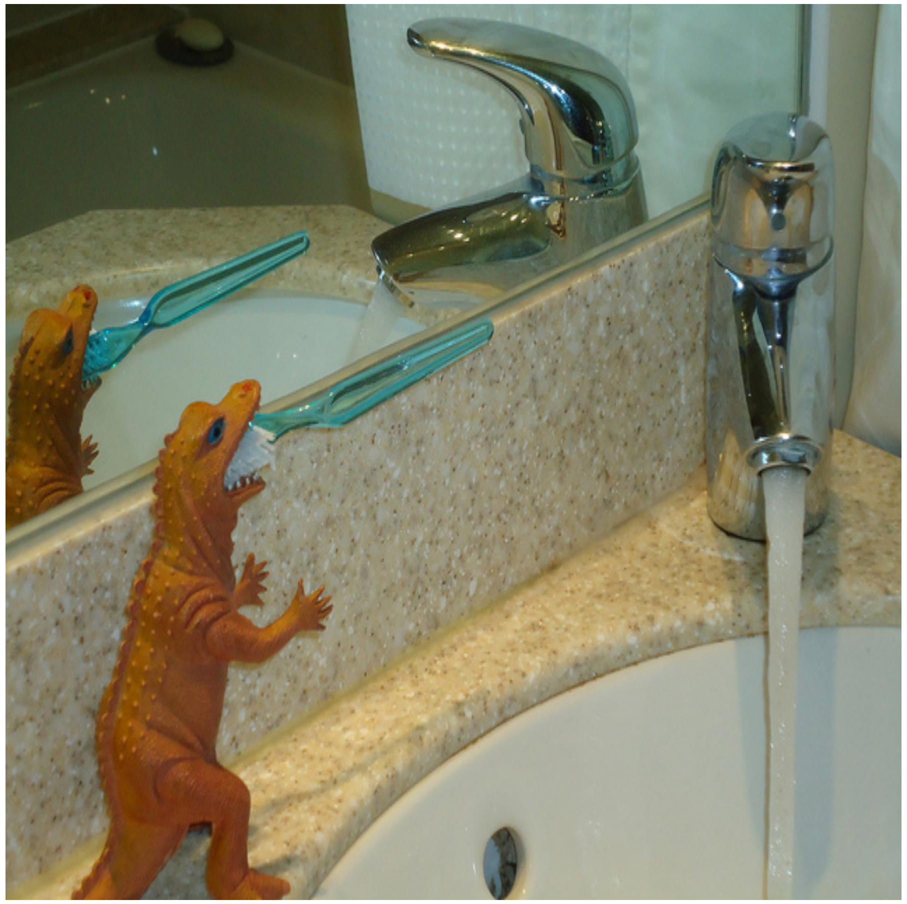
Image Captioning



→ A toy is standing on a sink.

