MACHINE LEARNING
PETNICA SUMMER INSTITUTE

Image captioning

Srđan Jovanović Radenko Pavlović

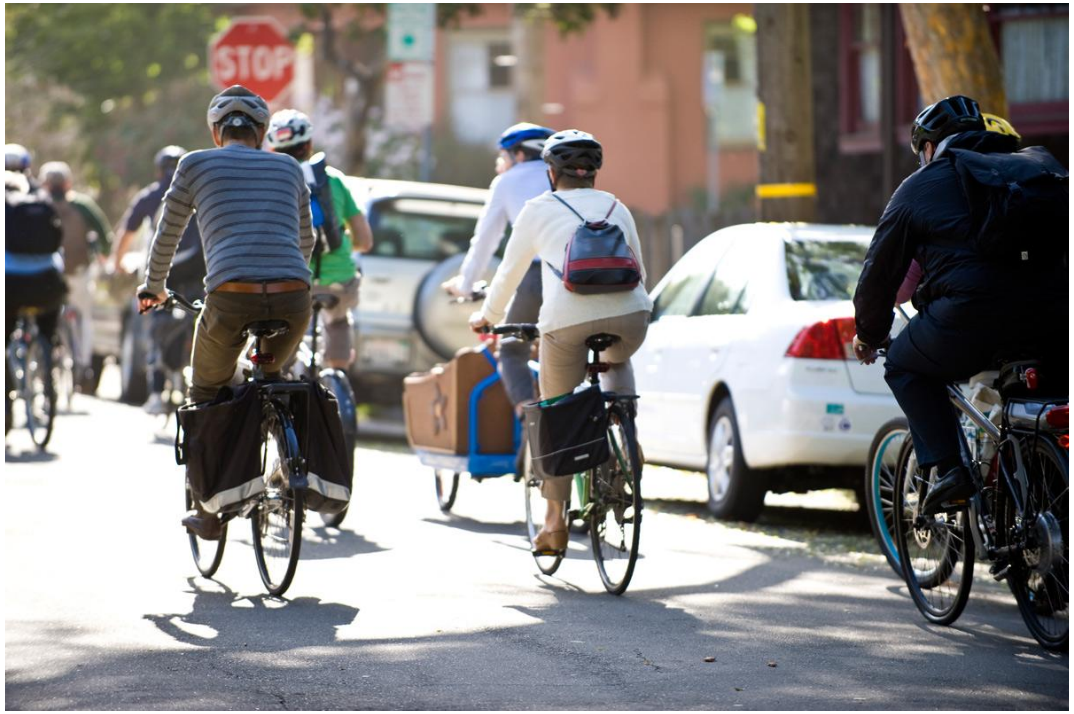
Mentors: Saša Galić, Momčilo Vasilijević

Microsoft | Development Center Serbia

Petnica

Problem to solve

Problem to solve



People ride bicycles on the street.

People with helmets ride bicycles.

A group of people ride bicycles.

Questions to be answered

Questions to be answered

1. What kind of architecture we need?

Questions to be answered

1. What kind of architecture we need?

2. **How accurate it can be?**

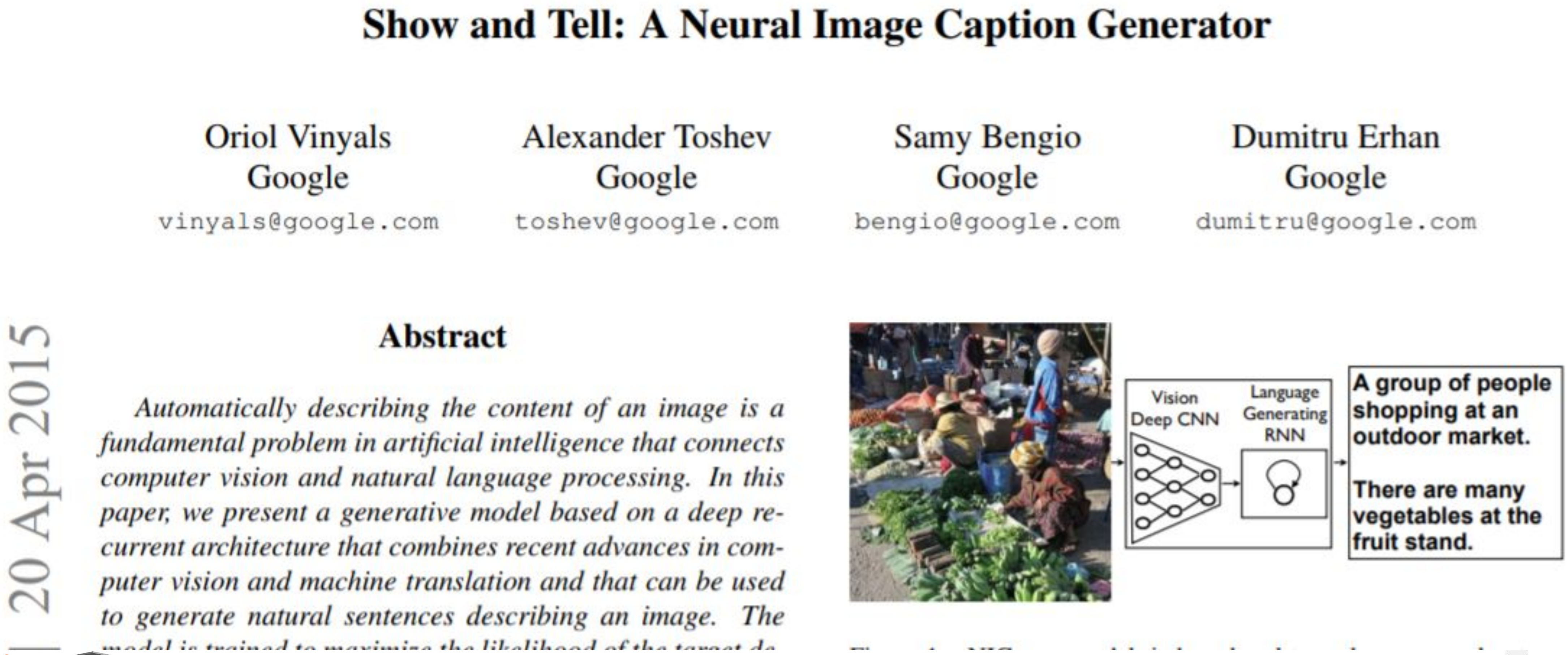
Questions to be answered

1. What kind of architecture we need?

2. How accurate it can be?

3. **Which parts of the image affect which word in the output?**

Solution



Show and Tell: A Neural Image Caption Generator

Oriol Vinyals
Google
vinyals@google.com

Alexander Toshev
Google
toshev@google.com

Samy Bengio
Google
bengio@google.com

Dumitru Erhan
Google
dumitru@google.com

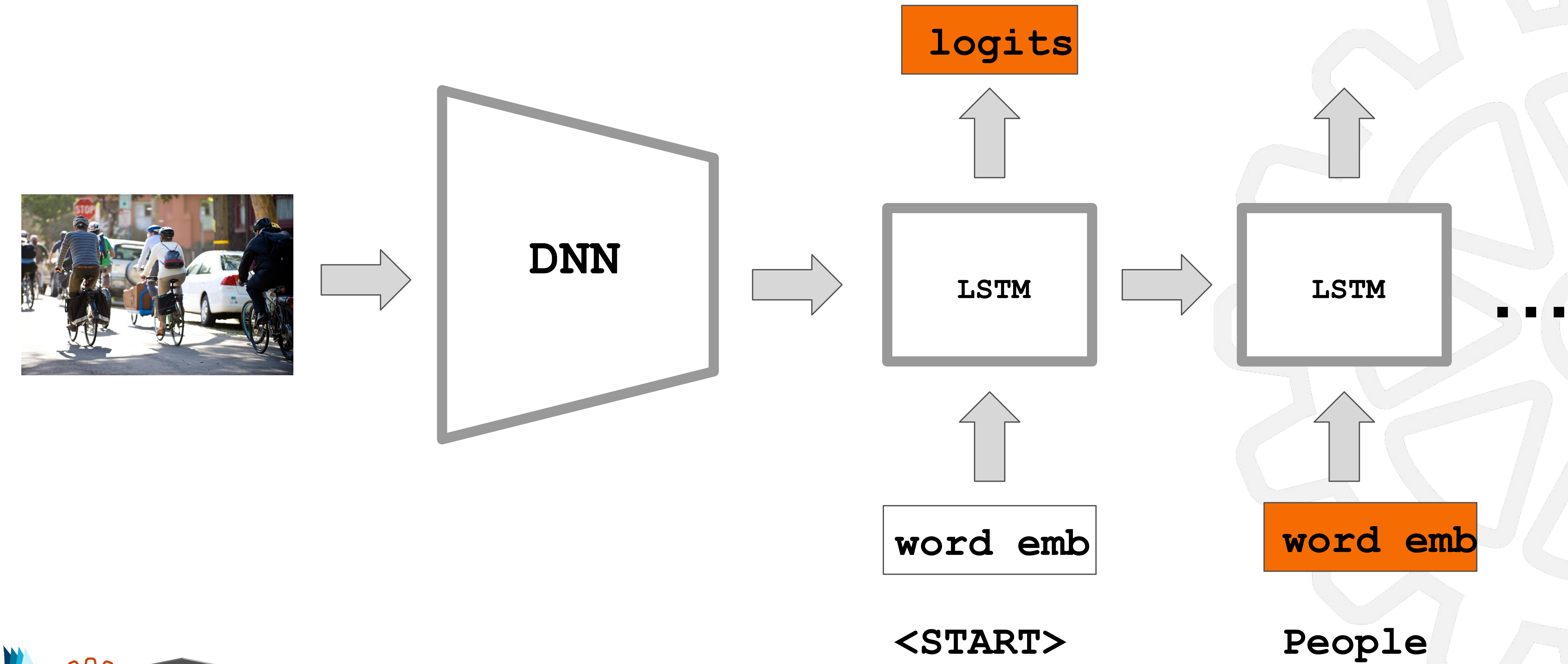
20 Apr 2015

Abstract

Automatically describing the content of an image is a fundamental problem in artificial intelligence that connects computer vision and natural language processing. In this paper, we present a generative model based on a deep recurrent architecture that combines recent advances in computer vision and machine translation and that can be used to generate natural sentences describing an image. The model is trained to maximize the likelihood of the target de-

Vision
Deep CNN
Language
Generating
RNN
A group of people shopping at an outdoor market.
There are many vegetables at the fruit stand.

