

<p>A young boy is playing basketball.</p> 	<p>Two dogs play in the grass.</p> 	<p>A dog swims in the water.</p> 	<p>A little girl in a pink shirt is swinging.</p> 
<p>A group of people walking down a street.</p> 	<p>A group of women dressed in formal attire.</p> 	<p>Two children play in the water.</p> 	<p>A dog jumps over a hurdle.</p> 

# Show, Attend and Tell

Neural Image Caption Generation with Visual Attention

Joanna Khek Cuina  
Huang Youqin  
Ju Qiaodan  
Wang Zhenhao  
Xu Shu  
Yang Changyong

**Group 35**

# Table of Contents



**01.**  
**Background  
Information**



**02.**  
**Technical Details**



**03.**  
**Reproduction  
and Extension**



**04.**  
**Main Results**



**05.**  
**Q & A**

# 01.



## **Background Information**

Introduction to Paper

# Paper Introduction

- Last revised on 19 April 2016 and cited by~8000

---

## Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

---

Kelvin Xu  
Jimmy Lei Ba  
Ryan Kiros  
Kyunghyun Cho  
Aaron Courville  
Ruslan Salakhutdinov  
Richard S. Zemel  
Yoshua Bengio

KELVIN.XU@UMONTREAL.CA  
JIMMY@PSI.UTORONTO.CA  
RKIROS@CS.TORONTO.EDU  
KYUNGHYUN.CHO@UMONTREAL.CA  
AARON.COURVILLE@UMONTREAL.CA  
RSALAKHU@CS.TORONTO.EDU  
ZEMEL@CS.TORONTO.EDU  
FIND-ME@THE.WEB

<https://arxiv.org> > cs

### Show, Attend and Tell: Neural Image Caption Generation with ...

by K Xu · 2015 · Cited by 8710 — Abstract: Inspired by recent work in machine translation and object detection, we introduce an attention based model that automatically ...

# Motivation

How would you  
teach young  
children a  
language?



# Image Captioning

- The idea of taking an image and then producing a sentence that describes the image

A young boy is playing basketball.



Two dogs play in the grass.



A dog swims in the water.



A little girl in a pink shirt is swinging.



A group of people walking down a street.



A group of women dressed in formal attire.



Two children play in the water.



A dog jumps over a hurdle.