Input Image

Output Caption

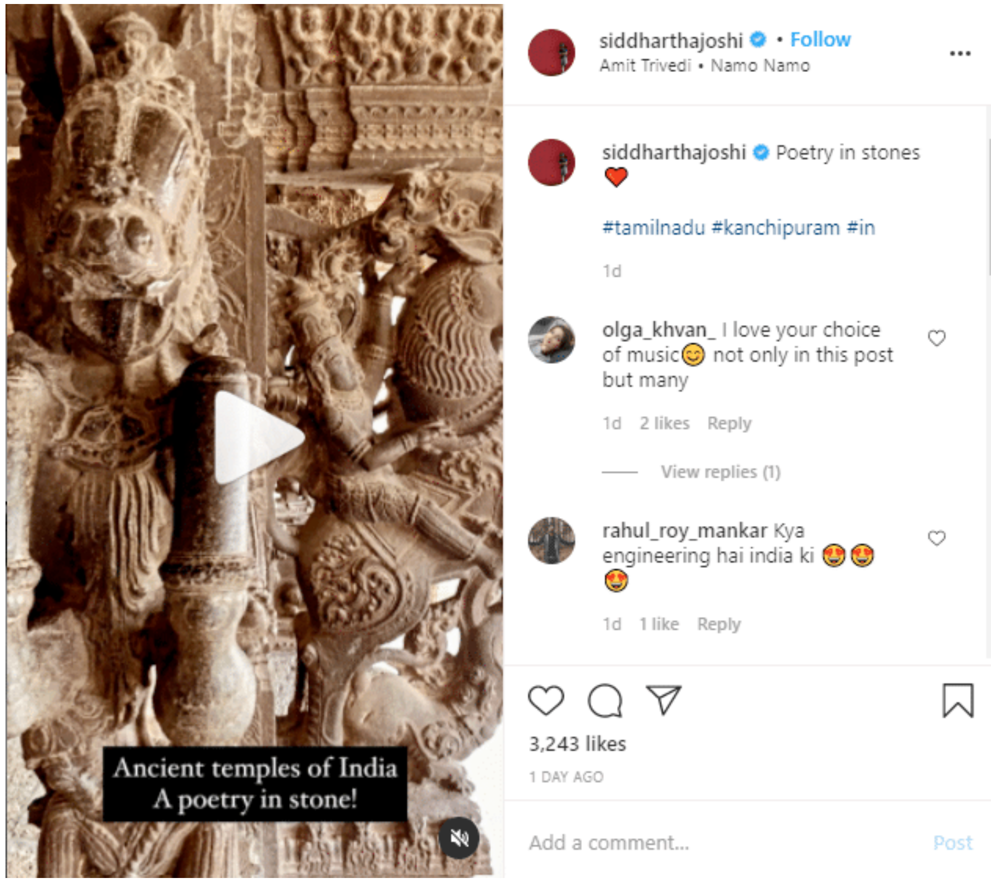
Application

Assistance for Visually Impaired

Input: GoPro images

Output: There is a rock by your side.

Advertisement bots



siddharthajoshi • Follow
Amit Trivedi • Namo Namo

siddharthajoshi Poetry in stones ❤️

#tamilnadu #kanchipuram #in

1d

olga_khvan_ I love your choice of music😊 not only in this post but many

1d 2 likes Reply

View replies (1)