Architecture



DNN

logits

LSTM

LSTM

...

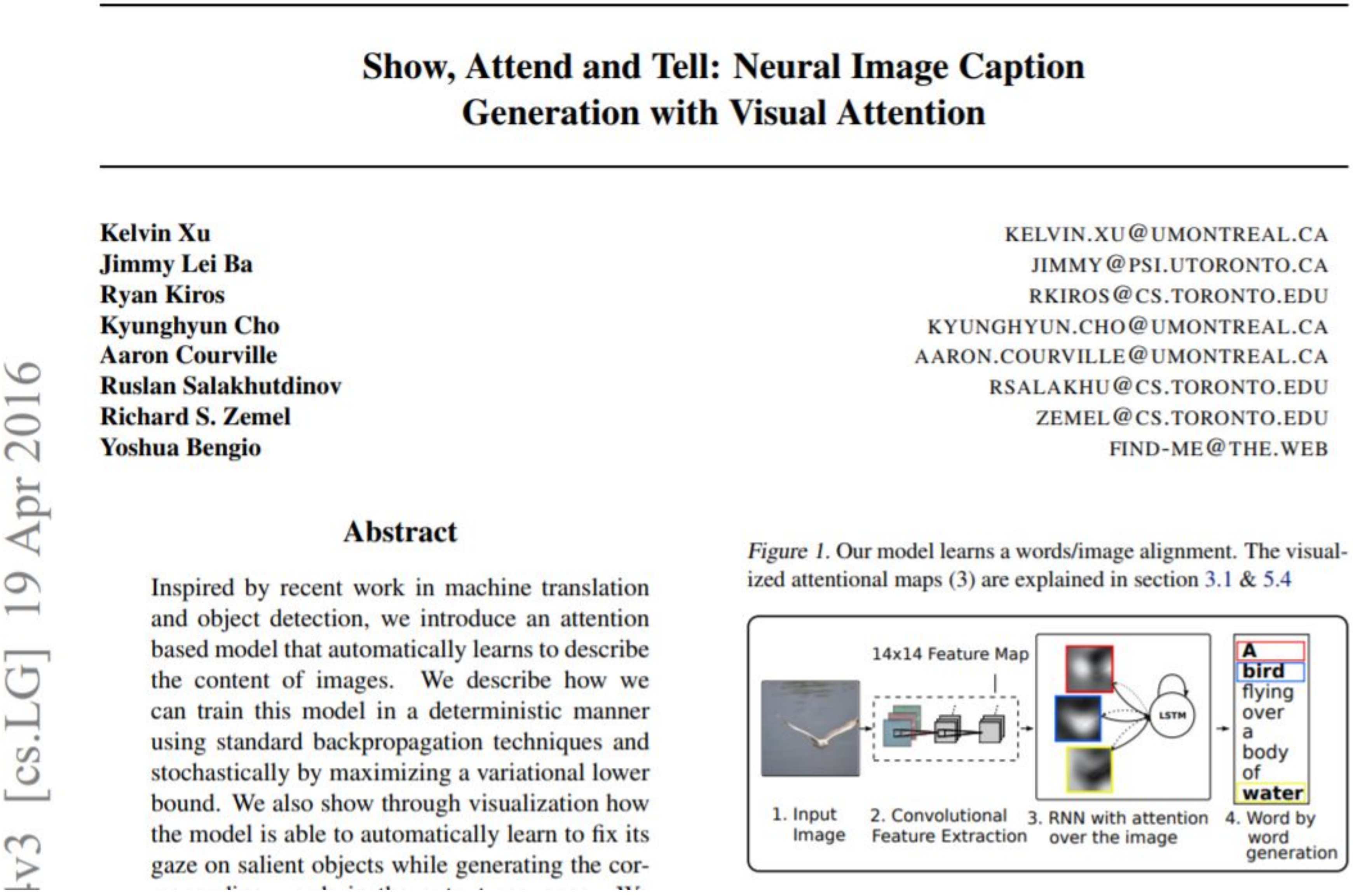
word emb

word emb

<START>

People

Attention model



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

Kelvin Xu KELVIN.XU@UMONTREAL.CA
Jimmy Lei Ba JIMMY@PSI.UTORONTO.CA
Ryan Kiros RKIROS@CS.TORONTO.EDU
Kyunghyun Cho KYUNGHYUN.CHO@UMONTREAL.CA
Aaron Courville AARON.COURVILLE@UMONTREAL.CA
Ruslan Salakhutdinov RSALAKHU@CS.TORONTO.EDU
Richard S. Zemel ZEMEL@CS.TORONTO.EDU
Yoshua Bengio FIND-ME@THE.WEB

v3 [cs.LG] 19 Apr 2016

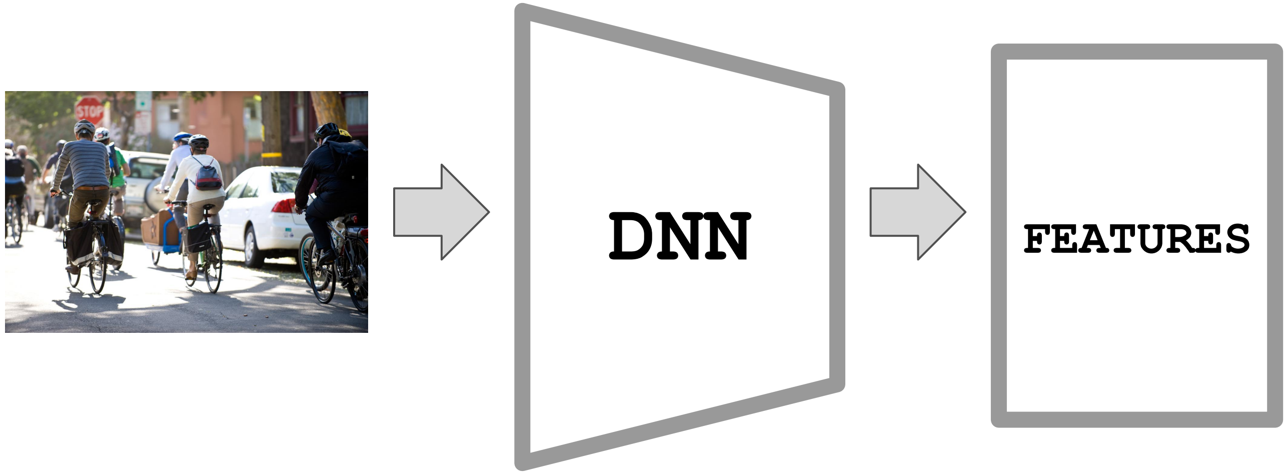
Abstract

Inspired by recent work in machine translation and object detection, we introduce an attention based model that automatically learns to describe the content of images. We describe how we can train this model in a deterministic manner using standard backpropagation techniques and stochastically by maximizing a variational lower bound. We also show through visualization how the model is able to automatically learn to fix its gaze on salient objects while generating the cor-

Figure 1. Our model learns a words/image alignment. The visualized attentional maps (3) are explained in section 3.1 & 5.4