# Challenges



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



**Objects**



**Objects  
+  
Relationships**



# Attention Framework



**Here!**



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



# 02.



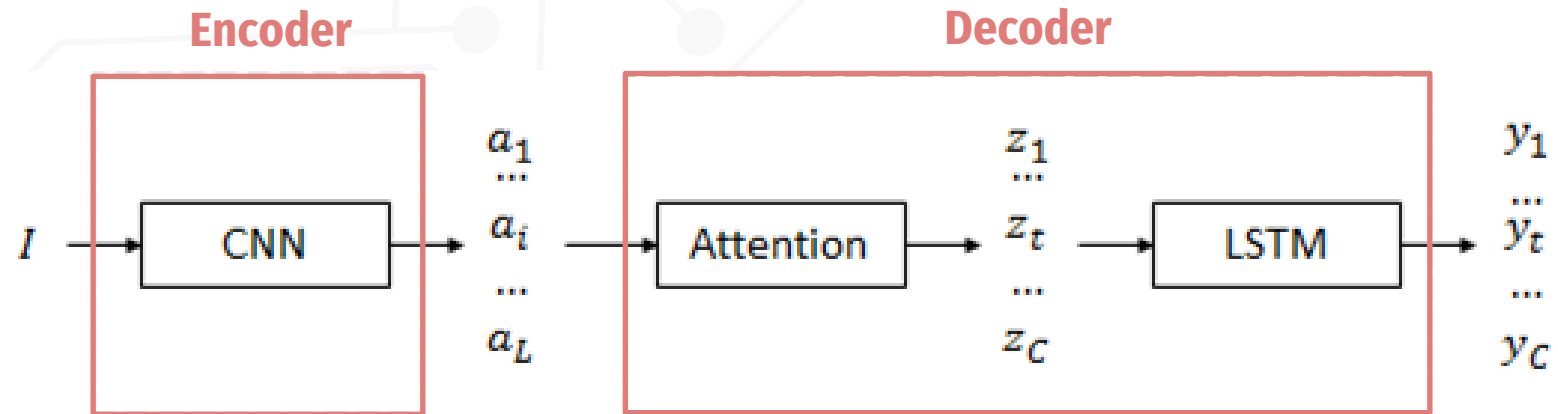
## Technical Details

Key concepts

# Overall Framework



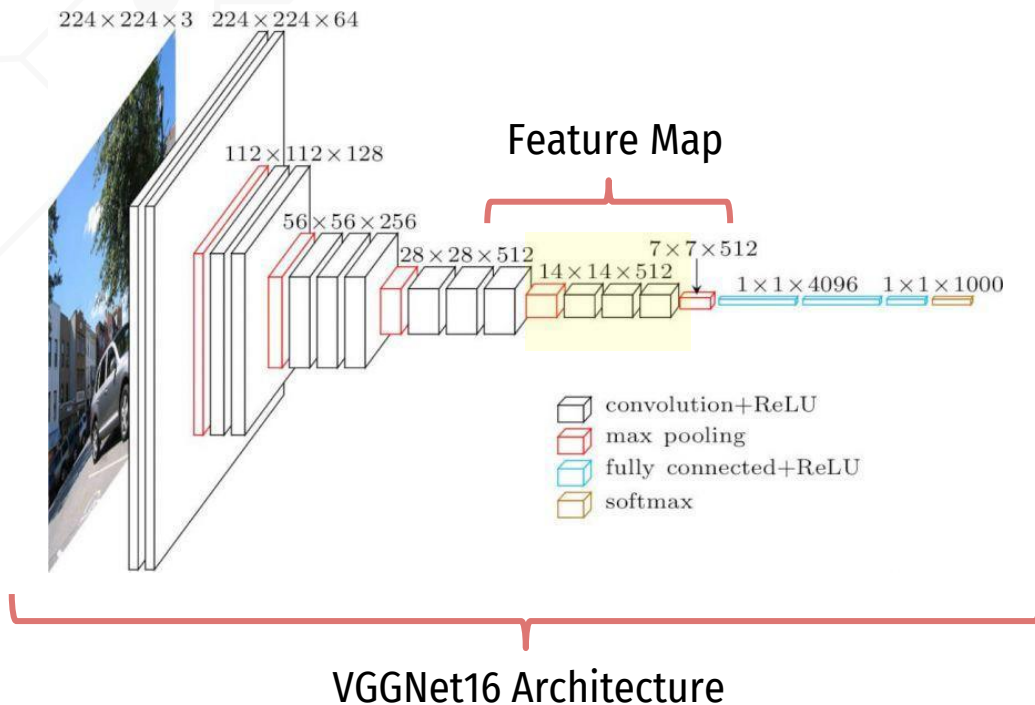
# Overall Framework



# Encoder: Convolutional Feature Map



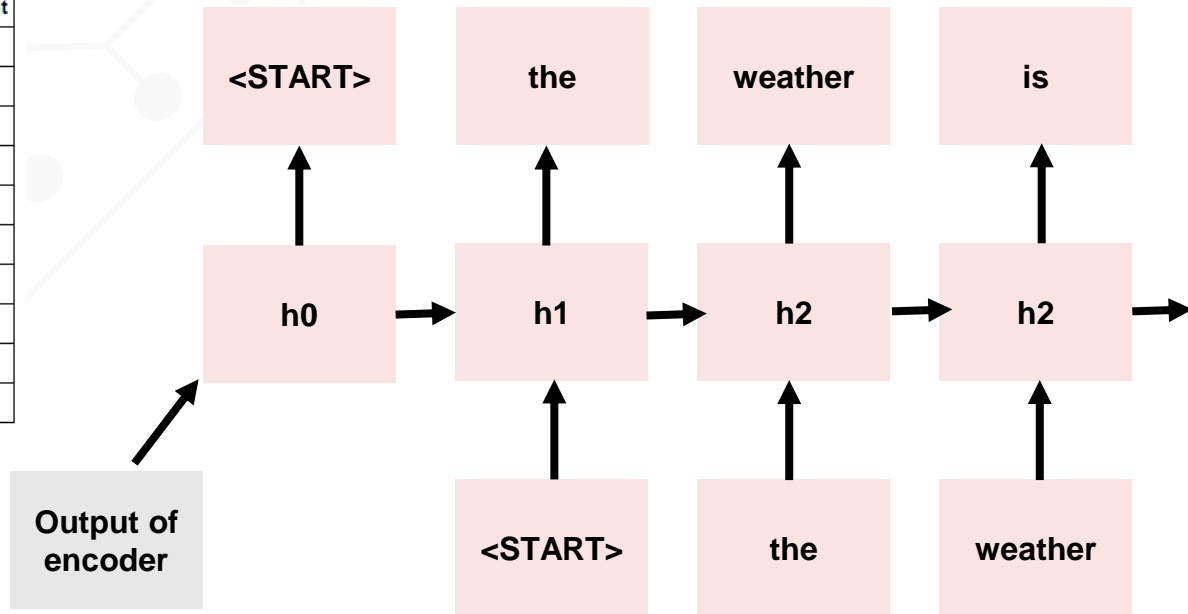
224 x 224 x 3 Image



# Decoder with Attention: Recurrent Neural Network

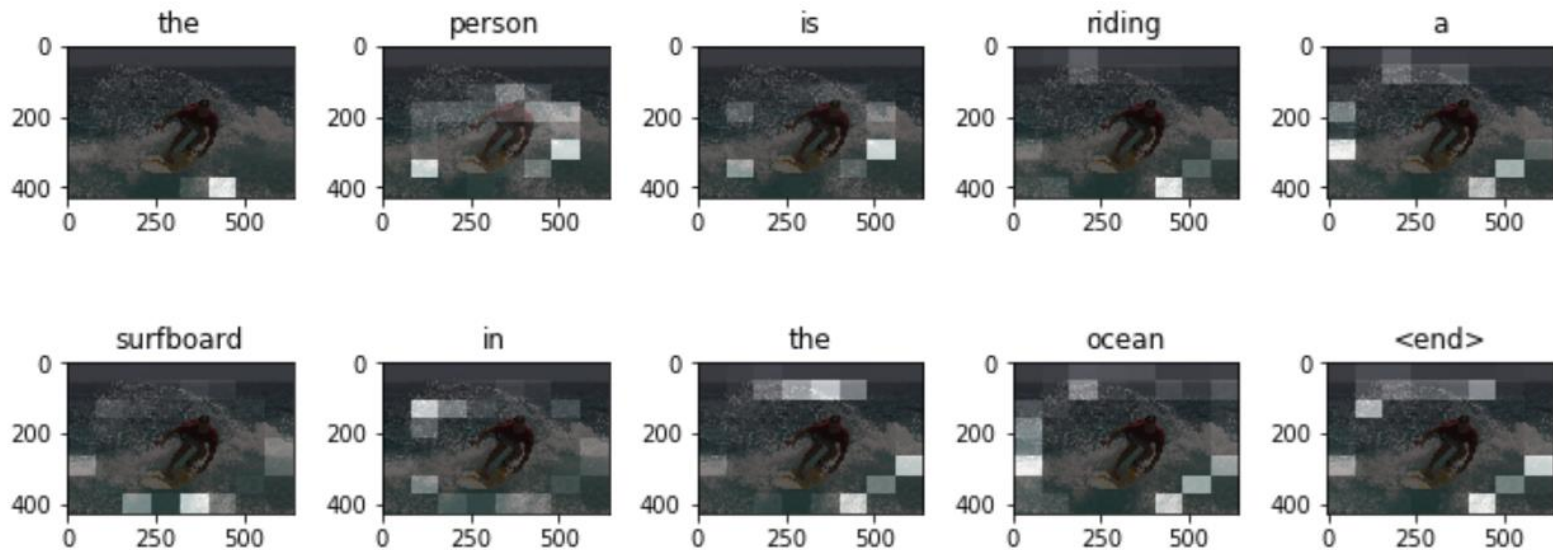
	Training Seq	Expected Output
1	"START"	the
2	the	weather
3	the weather	is
4	the weather is	great
5	the weather is great	and
6	the weather is great and	the
7	the weather is great and the	time
8	the weather is great and the time	is
9	the weather is great and the time is	now
10	the weather is great and the time is now	"END"

$P(\text{the} | \text{<START>})$   
 $p(\text{weather} | \text{<START>, the})$   
 $p(\text{is} | \text{<START>, the, weather})$