rahul_roy_mankar Kya engineering hai india ki 😍😍😍

1d 1 like Reply

3,243 likes

1 DAY AGO

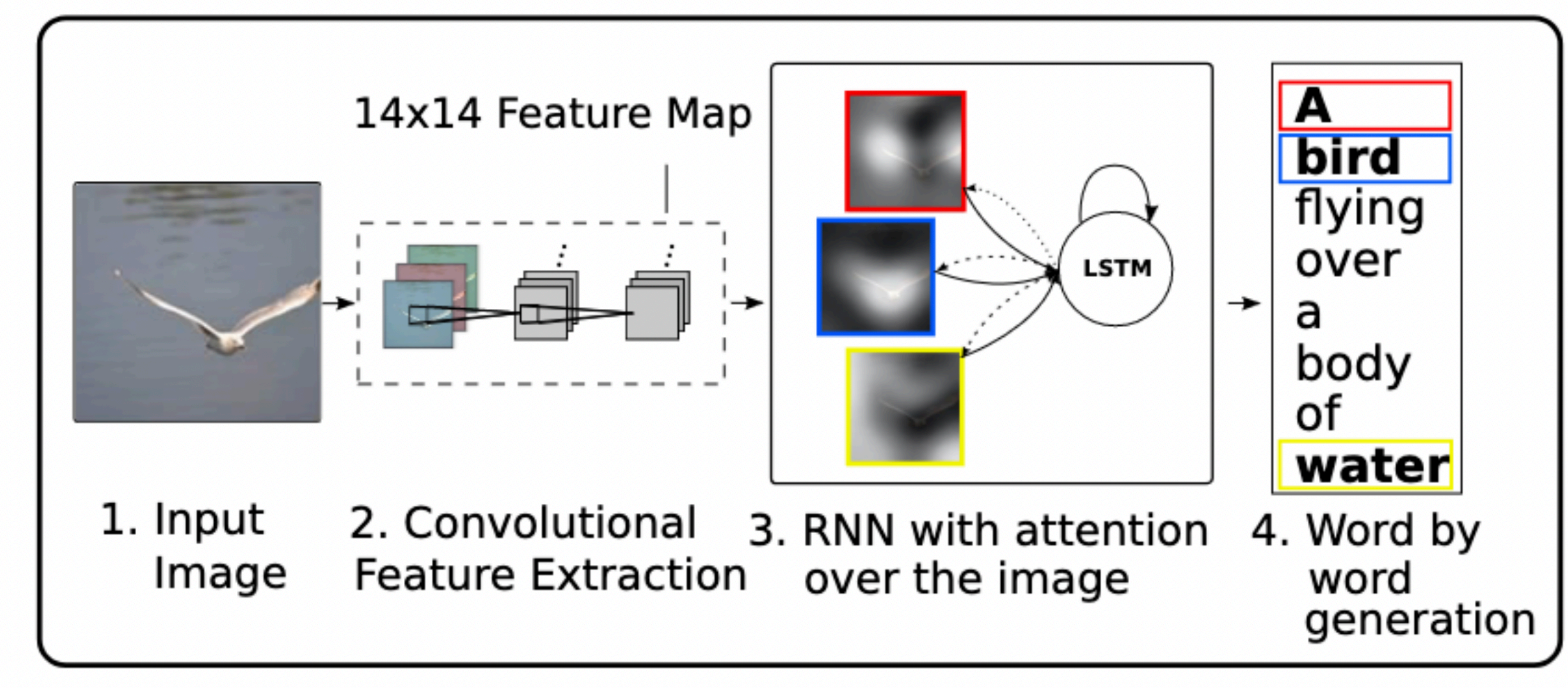
Add a comment... Post

Ancient temples of India
A poetry in stone!