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

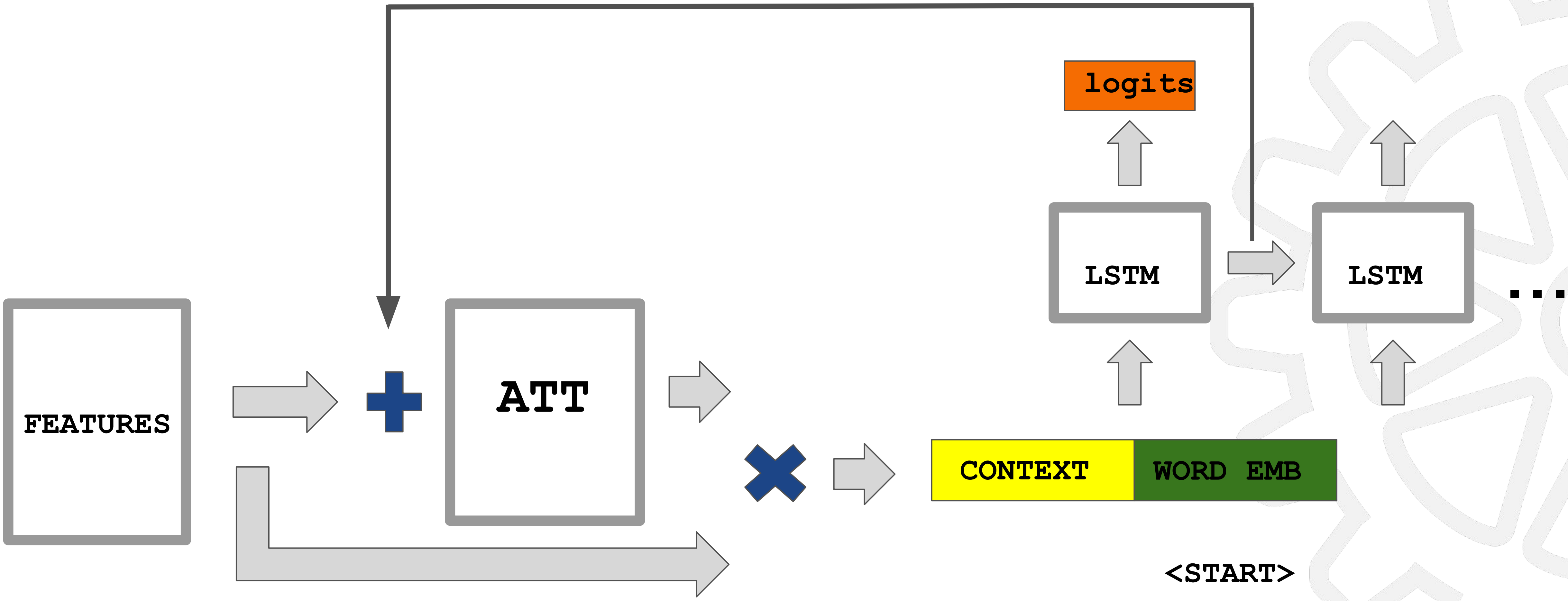
Architecture with attention



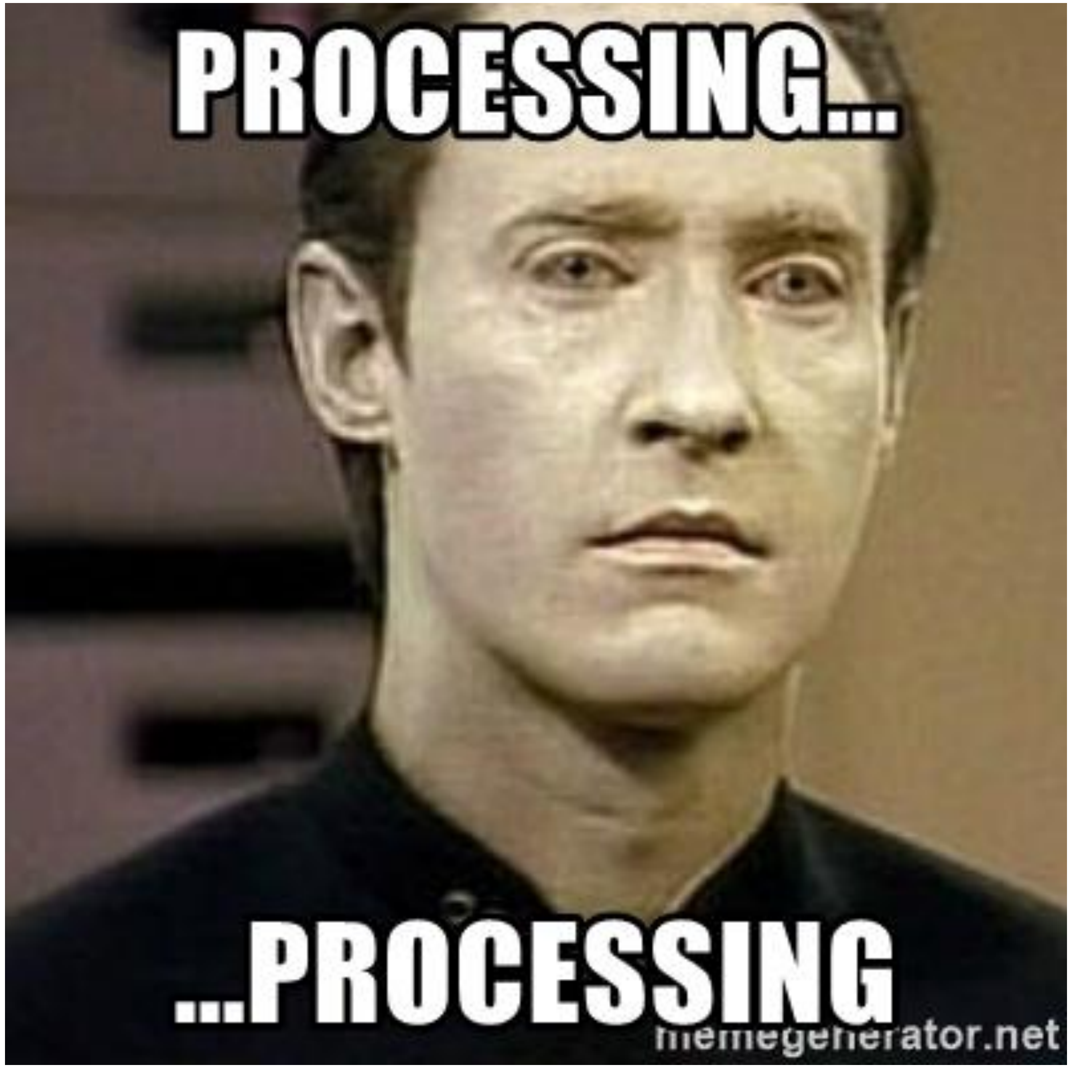
DNN

FEATURES

Architecture with attention

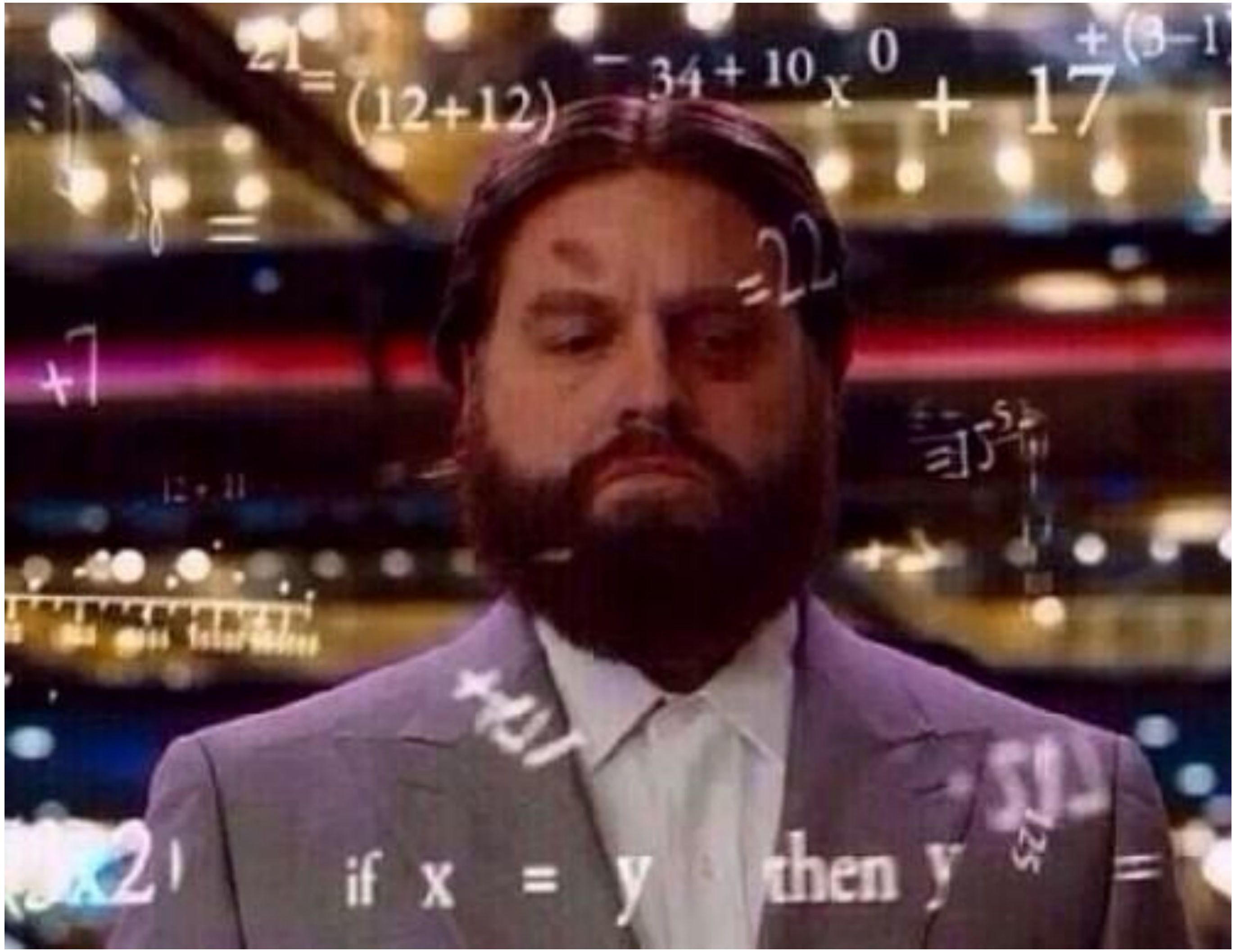


Struggles - processing dataset

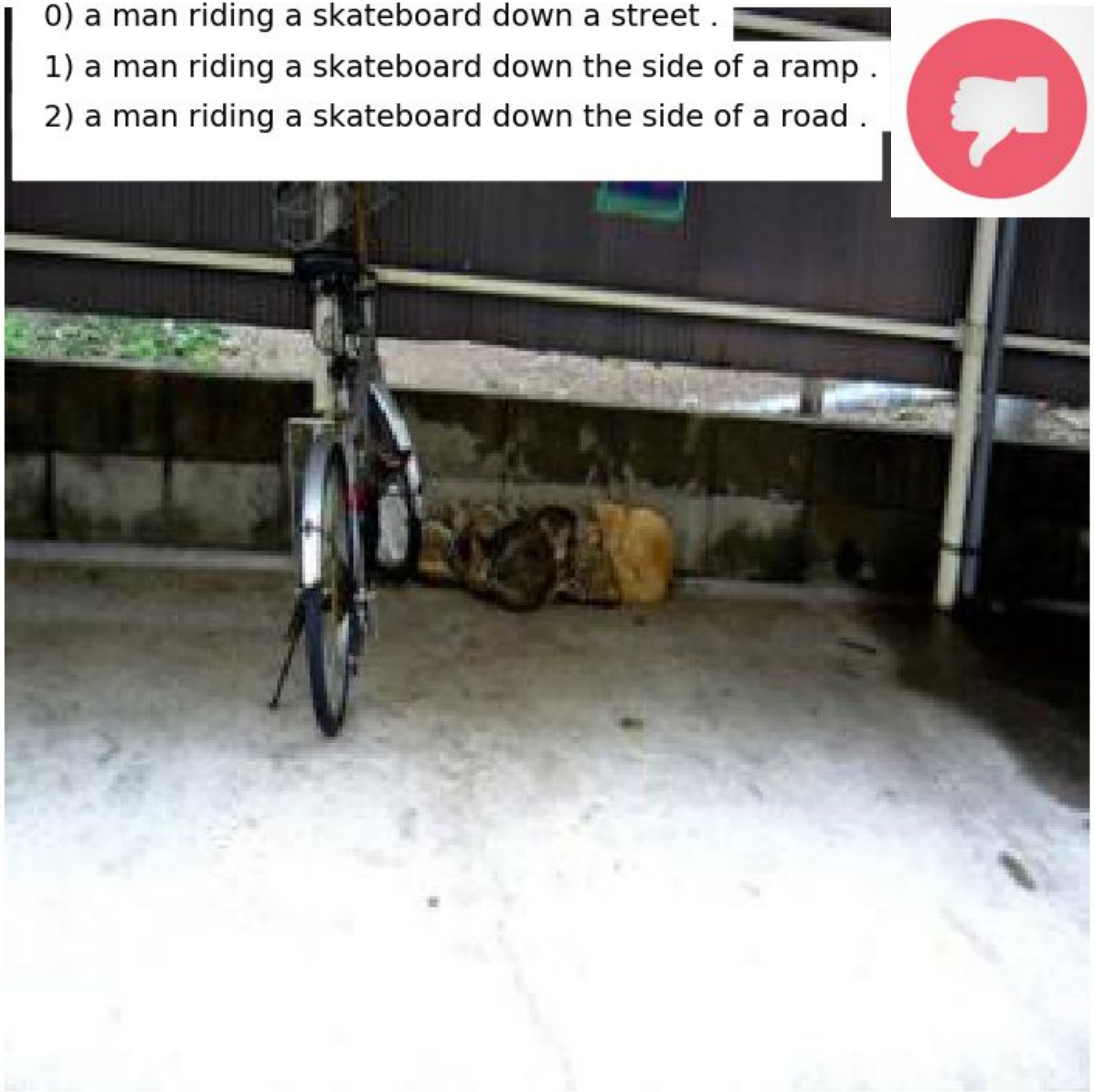


Struggles - getting things to work in tensorflow