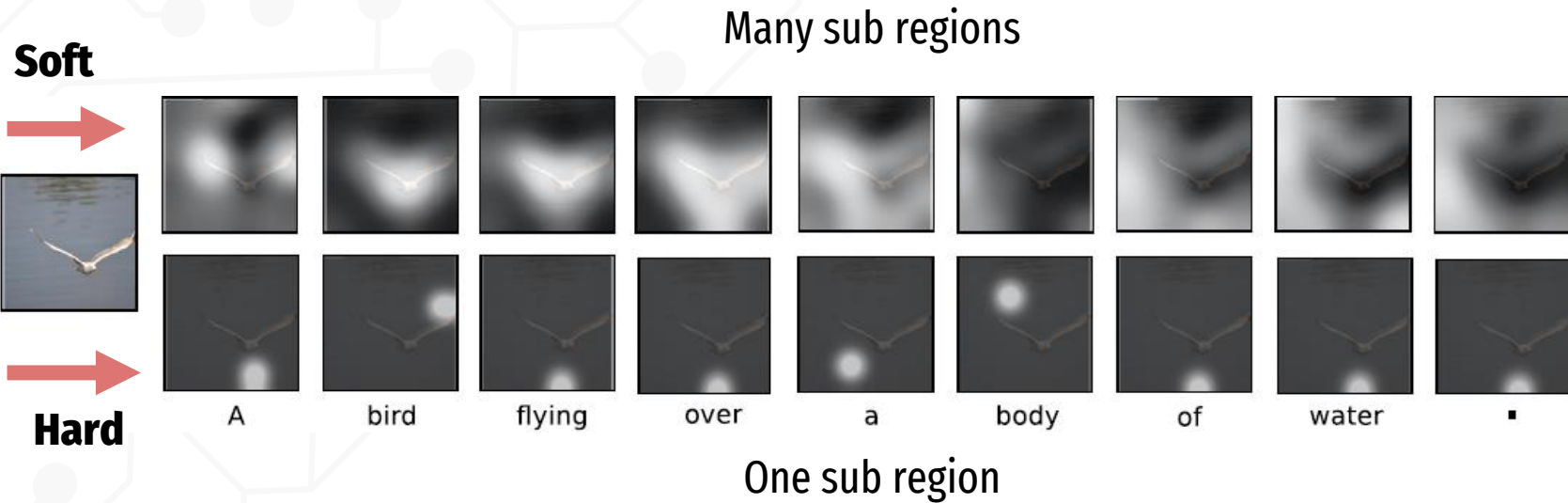


# Decoder with Attention: Recurrent Neural Network

Prediction Caption: the person is riding a surfboard in the ocean <end>



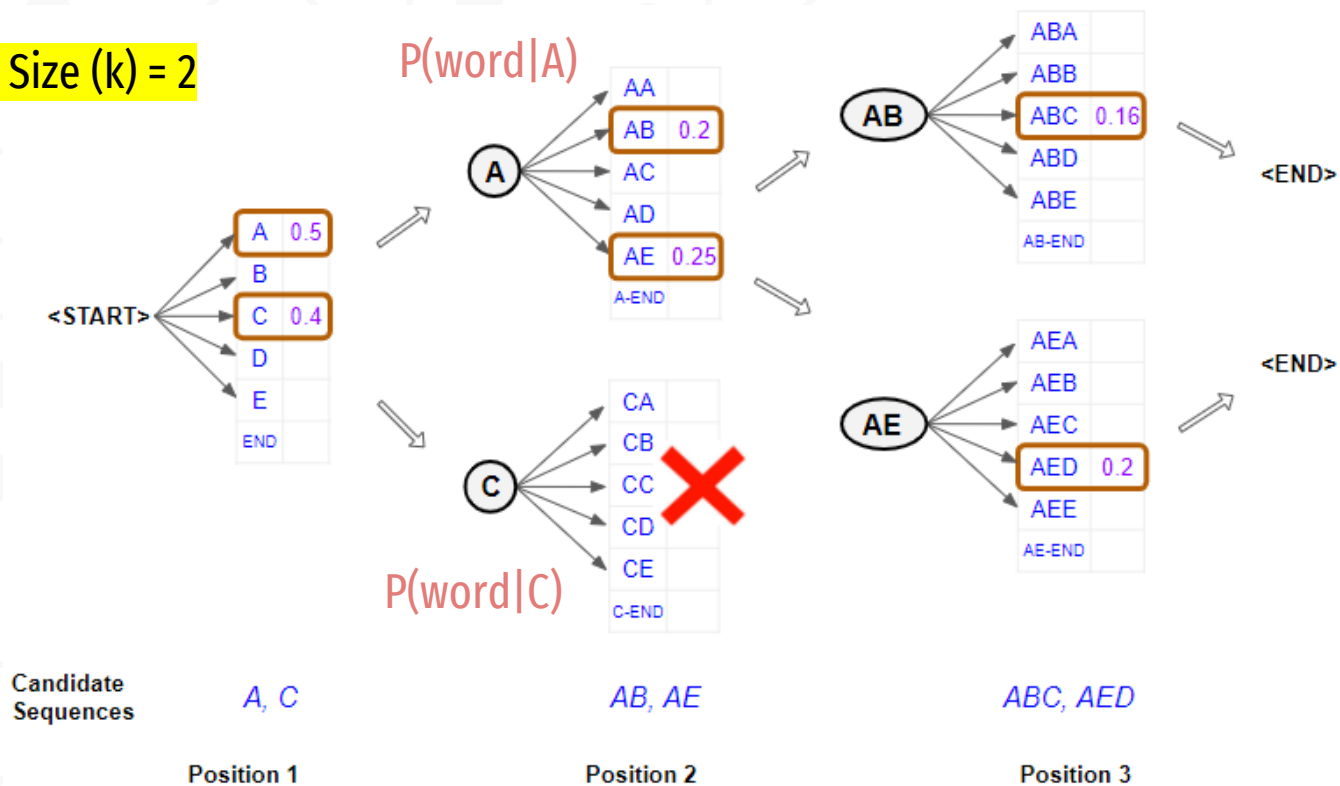
# “Soft” and “Hard” Attention





# Beam Search Algorithm

Beam Size ( $k$ ) = 2



# BLEU Score

- Evaluation Metric - **Bi**Lingual **E**valuation **U**nderstudy

**Input:** Bud Powell était un pianiste de légende.

**Reference:** Bud Powell was a legendary pianist.

**sentence BLEU**  
(0-100)

**Candidate 1:** Bud Powell was a legendary pianist.

100

**Candidate 2:** Bud Powell was a historic piano player.

46.7

**Candidate 3:** Bud Powell was a New Yorker.

54.1

# Model Comparison

Dataset	Model	BLEU				METEOR
		BLEU-1	BLEU-2	BLEU-3	BLEU-4	
Flickr8k	Google NIC(Vinyals et al., 2014) <sup>†<math>\Sigma</math></sup>	63	41	27	—	—
	Log Bilinear (Kiros et al., 2014a) <sup>o</sup>	65.6	42.4	27.7	17.7	17.31
	Soft-Attention	<b>67</b>	<b>44.8</b>	<b>29.9</b>	<b>19.5</b>	<b>18.93</b>
	Hard-Attention	<b>67</b>	<b>45.7</b>	<b>31.4</b>	<b>21.3</b>	<b>20.30</b>
Flickr30k	Google NIC <sup>†<math>\circ\Sigma</math></sup>	66.3	42.3	27.7	18.3	—
	Log Bilinear	60.0	38	25.4	17.1	16.88
	Soft-Attention	<b>66.7</b>	<b>43.4</b>	<b>28.8</b>	<b>19.1</b>	<b>18.49</b>
	Hard-Attention	<b>66.9</b>	<b>43.9</b>	<b>29.6</b>	<b>19.9</b>	<b>18.46</b>
COCO	CMU/MS Research (Chen & Zitnick, 2014) <sup>a</sup>	—	—	—	—	20.41
	MS Research (Fang et al., 2014) <sup>†<math>a</math></sup>	—	—	—	—	20.71
	BRNN (Karpathy & Li, 2014) <sup>o</sup>	64.2	45.1	30.4	20.3	—
	Google NIC <sup>†<math>\circ\Sigma</math></sup>	66.6	46.1	32.9	24.6	—
	Log Bilinear <sup>o</sup>	70.8	48.9	34.4	24.3	20.03
	Soft-Attention	<b>70.7</b>	<b>49.2</b>	<b>34.4</b>	<b>24.3</b>	<b>23.90</b>
	Hard-Attention	<b>71.8</b>	<b>50.4</b>	<b>35.7</b>	<b>25.0</b>	<b>23.04</b>

**03.**



## **Reproduction and Extension**

# Dataset

## VizWiz Caption Dataset



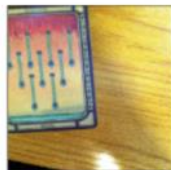
some sort of USDA choice beef sirloin steaks meant to eat

A box of four 6 oz USDA sirloin steaks.

A photo of Schwan's USDA choice Beef Sirloin steaks.

A box of USDA choice beef sirloin steak is on a counter.

A package of USDA Choice Beef Sirloin Steaks, containing four six-ounce steaks.



A tarot card labeled nine on top and wands on the bottom with a picture of nine wands in the middle.

A playing card showing "nine wands" is on a wooden surface.

Nine wands tarot card with a picture of nine wands over fire.

A trading card is lying on top of a wooden table.

A card with objects symbolizing the number nine placed near the edge of a wooden table

## MSCOCO Dataset



A person riding a motorcycle on a dirt road.



Two dogs play in the grass.

# Dataset

## VizWiz Caption Dataset



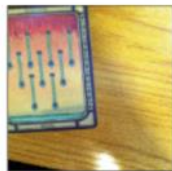
some sort of USDA choice beef sirloin steaks meant to eat

A box of four 6 oz USDA sirloin steaks.

A photo of Schwan's USDA choice Beef Sirloin steaks.

A box of USDA choice beef sirloin steak is on a counter.