Input: Selfie

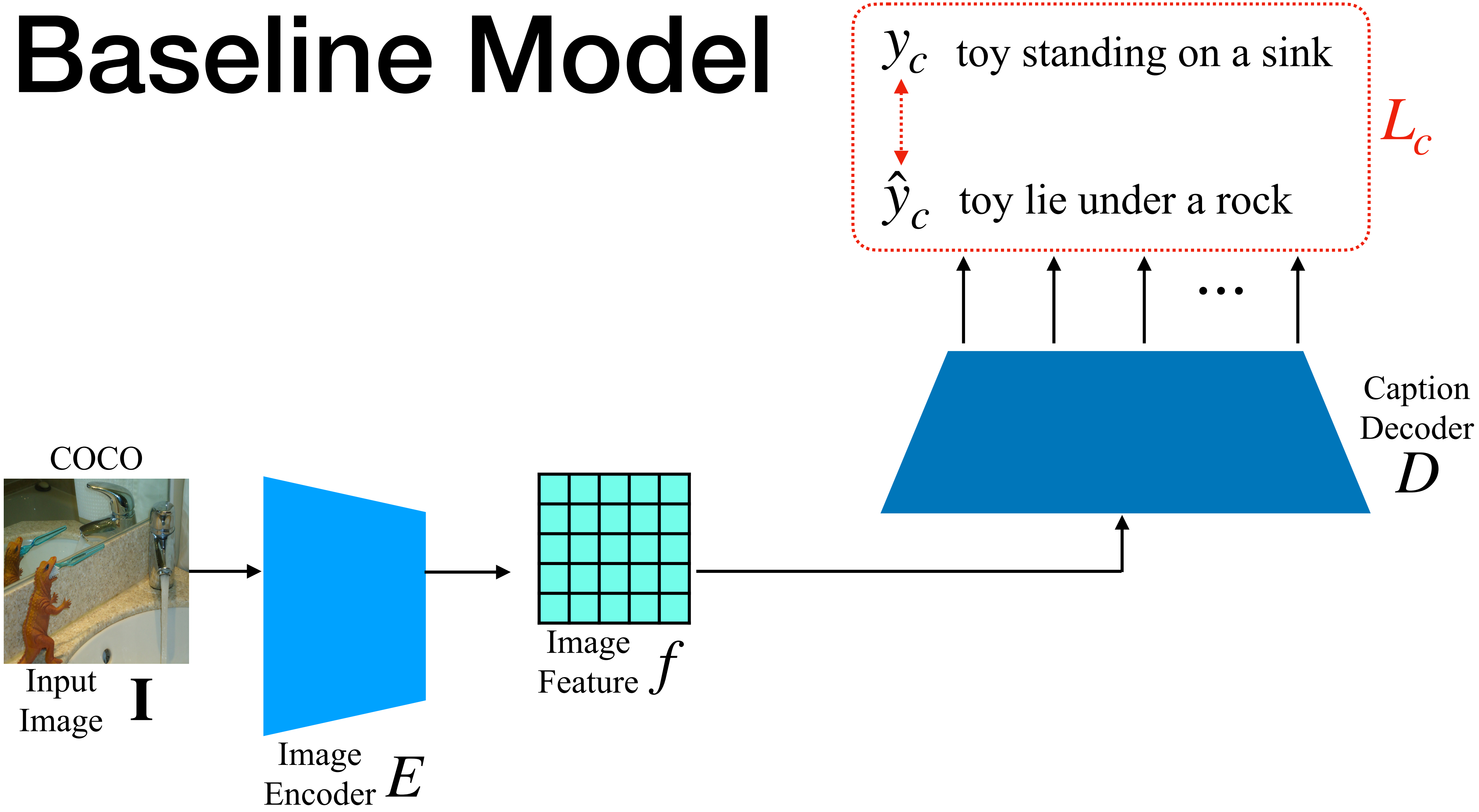
Output: Lots of comments.

Baseline: Show, tell, and attend


14x14 Feature Map
LSTM
A
bird
flying
over
a
body
of
water
1. Input Image
2. Convolutional Feature Extraction
3. RNN with attention over the image
4. Word by word generation