Struggles - understanding model complexity



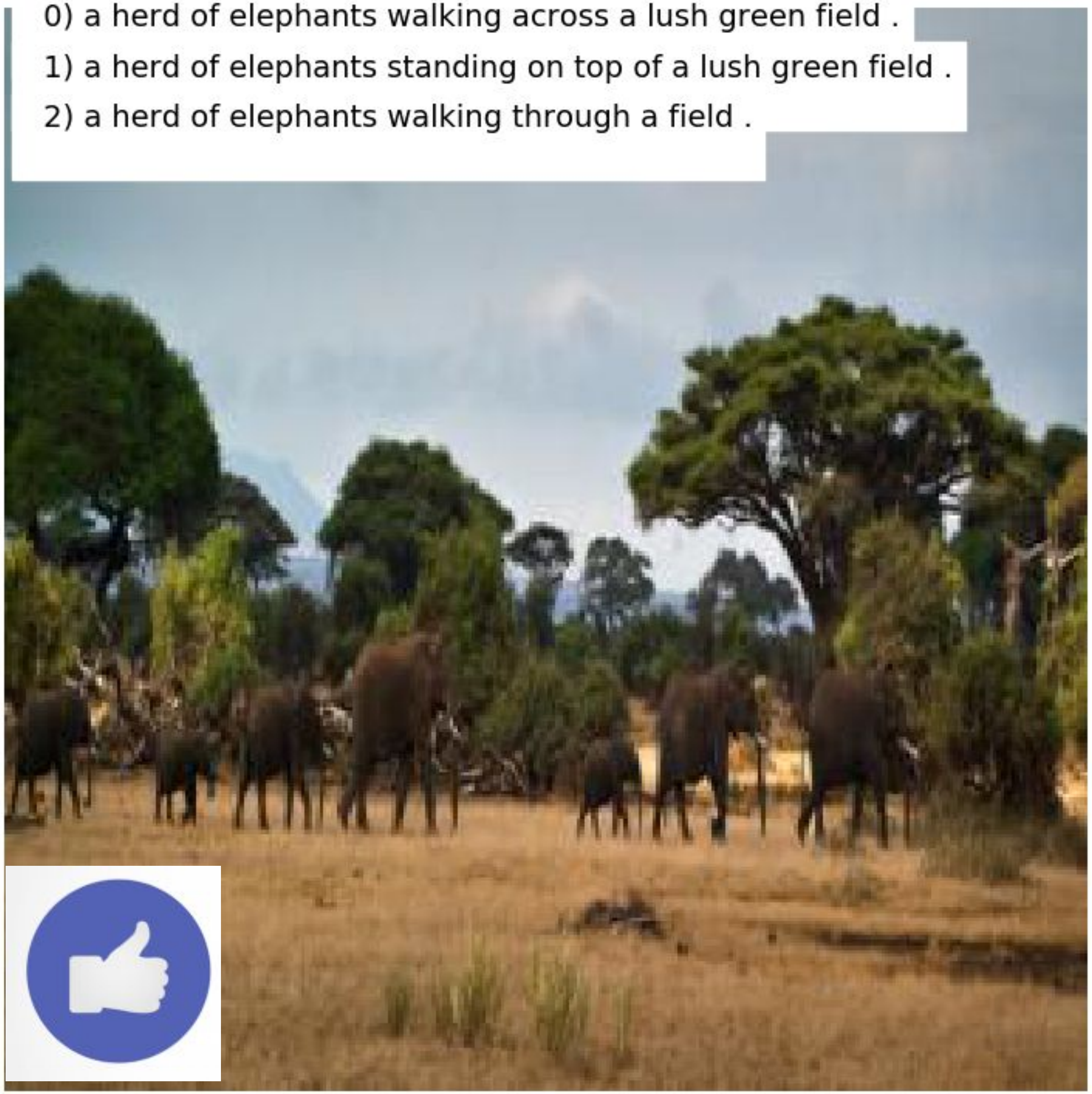
Results



0) a man riding a skateboard down a street .

1) a man riding a skateboard down the side of a ramp .

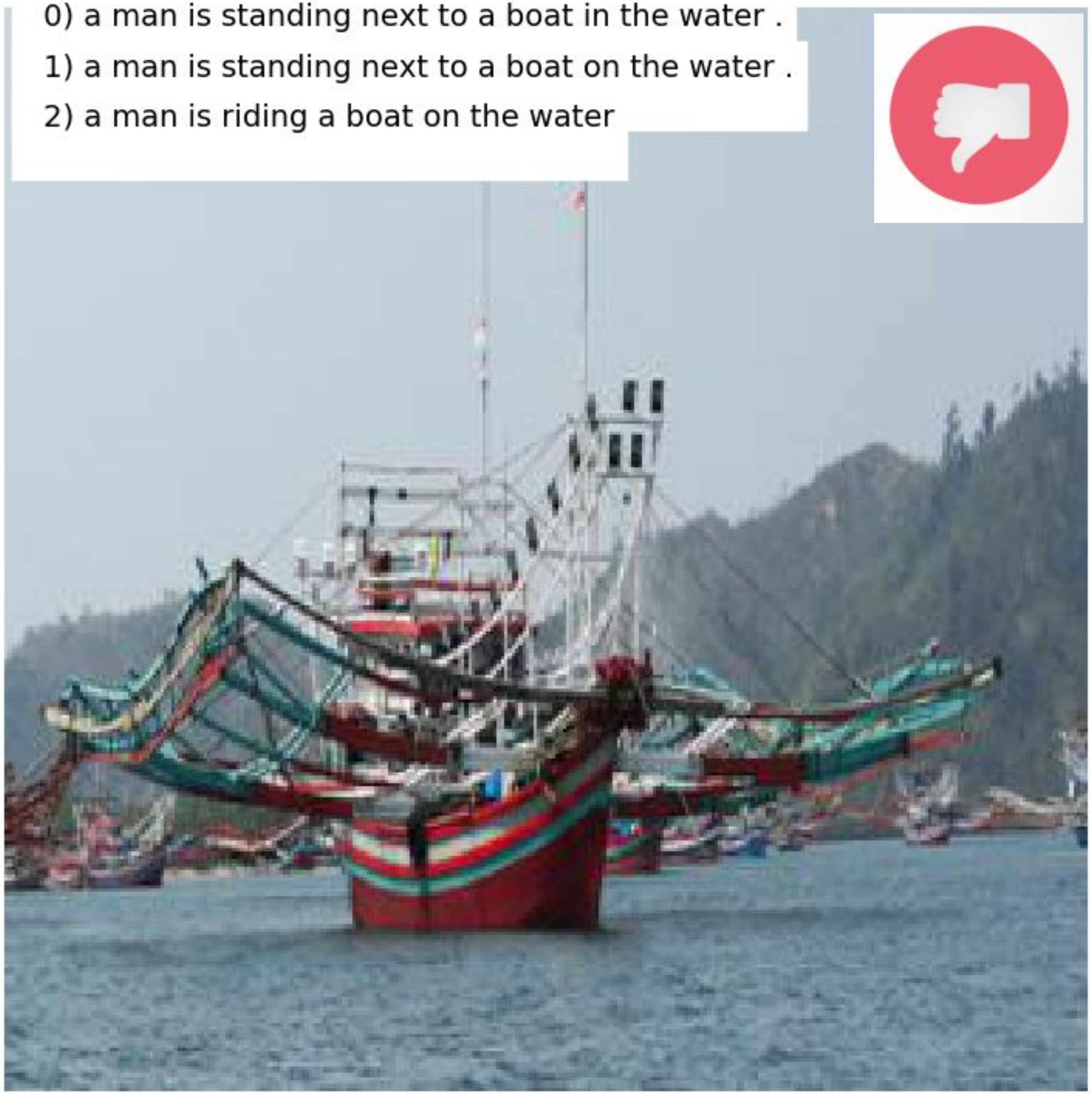
2) a man riding a skateboard down the side of a road .



0) a herd of elephants walking across a lush green field .

1) a herd of elephants standing on top of a lush green field .