A package of USDA Choice Beef Sirloin Steaks, containing four six-ounce steaks.



A tarot card labeled nine on top and wands on the bottom with a picture of nine wands in the middle.

A playing card showing "nine wands" is on a wooden surface.

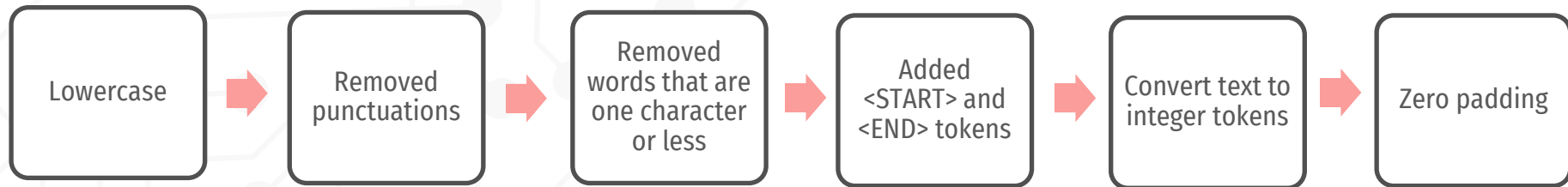
Nine wands tarot card with a picture of nine wands over fire.

A trading card is lying on top of a wooden table.

A card with objects symbolizing the number nine placed near the edge of a wooden table

- ~23K training images
- ~100K training captions

# Data Pre-processing



```
['<start> frozen meal sitting on flat horizontal surface <end>',  
'<start> package of corned beef sitting on the table <end>',  
'<start> an boxed package of corned beef frozen entree <end>',  
'<start> box of microwave corned beef meal <end>',
```



# Reproduction and Extension

	Used in paper	Reproduction	Extension
CNN Encoder	VGGNet16	VGGNet16, EfficientNet-B0 InceptionNetV3, MobileNetv2, DenseNet121, ResNet101	
Attention	Soft and Hard Attention	Soft Attention	
RNN Decoder	LSTM	LSTM	GRU
Evaluation Metric	BLEU (Beam Search)	BLEU (Beam Search)	BLEU (Greedy Search) and Training Time

04.



## **Main Results**

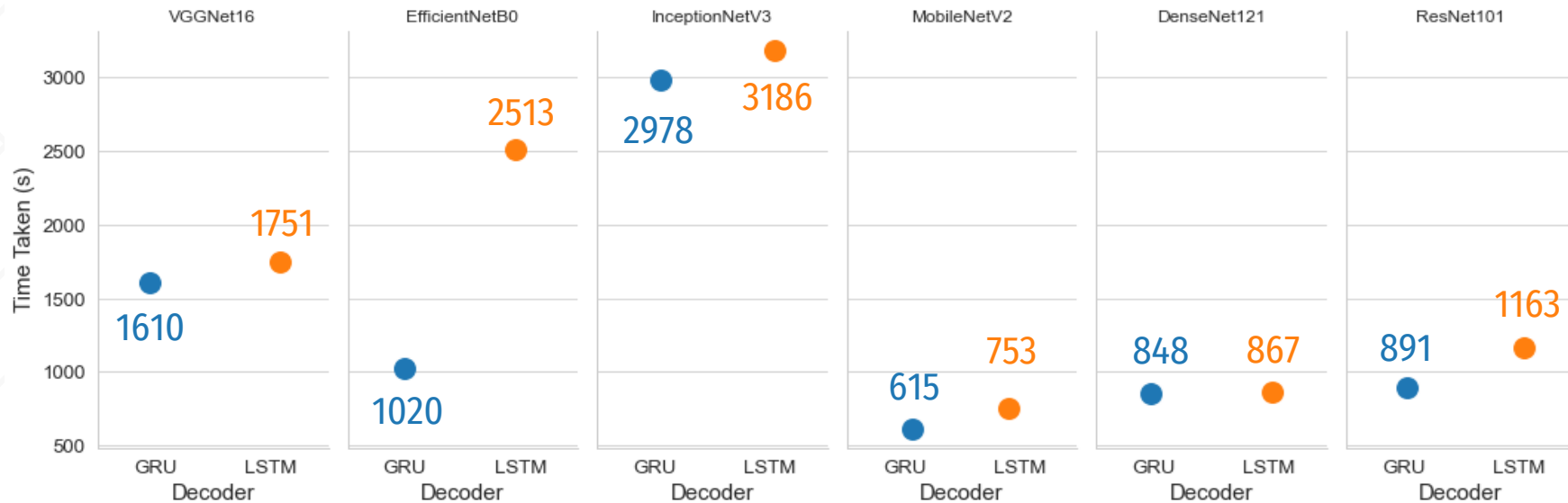
# LSTM V.S GRU Performance

- Batch size = 64
- Top words = 5000

Architecture	Decoder	Train Time (s)	Beam Search (k=3)			
			BLEU-1	BLEU-2	BLEU-3	BLEU-4
VGGNet16	GRU	1610	40.50	33.05	27.49	23.94
	LSTM	1751	42.41	34.15	27.75	23.90
EfficientNetB0	GRU	1020	45.14	37.29	30.27	26.25
	LSTM	2513	39.29	31.44	23.89	20.09
InceptionNetV3	GRU	2978	41.38	33.20	25.65	21.87
	LSTM	3186	43.30	38.50	33.20	27.89
MobileNetV2	GRU	615	46.15	37.41	29.21	24.52
	LSTM	753	46.13	36.78	28.33	23.23
DenseNet121	GRU	848	49.40	34.14	20.80	14.06
	LSTM	867	43.33	29.92	18.14	11.90
ResNet101	GRU	891	44.13	36.83	30.20	26.49
	LSTM	1163	51.03	41.29	33.15	28.74