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

Baseline Model



y_c toy standing on a sink

$\hat{y}_c$ toy lie under a rock

L_c

…

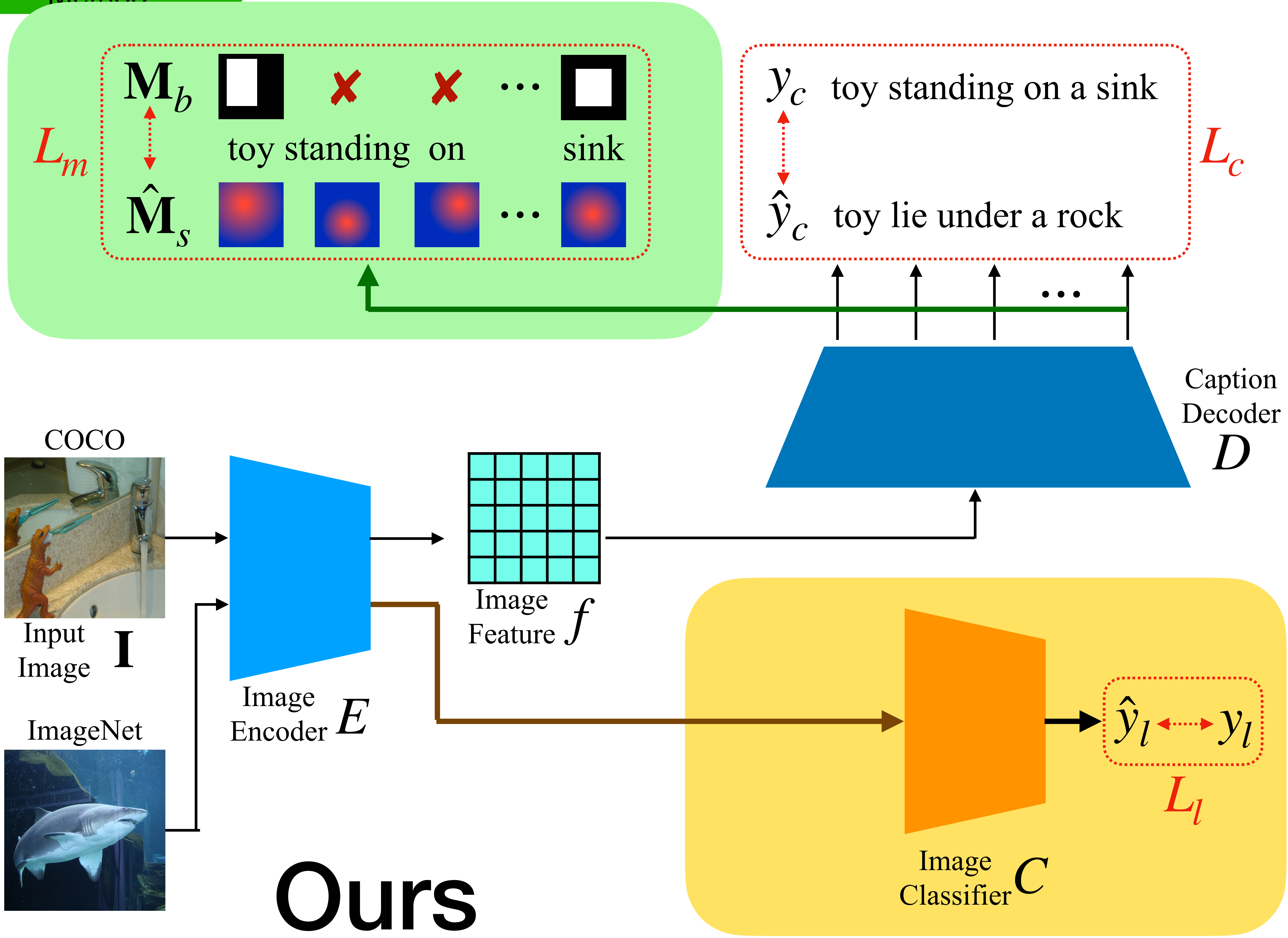
Caption Decoder D

COCO

Input Image $\mathbf{I}$

Image Encoder E

Image Feature f


L_m
$\mathbf{M}_b$
toy standing on
sink
$\hat{\mathbf{M}}_s$
y_c toy standing on a sink
$\hat{y}_c$ toy lie under a rock
L_c
Caption Decoder D
COCO
Input Image $\mathbf{I}$
ImageNet
Image Encoder E
Image Feature f
Image Classifier C
$\hat{y}_l$
y_l
L_l