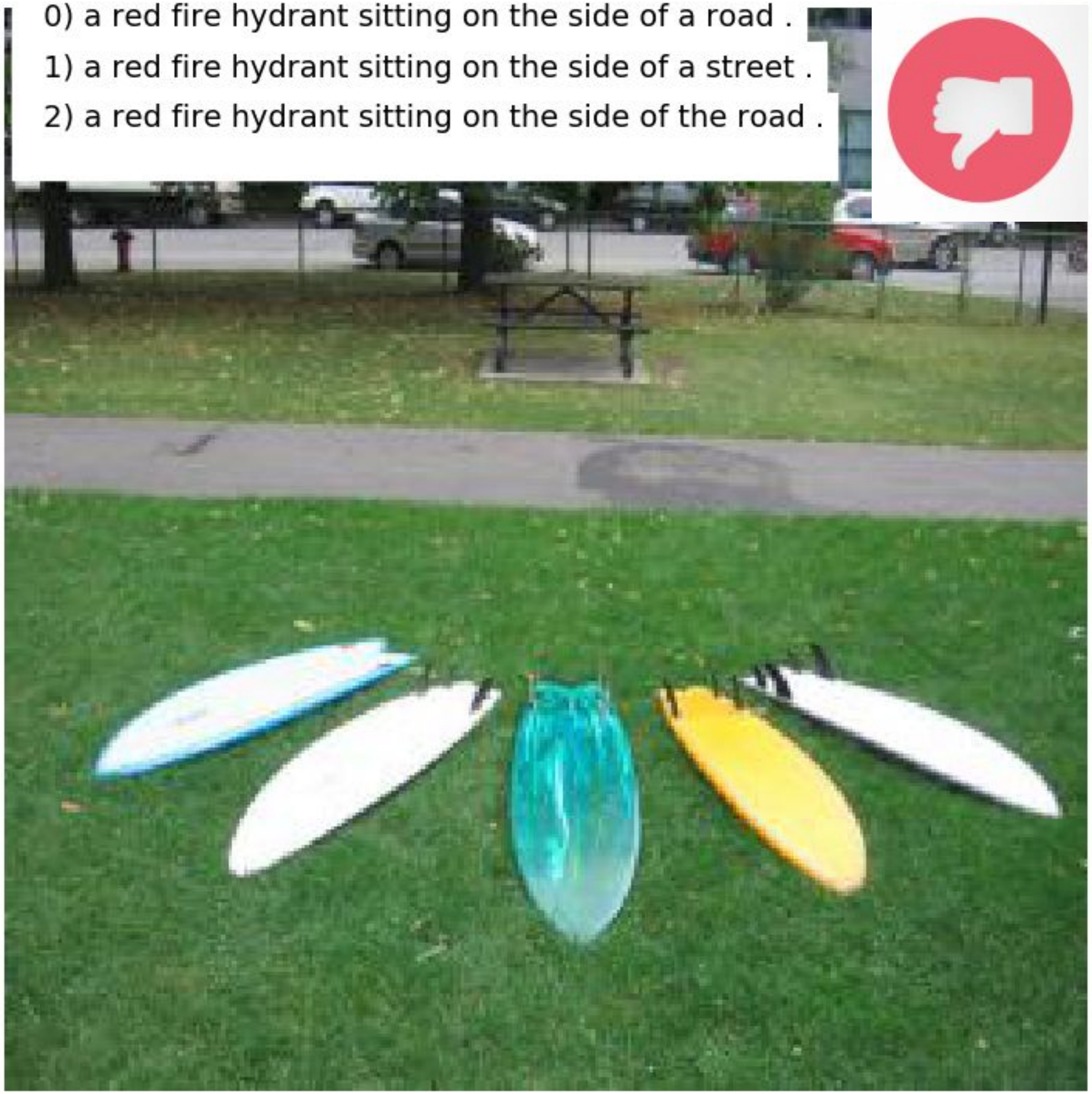
2) a herd of elephants walking through a field .



0) a man is standing next to a boat in the water .

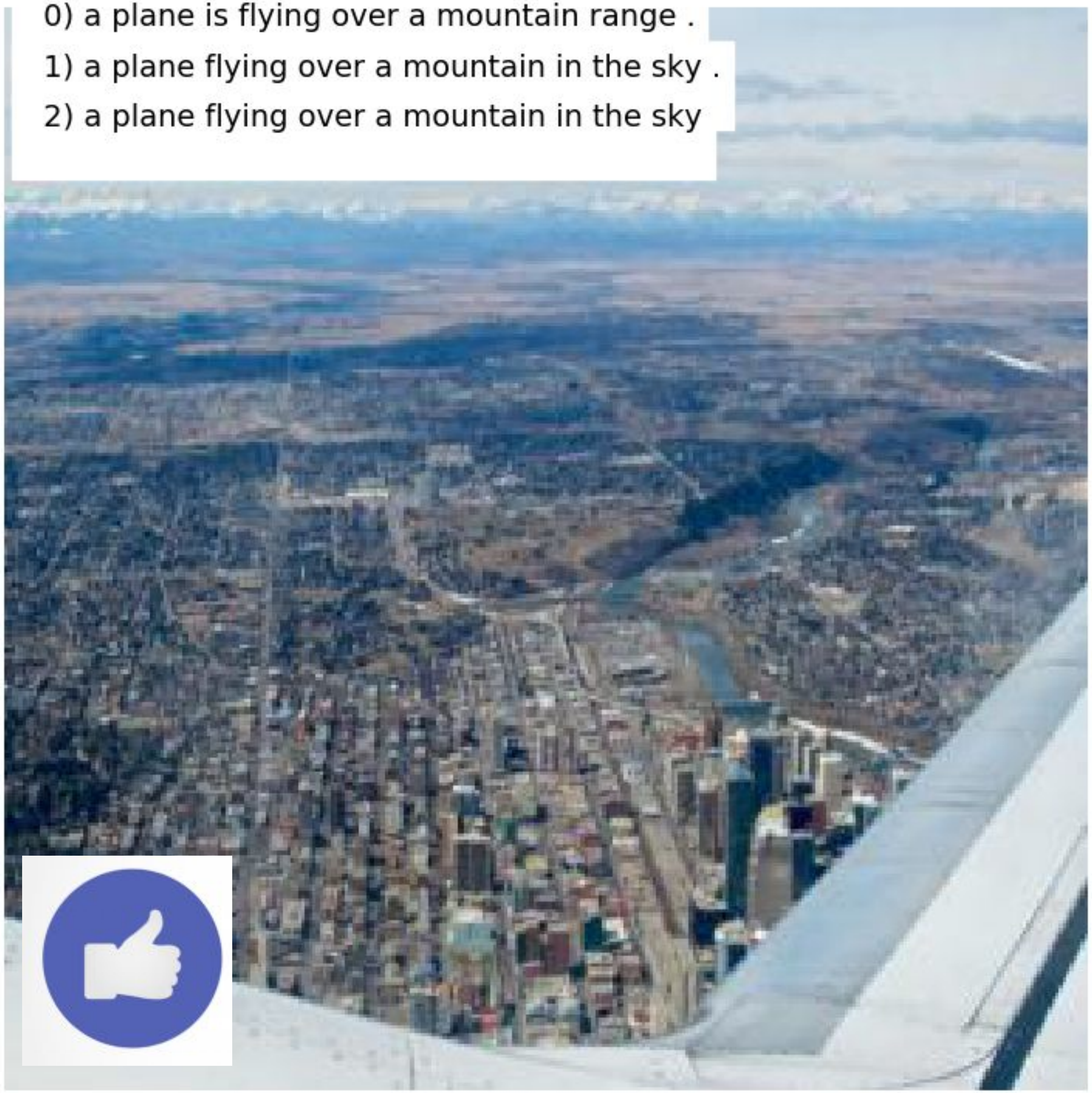
1) a man is standing next to a boat on the water .

2) a man is riding a boat on the water



0) a red fire hydrant sitting on the side of a road .
1) a red fire hydrant sitting on the side of a street .
2) a red fire hydrant sitting on the side of the road .

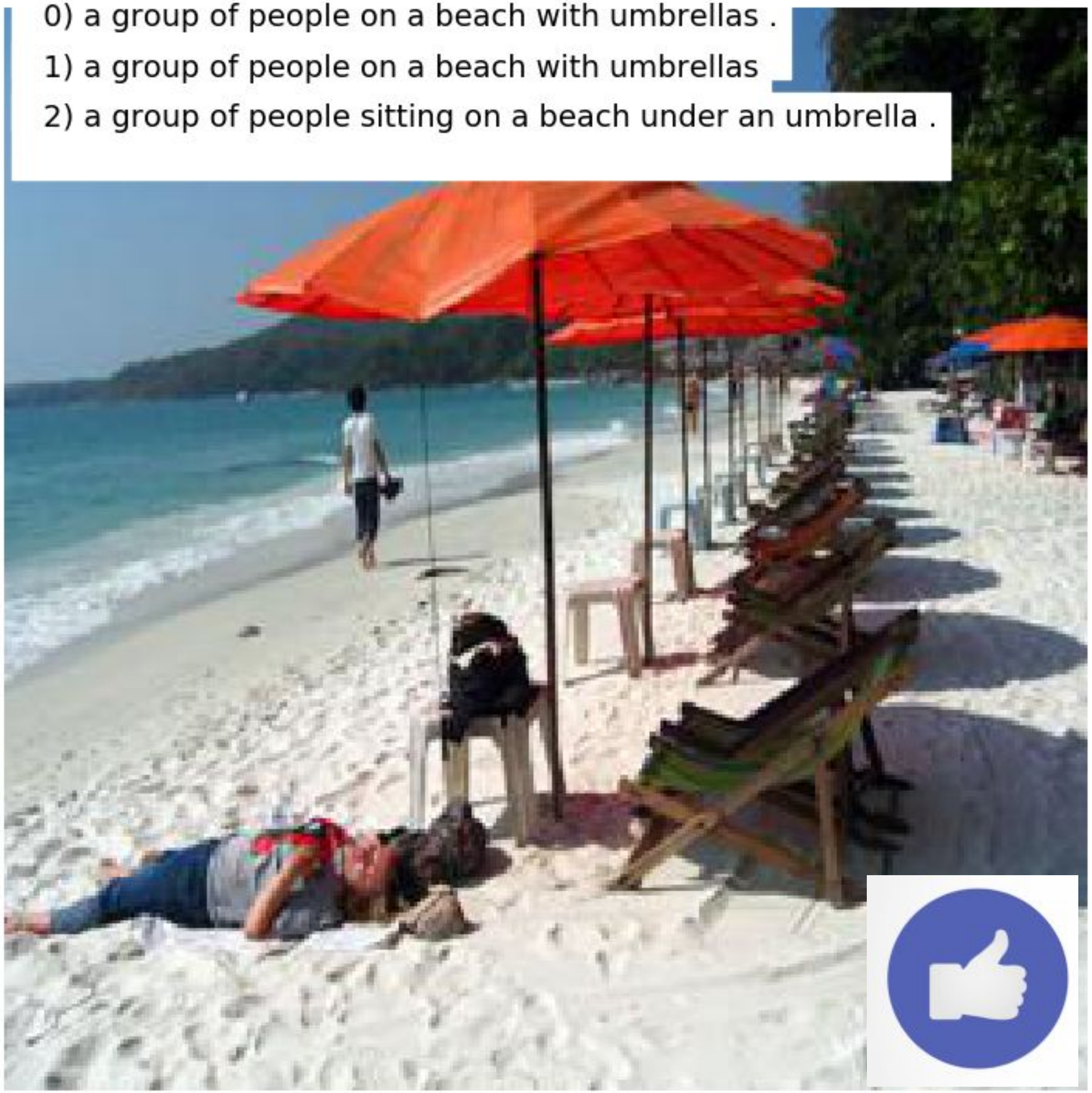
👎



0) a plane is flying over a mountain range .

1) a plane flying over a mountain in the sky .