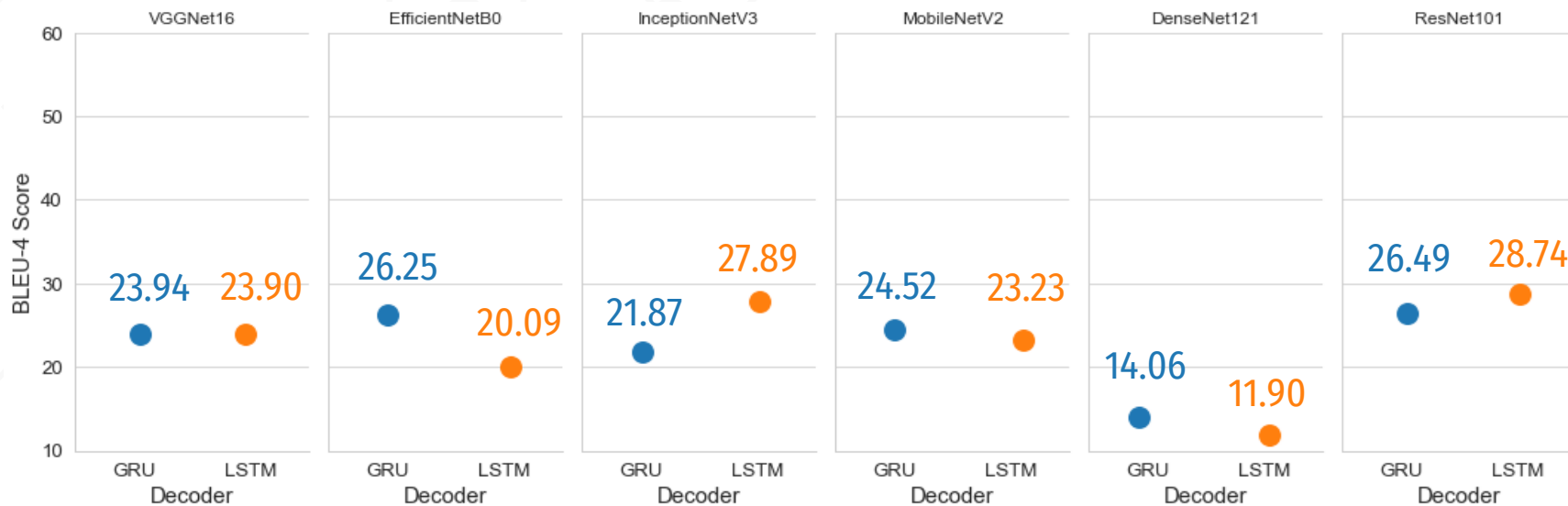
# LSTM V.S GRU Performance

- Training Time



# LSTM V.S GRU Performance

- BLEU-4 Score



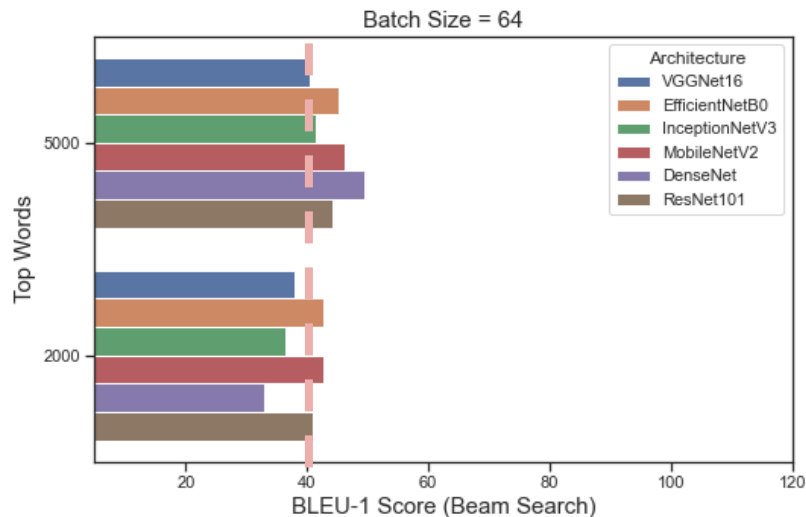
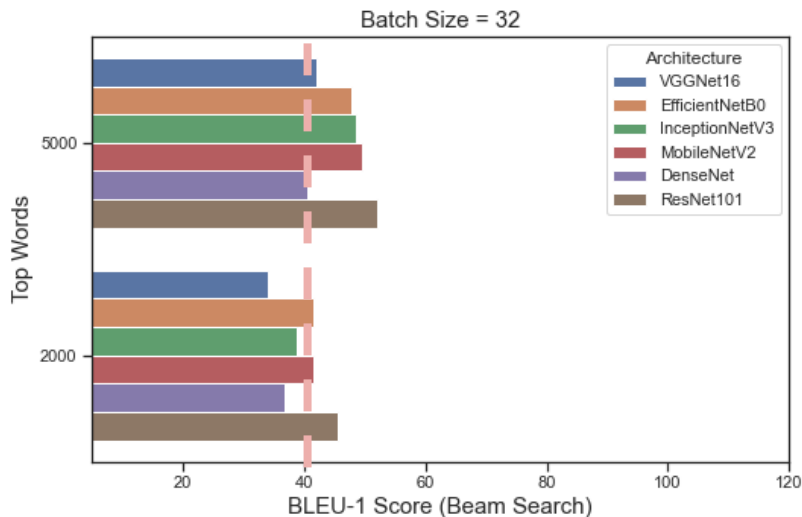
# Hyperparameters Tuning

- GRU
- Batch Size = 32, 64
- Top Words = 2000, 5000

Architecture	Batch Size	Top Words	BLEU-1	Beam Search (k=3)		BLEU-4
				BLEU-2	BLEU-3	
VGGNet16	32	2000	34.14	27.49	22.87	19.95
	64	2000	37.83	29.89	24.48	21.08
	32	5000	41.94	34.27	28.31	24.72
	64	5000	40.50	33.05	27.49	23.94
EfficientNetB0	32	2000	41.66	34.12	27.69	24.08
	64	2000	42.58	24.53	38.01	24.04
	32	5000	47.70	39.20	31.67	27.45
	64	5000	45.14	37.29	30.27	26.25
InceptionNetV3	32	2000	38.77	28.71	21.70	17.49
	64	2000	36.35	27.98	22.40	19.57
	32	5000	48.45	38.67	29.72	25.37
	64	5000	41.38	33.20	25.65	21.87
MobileNetV2	32	2000	41.58	32.82	25.68	21.18
	64	2000	42.78	33.90	26.27	21.52
	32	5000	49.54	40.34	31.95	26.95
	64	5000	46.15	37.41	29.21	24.52
DenseNet	32	2000	36.86	23.73	14.75	10.04
	64	2000	32.86	20.60	13.24	9.17
	32	5000	40.51	28.29	17.46	11.97
	64	5000	49.40	34.14	20.80	14.06
ResNet101	32	2000	45.66	37.01	29.56	25.47
	64	2000	41.05	33.07	27.26	23.62
	32	5000	52.02	42.52	34.13	29.23
	64	5000	44.13	36.83	30.20	26.49

# Hyperparameters Tuning (BLEU-1)

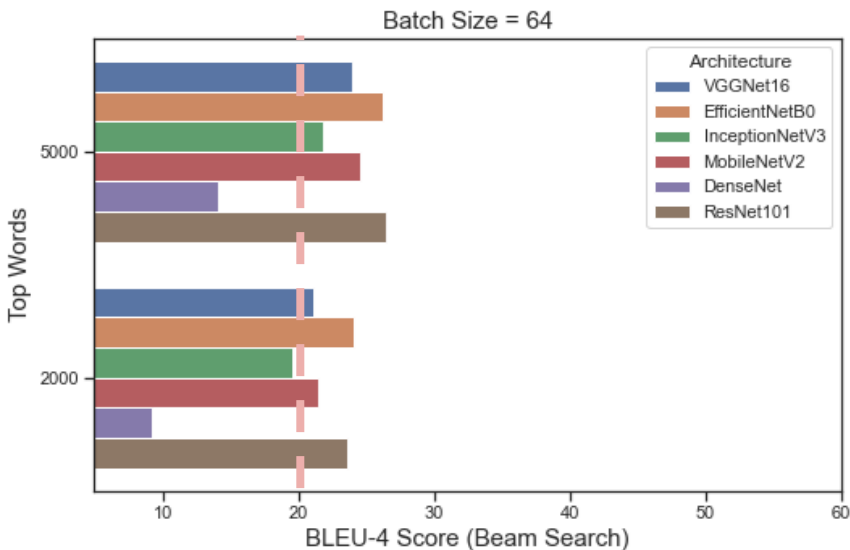
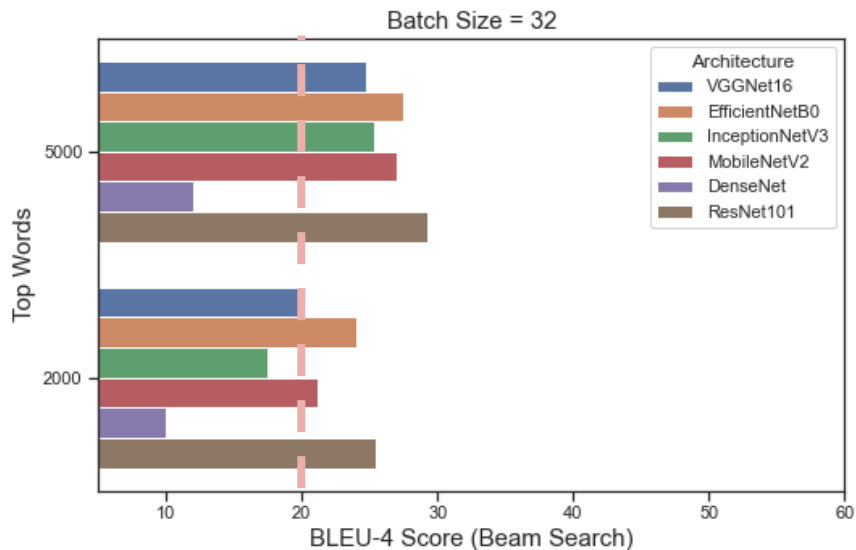
- Smaller batch size and larger pool of words led to a slightly better performance





