green.d) A vivid orange.</td></tr><tr><td></td><td><p>Q: What animal is running in the background ?</p><p>Answer:</p><ul style="list-style-type: none">a) A dog.b) A horse.c) A llama.d) A kangaroo.</td><td><p>Q: Which animal is running in the background ?</p><p>Answer:</p><ul style="list-style-type: none">a) A dog.b) A horse.c) A llama.d) A kangaroo.</td></tr></tbody></table>	Image	Original	After rule		<p>Q: What color are the pots ?</p> <p>Answer:</p> <ul style="list-style-type: none">a) Silver.b) Black.c) White.d) Gold.	<p>Q: Which color are the pots ?</p> <p>Answer:</p> <ul style="list-style-type: none">a) Silver.b) Black.c) White.d) Gold.		<p>Q: What color is the lampshade ?</p> <p>Answer:</p> <ul style="list-style-type: none">a) A light yellow.b) A bright red.c) A subtle green.d) A vivid orange.	<p>Q: Which color is the lampshade ?</p> <p>Answer:</p> <ul style="list-style-type: none">a) A light yellow.b) A bright red.c) A subtle green.d) A vivid orange.		<p>Q: What animal is running in the background ?</p> <p>Answer:</p> <ul style="list-style-type: none">a) A dog.b) A horse.c) A llama.d) A kangaroo.	<p>Q: Which animal is running in the background ?</p> <p>Answer:</p> <ul style="list-style-type: none">a) A dog.b) A horse.c) A llama.d) A kangaroo.
Image	Original	After rule											
	<p>Q: What color are the pots ?</p> <p>Answer:</p> <ul style="list-style-type: none">a) Silver.b) Black.c) White.d) Gold.	<p>Q: Which color are the pots ?</p> <p>Answer:</p> <ul style="list-style-type: none">a) Silver.b) Black.c) White.d) Gold.											
	<p>Q: What color is the lampshade ?</p> <p>Answer:</p> <ul style="list-style-type: none">a) A light yellow.b) A bright red.c) A subtle green.d) A vivid orange.	<p>Q: Which color is the lampshade ?</p> <p>Answer:</p> <ul style="list-style-type: none">a) A light yellow.b) A bright red.c) A subtle green.d) A vivid orange.											
	<p>Q: What animal is running in the background ?</p> <p>Answer:</p> <ul style="list-style-type: none">a) A dog.b) A horse.c) A llama.d) A kangaroo.	<p>Q: Which animal is running in the background ?</p> <p>Answer:</p> <ul style="list-style-type: none">a) A dog.b) A horse.c) A llama.d) A kangaroo.											