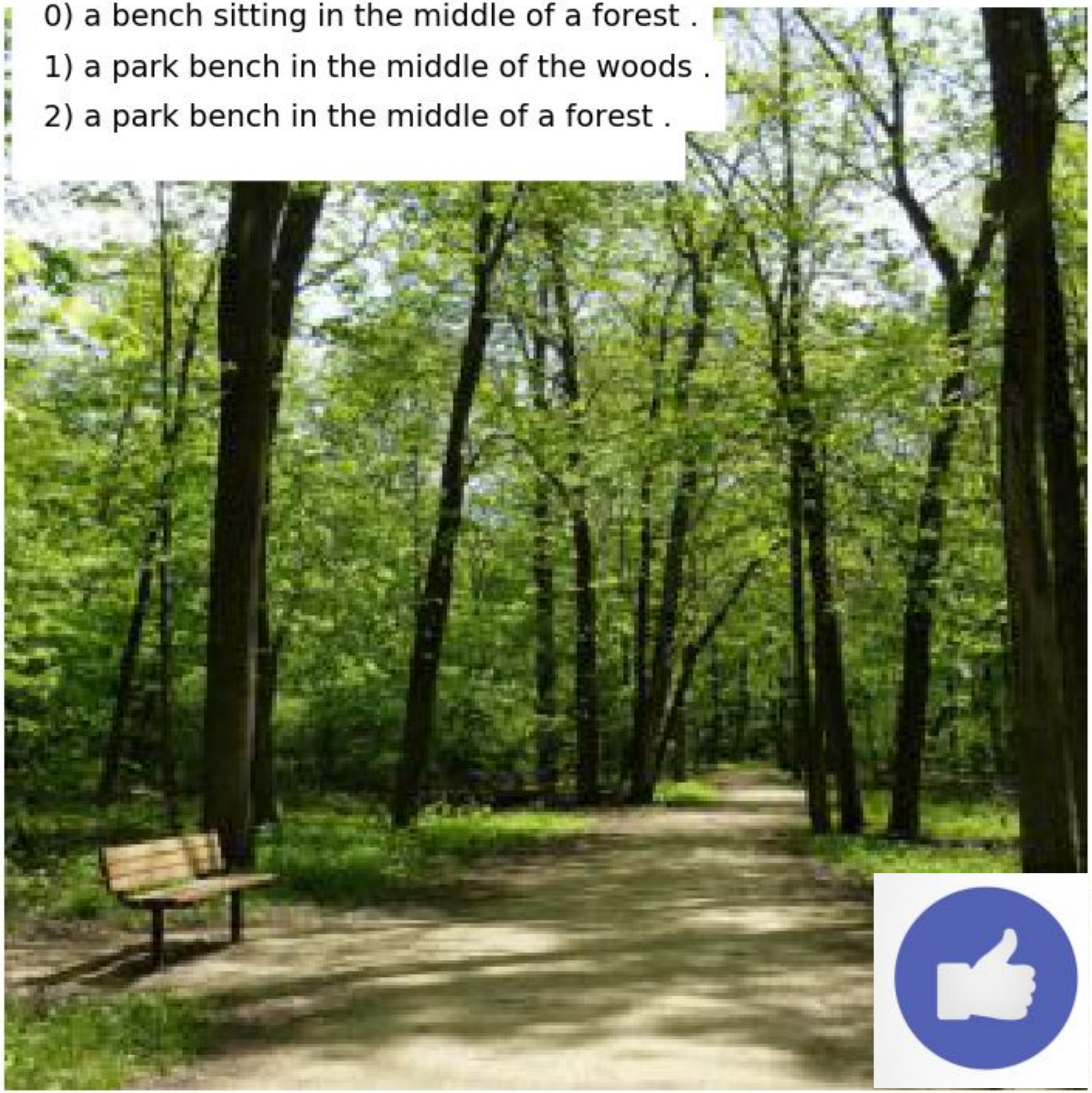
2) a plane flying over a mountain in the sky



0) a group of people on a beach with umbrellas .

1) a group of people on a beach with umbrellas

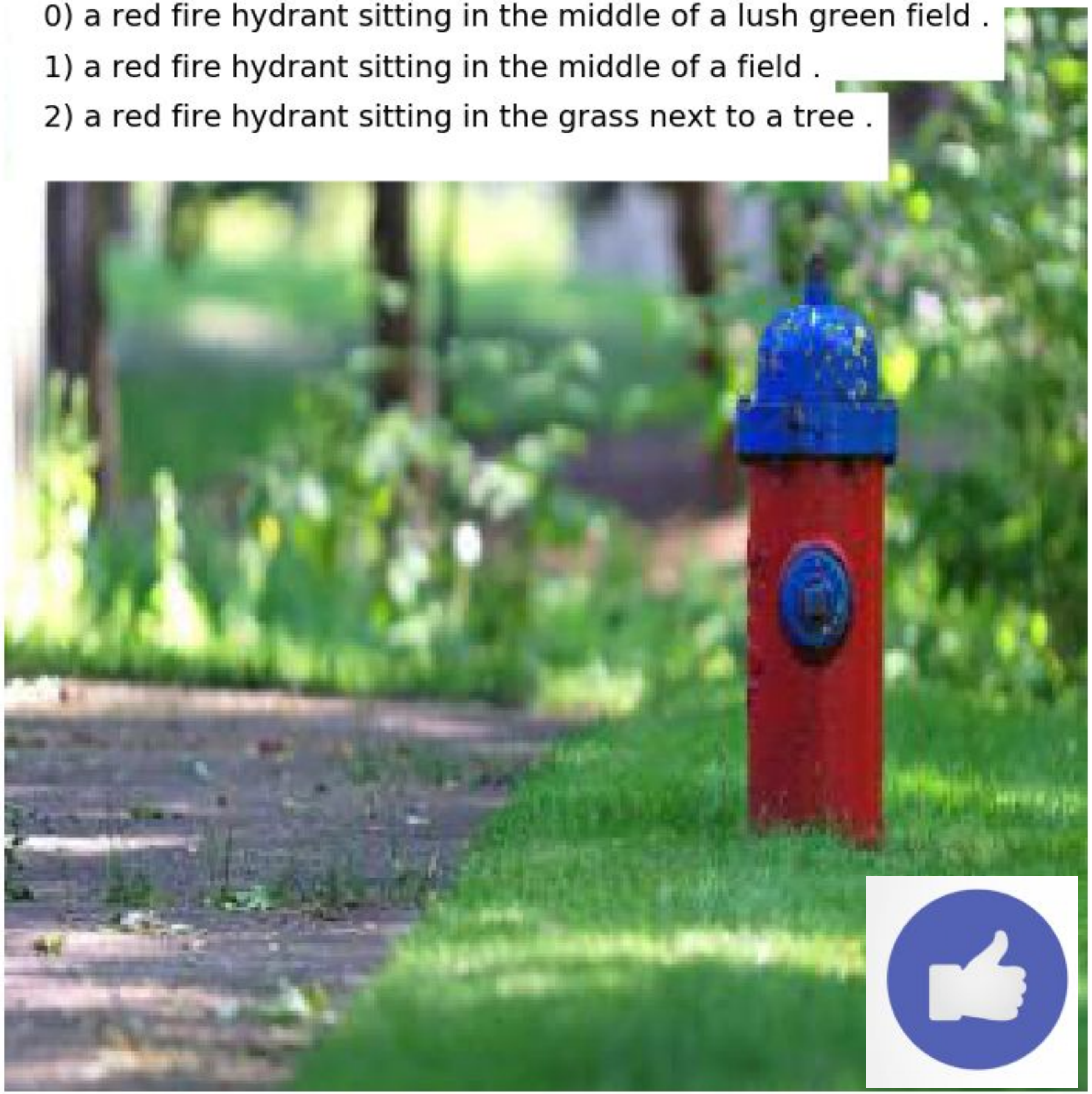
2) a group of people sitting on a beach under an umbrella .



0) a bench sitting in the middle of a forest .

1) a park bench in the middle of the woods .

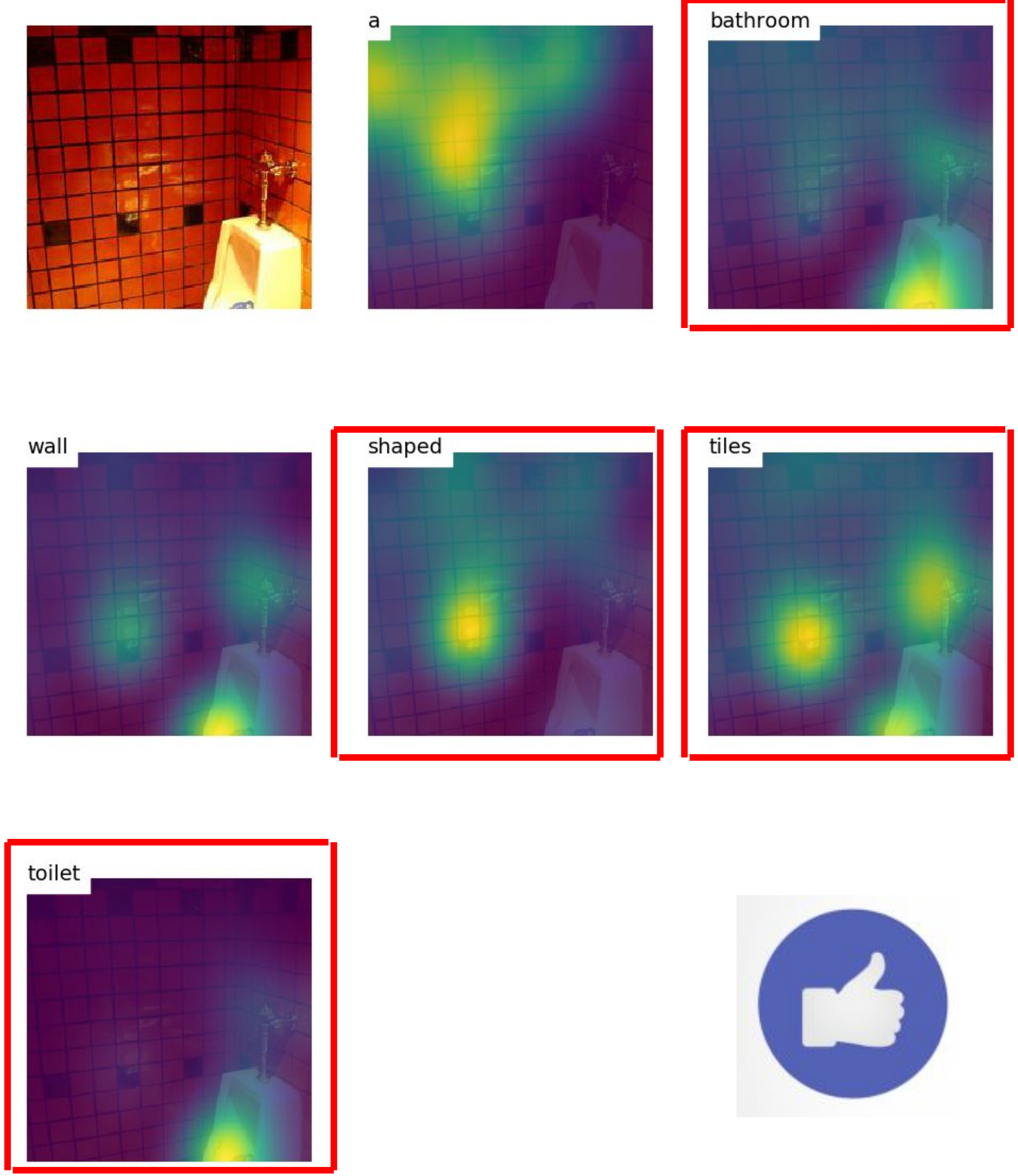
2) a park bench in the middle of a forest .