# Hyperparameters Tuning (BLEU-4)

- Smaller batch size and larger pool of words led to a slightly better performance



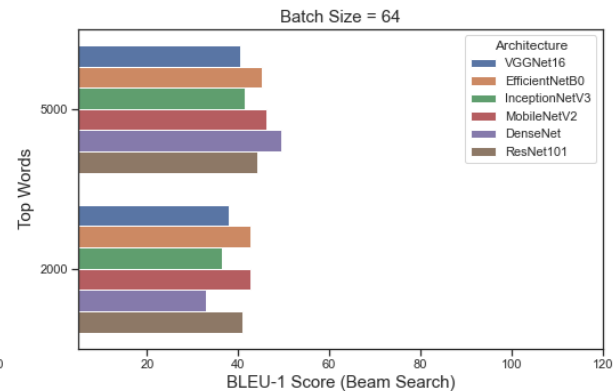
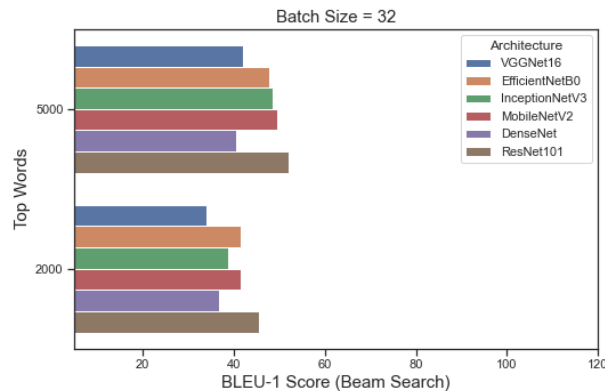
# Greedy Search

- GRU
- Batch Size = 32, 64
- Top Words = 2000, 5000

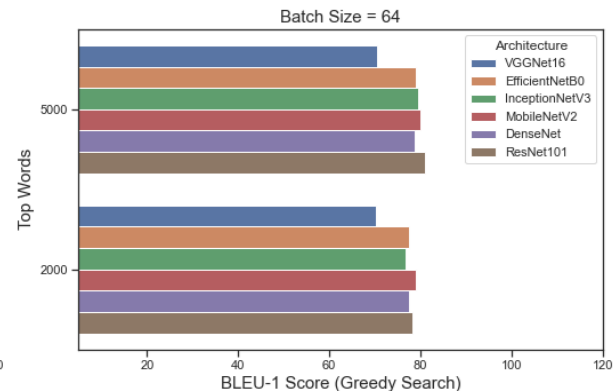
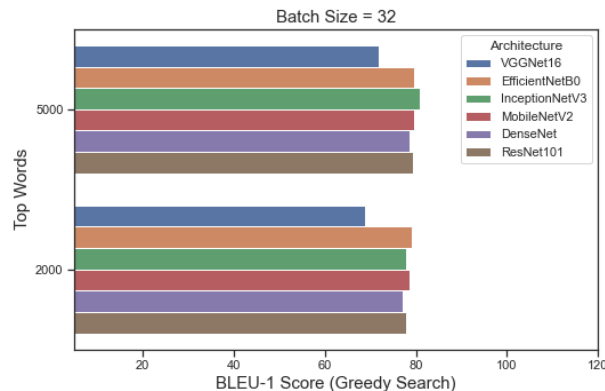
Architecture	Batch Size	Top Words	BLEU-1	Greedy Search		BLEU-4
				BLEU-2	BLEU-3	
VGGNet16	32	2000	68.83	49.81	32.55	25.13
	64	2000	70.34	51.47	33.78	26.00
	32	5000	71.83	53.38	36.00	28.36
	64	5000	70.52	52.27	34.96	27.25
EfficientNetB0	32	2000	79.17	58.07	39.69	32.39
	64	2000	77.49	55.38	35.71	27.98
	32	5000	79.56	59.74	41.35	33.72
	64	5000	78.88	57.65	37.40	29.21
InceptionNetV3	32	2000	77.88	53.80	32.53	23.48
	64	2000	76.73	52.71	31.58	24.17
	32	5000	80.81	53.49	36.92	31.30
	64	5000	79.58	56.32	34.54	26.32
MobileNetV2	32	2000	78.60	56.23	36.73	28.21
	64	2000	78.91	56.35	36.03	27.05
	32	5000	79.64	59.24	40.06	31.70
	64	5000	80.10	58.48	38.06	29.27
DenseNet	32	2000	77.19	48.50	22.95	13.29
	64	2000	77.59	49.58	23.79	13.79
	32	5000	78.72	50.30	23.45	13.13
	64	5000	78.83	50.47	23.83	13.75
ResNet101	32	2000	77.96	56.85	37.99	30.34
	64	2000	78.28	56.03	36.62	29.32
	32	5000	79.44	59.64	41.12	33.29
	64	5000	80.90	60.56	41.47	33.84

# Beam Search V.S Greedy Search (BLEU-1)

Beam  
Search

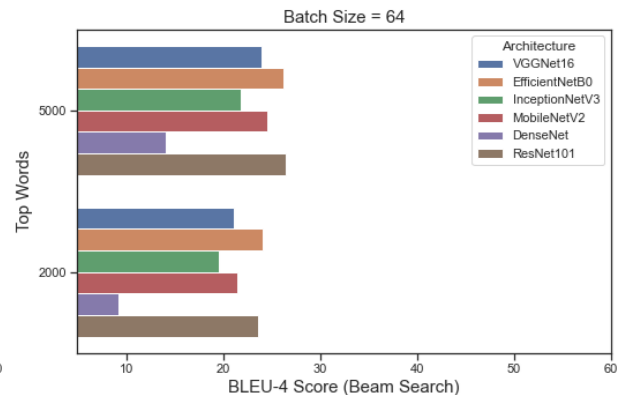
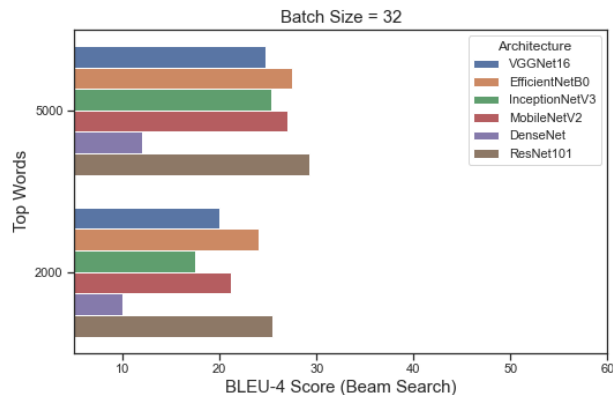