Ours

Results

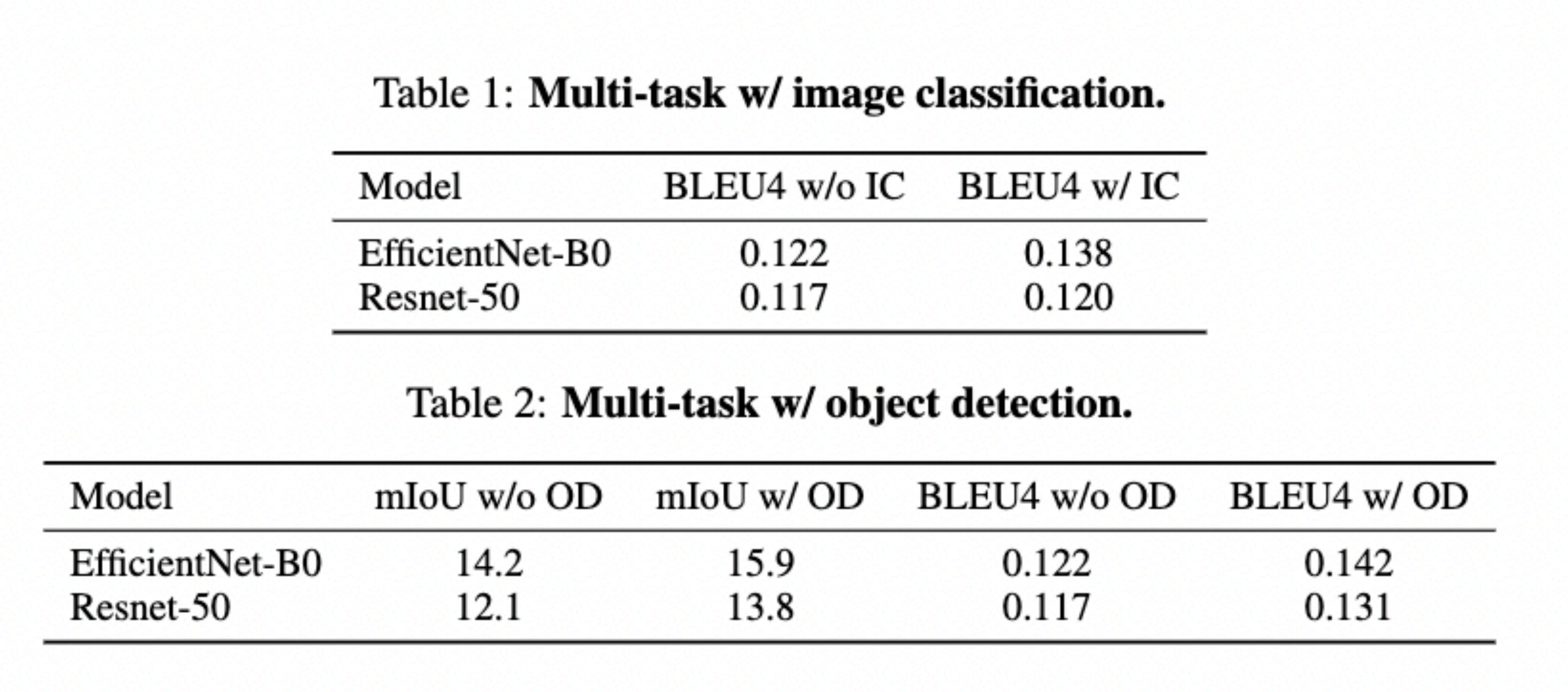


Table 1: **Multi-task w/ image classification.**

Model	BLEU4 w/o IC	BLEU4 w/ IC
EfficientNet-B0	0.122	0.138
Resnet-50	0.117	0.120

Table 2: **Multi-task w/ object detection.**

Model	mIoU w/o OD	mIoU w/ OD	BLEU4 w/o OD	BLEU4 w/ OD
EfficientNet-B0	14.2	15.9	0.122	0.142
Resnet-50	12.1	13.8	0.117	0.131

- The experiments (1) strength our claim that more visual information results in better text understanding in image captioning and (2) shows the importance of localizing objects in images on language captioning.

Discussion

- What we've learned in this project?
  - basic approach to image captioning
  - a basis to explore other cross-modal (visual and language) learning tasks such as text2image generation and VQA
  - In cross-modal learning tasks, it is important to design a model that is able to connect the feature between images and texts
- What we wish to accomplish in the future?
  - Besides image classification and object detection, there are more visual tasks that we can think about using using visual data, image generation for example. A possible idea is to design a bi-directional generation model with cycle consistency.
  - There are also more powerful language models that we can experiment with nowadays. We wish to find out how the learned features in ChatGPT interact with visual data.