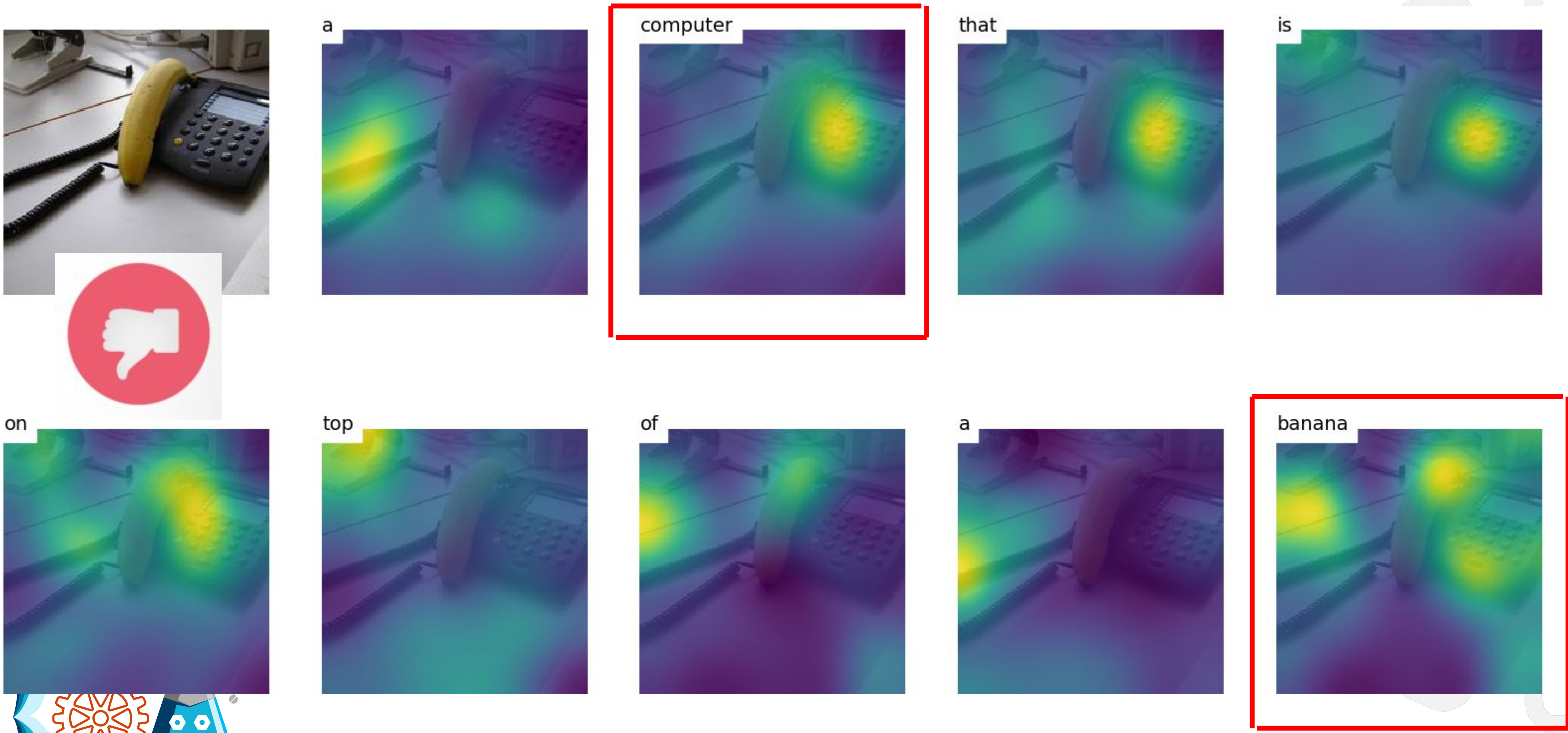


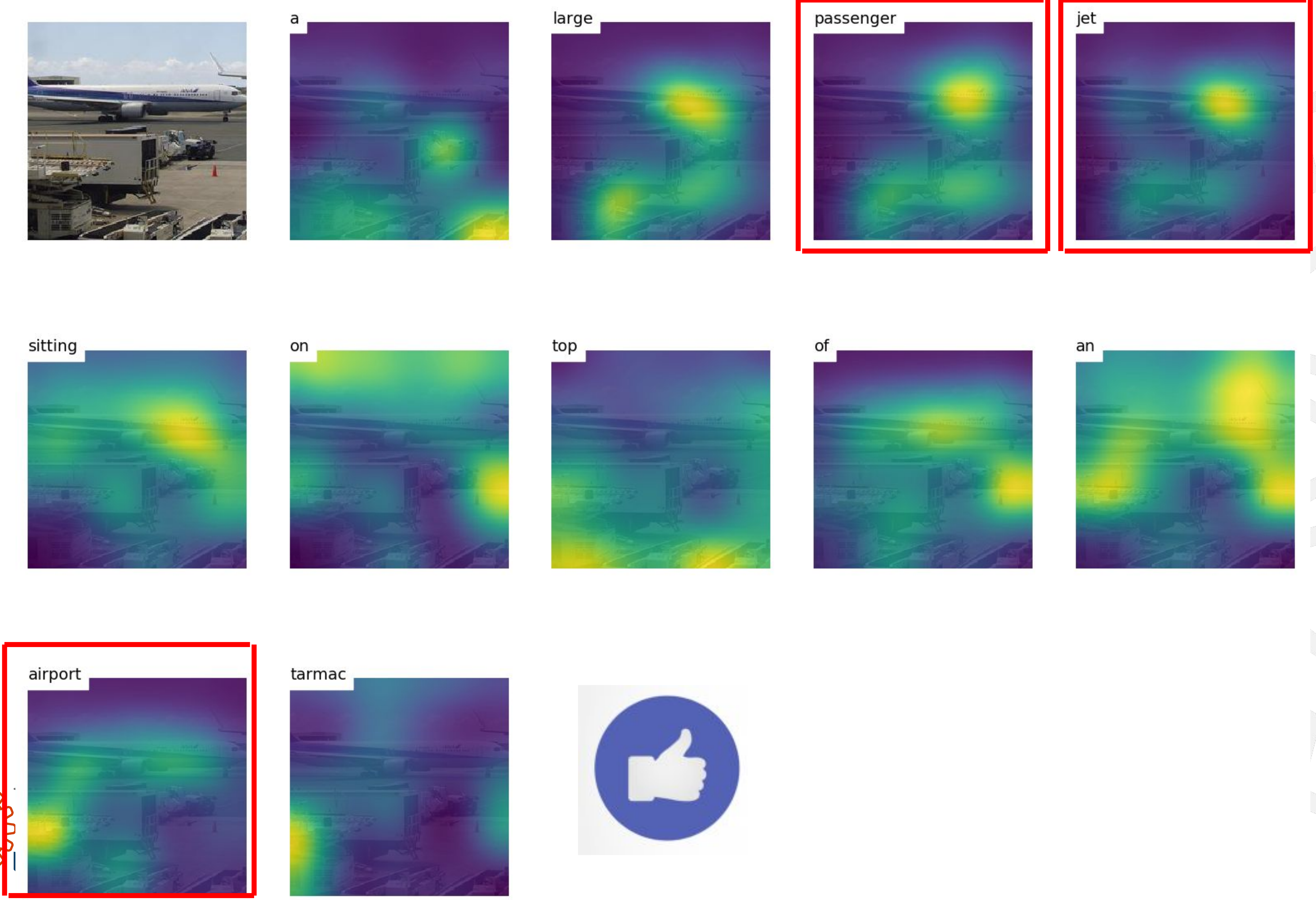
0) a red fire hydrant sitting in the middle of a lush green field .