Greedy  
Search

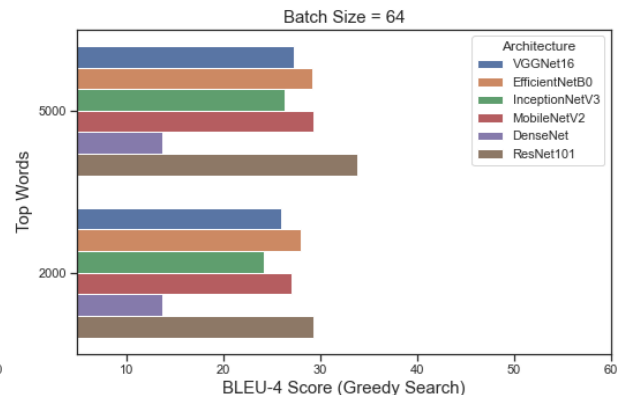
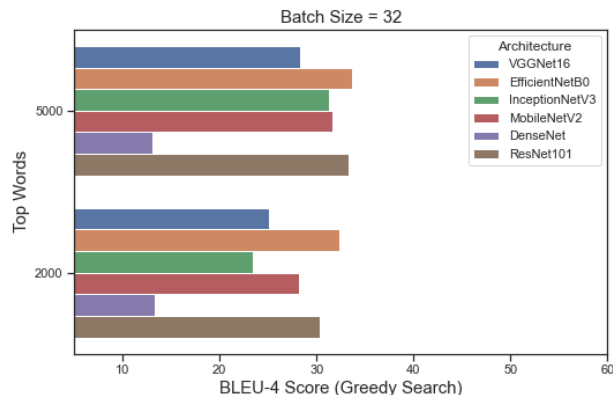


# Beam Search V.S Greedy Search (BLEU-4)

Beam  
Search



Greedy  
Search



# Good Results

Real Captions:

see pack of food sitting on the table

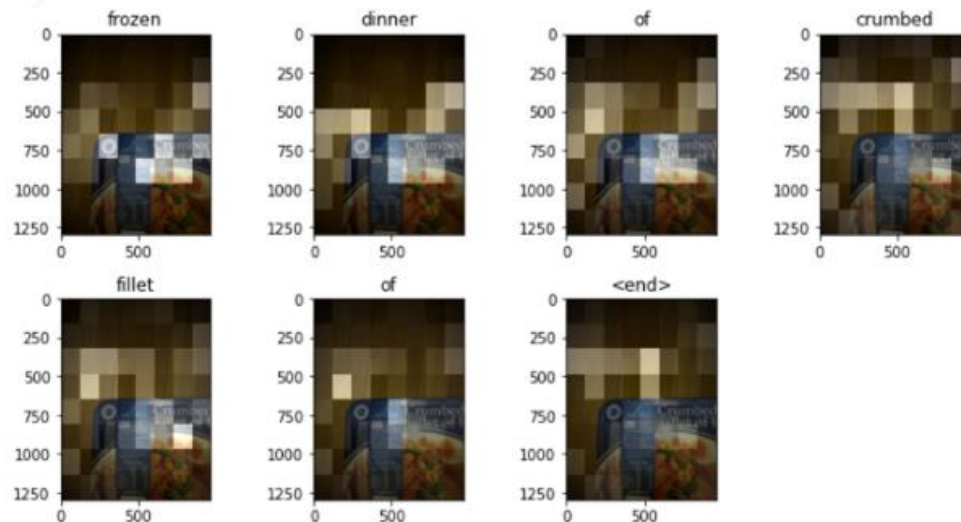
heat up food that is put in plastic blue

frozen food on top of wooden

frozen dinner of crumbed fillet of

blue package of microwave dinner with nutritional information on

Prediction Caption: frozen dinner of crumbed fillet of <end>

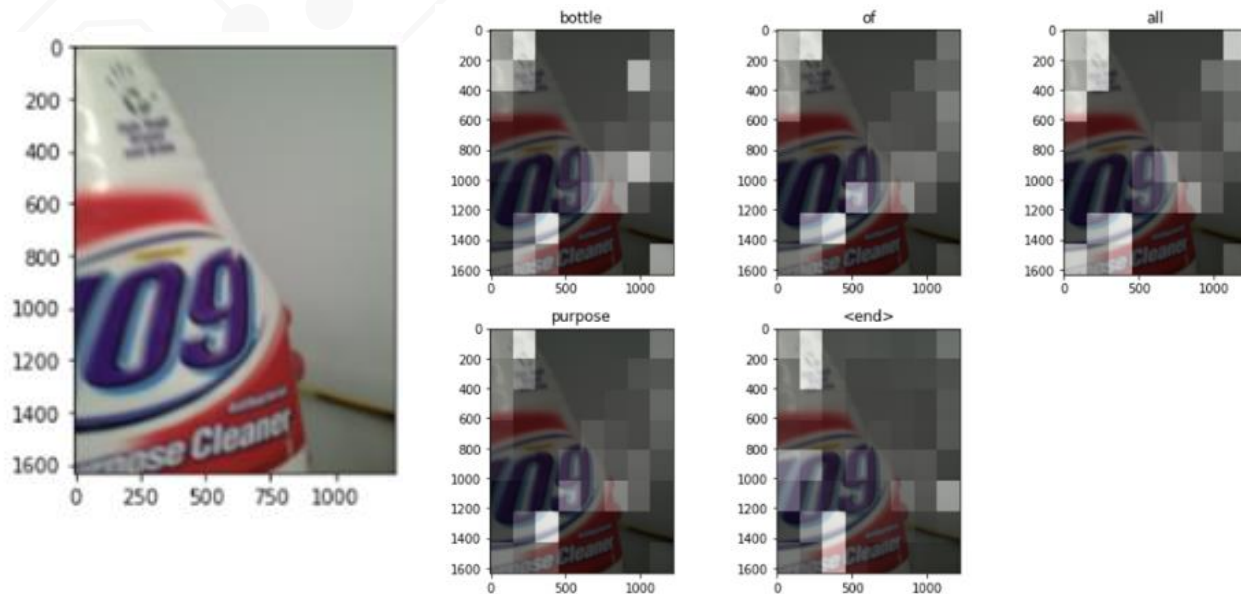


# Good Results

Real Captions:

white and red bottle of all purpose antibacterial  
the front of bottle of household  
person holding bottle of all purpose  
bottle of all purpose cleaner in front of white  
closeup of bottle of brand

Prediction Caption: bottle of all purpose <end>

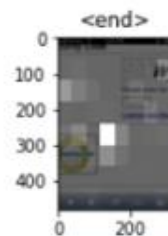
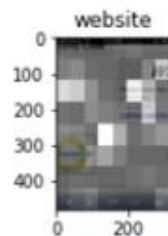
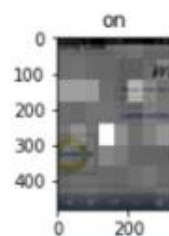
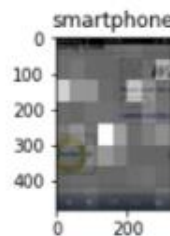
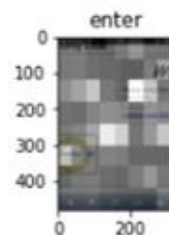
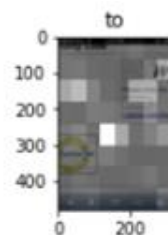
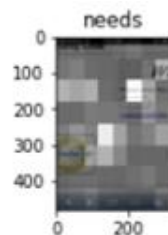
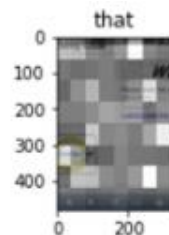
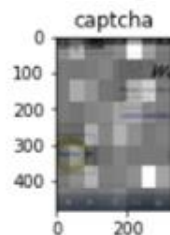
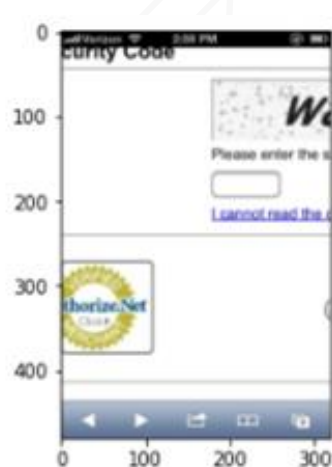