Results

% correct predictions flipped

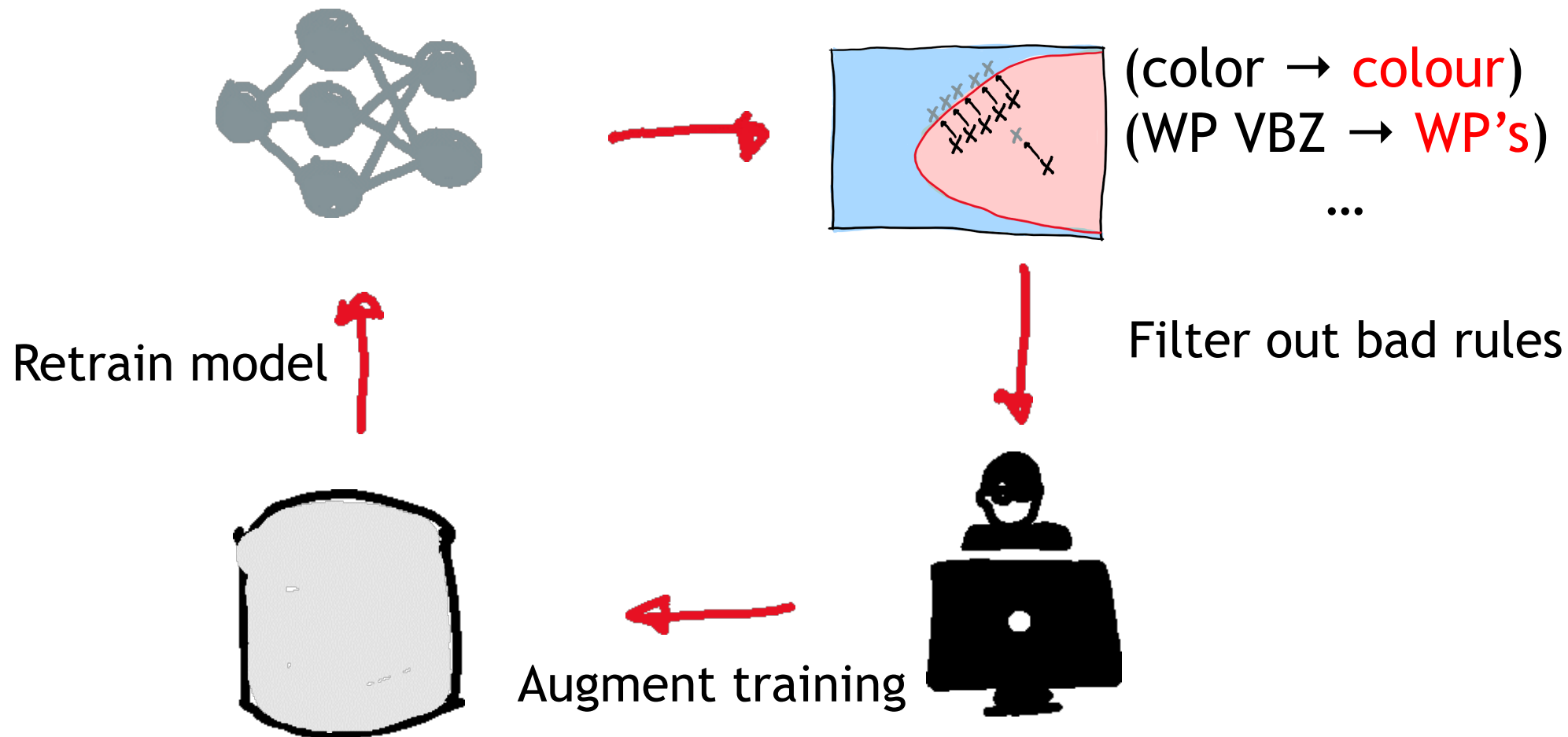


Time (minutes)