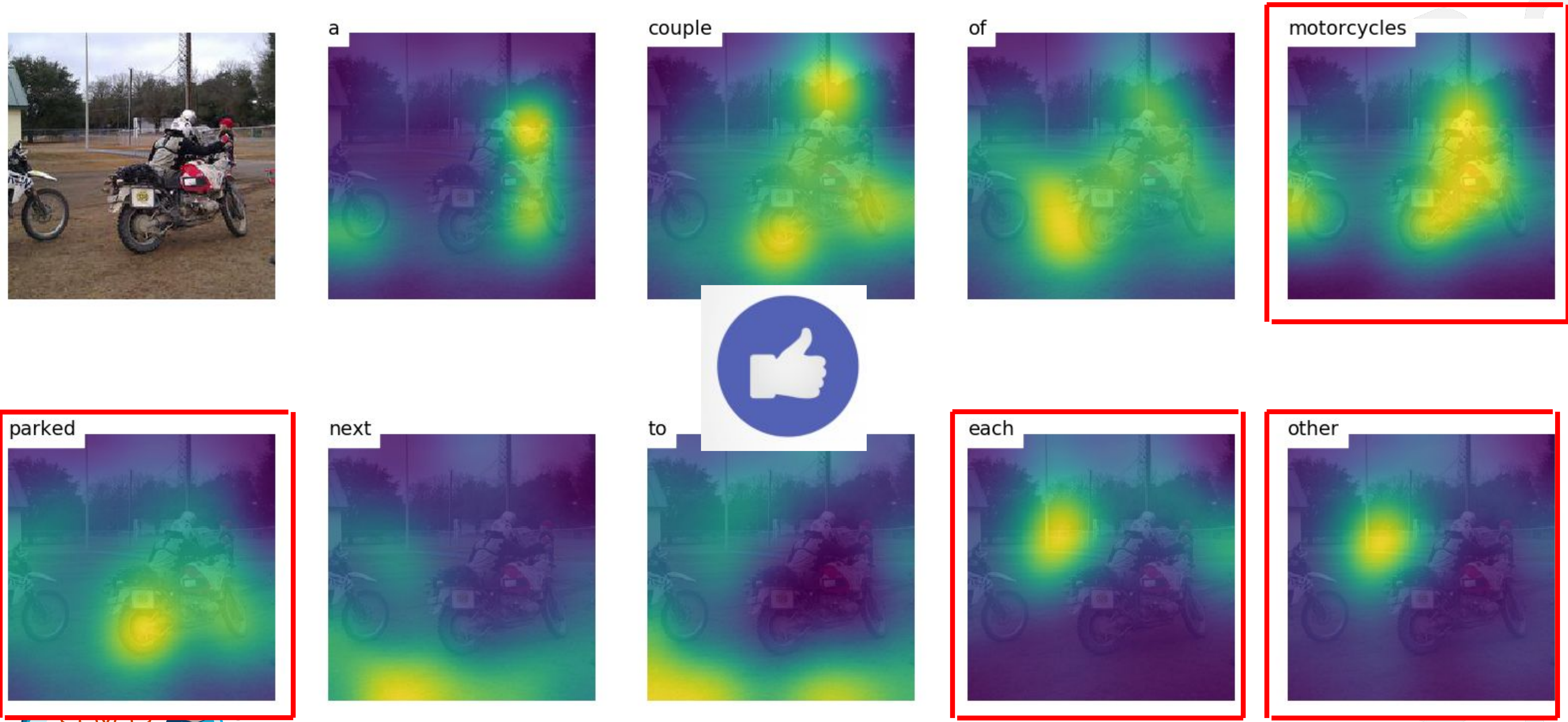
1) a red fire hydrant sitting in the middle of a field .

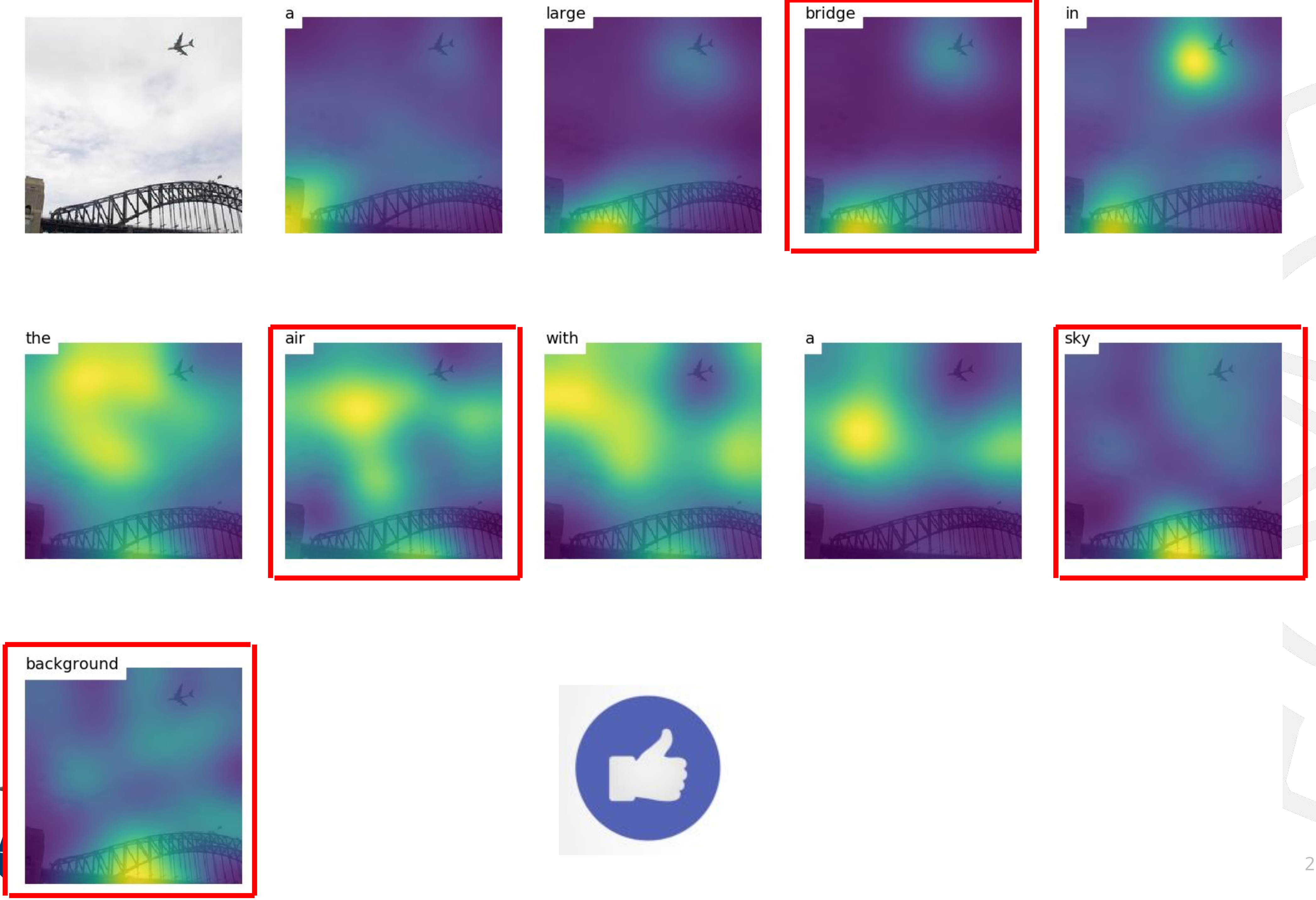
2) a red fire hydrant sitting in the grass next to a tree .


a
bathroom
with
a
wall
shaped
tiles
and
a
toilet


a
computer
that
is
on
top
of
a
banana


a
large
passenger
jet
sitting
on
top
of
an
airport
tarmac


a
couple
of
motorcycles
parked
next
to
each
other


a
large
bridge
in
the
air
with
a
sky
background

Benefits

Benefits

1. Learned to preprocess dataset wrt. input of the network

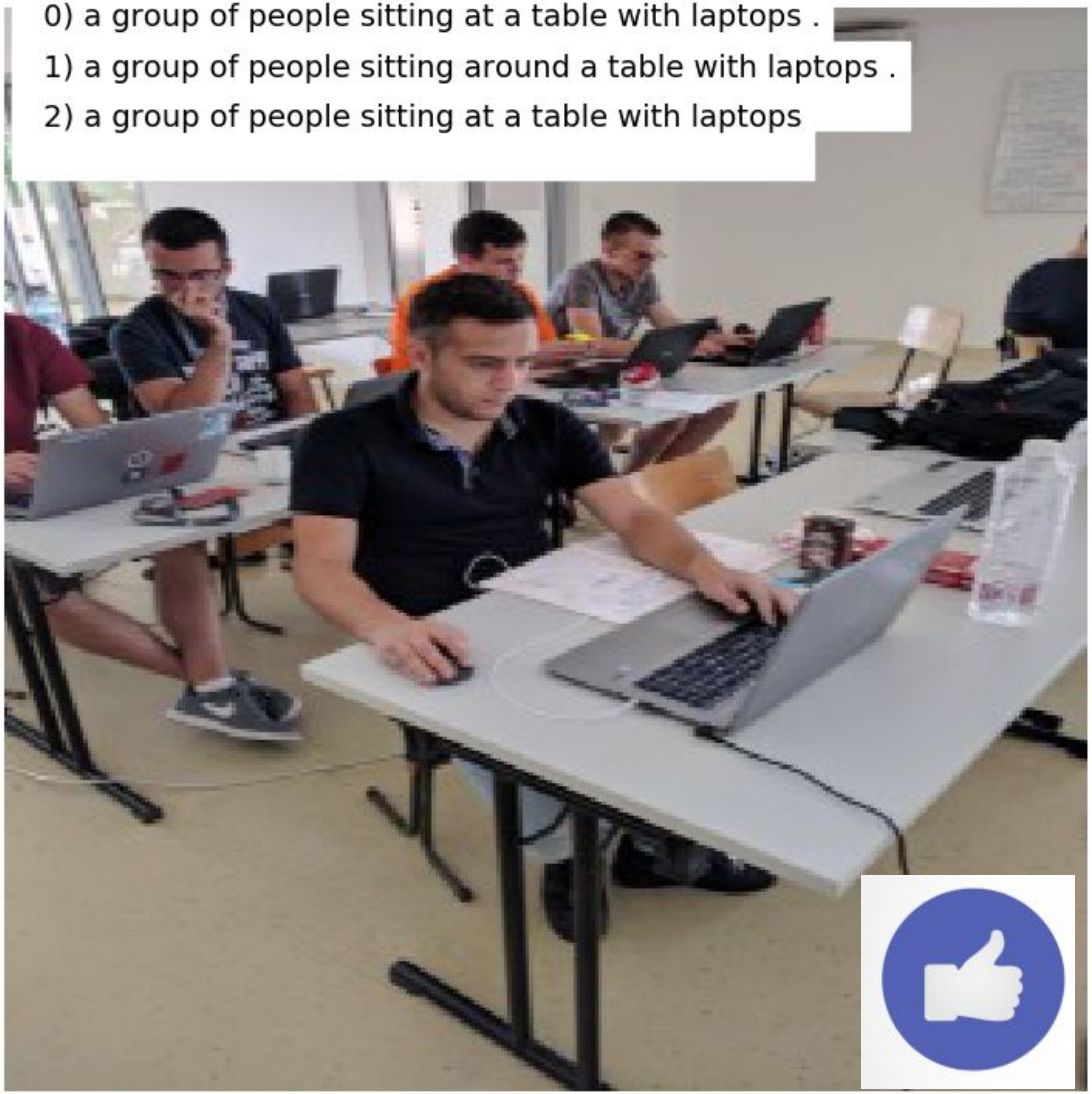
Benefits

1. Learned to preprocess dataset wrt. input of the network
2. **Improved tensorflow skills**

Benefits

1. Learned to preprocess dataset wrt. input of the network
2. Improved tensorflow skills
3. **Developed understanding of convolutional and recurrent neural networks**

Real world example



0) a group of people sitting at a table with laptops .

1) a group of people sitting around a table with laptops .

2) a group of people sitting at a table with laptops

Thank you for your ATTENTION !