# Good Results

Real Captions:

section of smartphone screen showcasing verified merchant and captcha  
security code verification on mobile  
screenshot of smartphone display asking for security code  
part of computer screen that reads enter  
verizon phone screenshot on website commanding info

Prediction Caption: captcha that needs to enter smartphone on website <end>





# Close Results

Real Captions:

united states postal service box that is blue

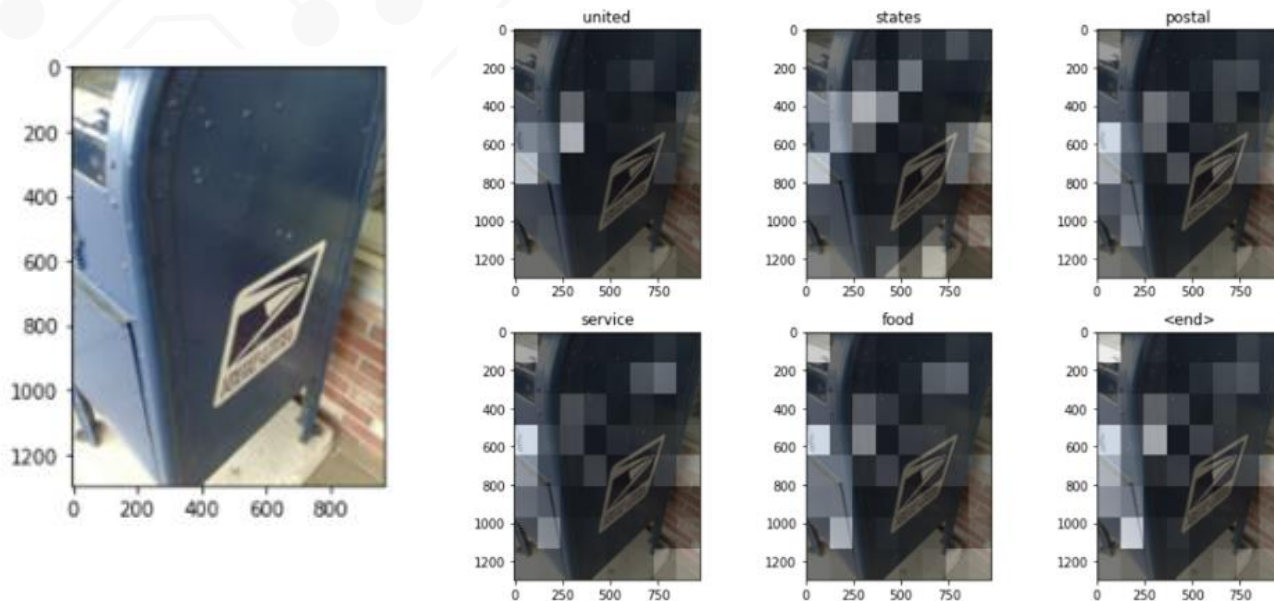
blue united states postal service mailbox in front of brick siding

blue metal united states postal service mail box attached to cement block with short brick wall behind

blue mail drop box with usps logo on

blue united states postal service mail box bolted to

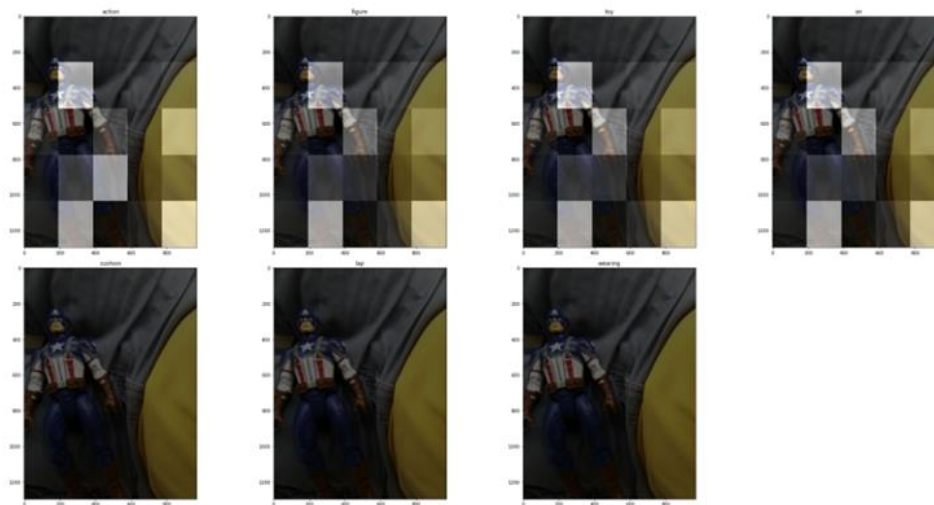
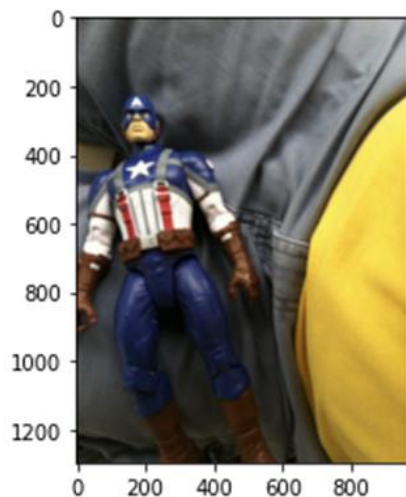
Prediction Caption: united states postal service food <end>



# Close Results

Real Captions:

action figure of captain america on lap  
captain america plastic kids action figure on mans lap  
toy figure of captain america lying in someone lap  
an image of laying toy figurine on person lap  
action figure wearing blue on top of lap wearing grey pants and yellow shirt  
Prediction Caption: action figure toy on cushion lap wearing jeans

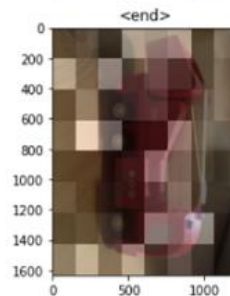
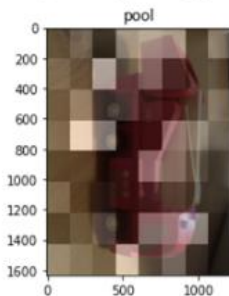
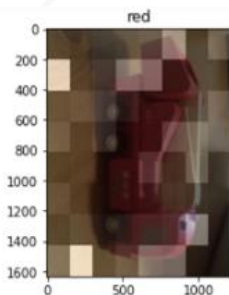


# Bad Results

Real Captions:

red toy vehicle placed on the floor  
kids wooden red truck with three black wheels on wooden  
red wooden toy fire truck with black  
red toy truck is on wooden  
red wooden fire truck toy with removable

Prediction Caption: red wooden wooden pool <end>




???





**05.**

 **Q & A**