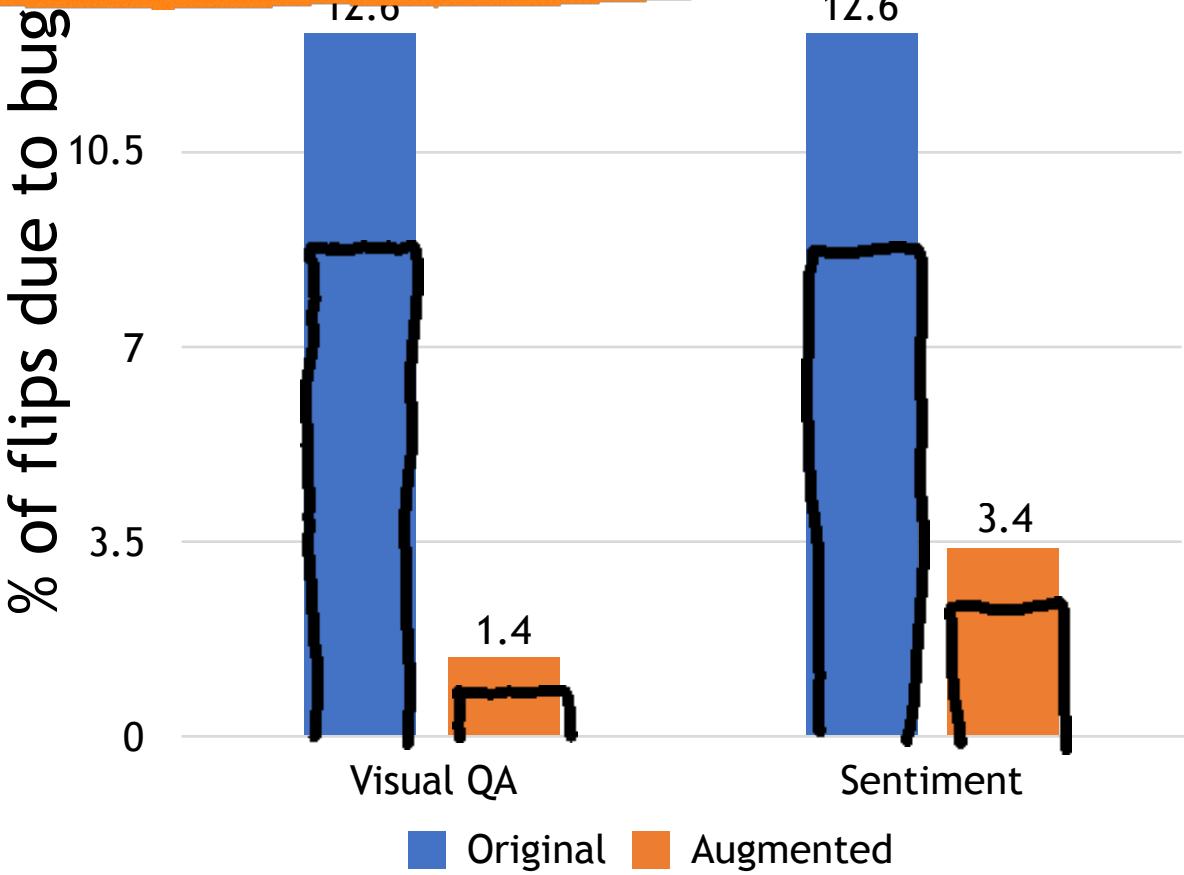


3. Fixing bugs

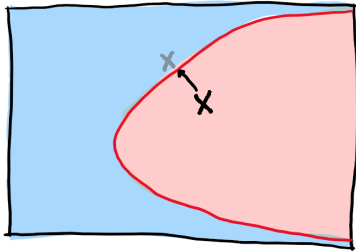
Closing the loop



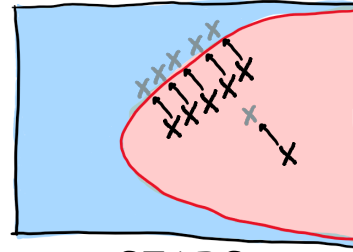
Results



Conclusion



SEA



SEARS



M



SS



them

Semantically Equivalent Adversarial Rules for Debugging NLP Models

Sameer Singh (UC Irvine)



Carlos Guestrin



Semantic scoring is still a research problem...

SEAR	Questions / SEAs	f(x)
on → in	What is on in the background? What is on in?	A building Mountains Lights The television
VBP → is	Where are is the water bottles? Where are is the people gathered?	Table Vending machine Living room Kitchen
VERB on → VERB	What is on the backbround? Where are the planes parked on ?	A building Mountains Concrete Landing strip



Problem: not comparable across instances



$$S(x, x') = \min \left(1, \frac{P(X'|X)}{P(X|X)} \right)$$

Examples: VQA

SEAR	Questions / SEAs	f(x)	Flips
WP VBZ → WP's	What has What's been cut?	Cake Pizza	3.3%
	Who is Who's holding the baby?	Woman Man	
What NOUN → Which NOUN	What Which kind of floor is it?	Wood Marble	3.9%
	What Which color is the jet?	Gray White	
color → colour	What color colour is the tray?	Pink Green	2.2%
	What color colour is the jet?	Gray Blue	
ADV is → ADV's	Where is Where's the jet?	Sky Airport	2.1%
	How is How's the desk?	Messy Empty	

Examples: Movie Review Sentiment Analysis

FastText [Joulin et al 2016]

SEAR	Reviews / SEAs	f(x)	Flips
movie → film	Yeah, the <i>movie</i> - film pretty much sucked .	Neg Pos	2%
	This is not <i>movie</i> - film making .	Neg Pos	
film → movie	Excellent <i>film</i> - movie .	Pos Neg	1%
	I'll give this <i>film</i> - movie 10 out of 10 !	Pos Neg	
is → was	Ray Charles <i>is</i> was legendary .	Pos Neg	4%
	It <i>is</i> was a really good show to watch .	Pos Neg	
this → that	Now <i>this</i> that is a movie I really dislike .	Neg Pos	1%
	The camera really likes her in <i>this</i> - that movie.	Pos Neg	

$$SEA(x, x') = 1 \left[S(x, x') > \tau \wedge f(x) \neq f(x') \right]$$

$$\max_{x, x'} S(x, x') > \tau \text{ s.t. } SEA(x, x') =$$