Deep Learning for Vision & Language

Referring Expression Comprehension (Visual Grounding),

Visual Question Answering, Explainable Heatmaps

RICE UNIVERSITY

Today

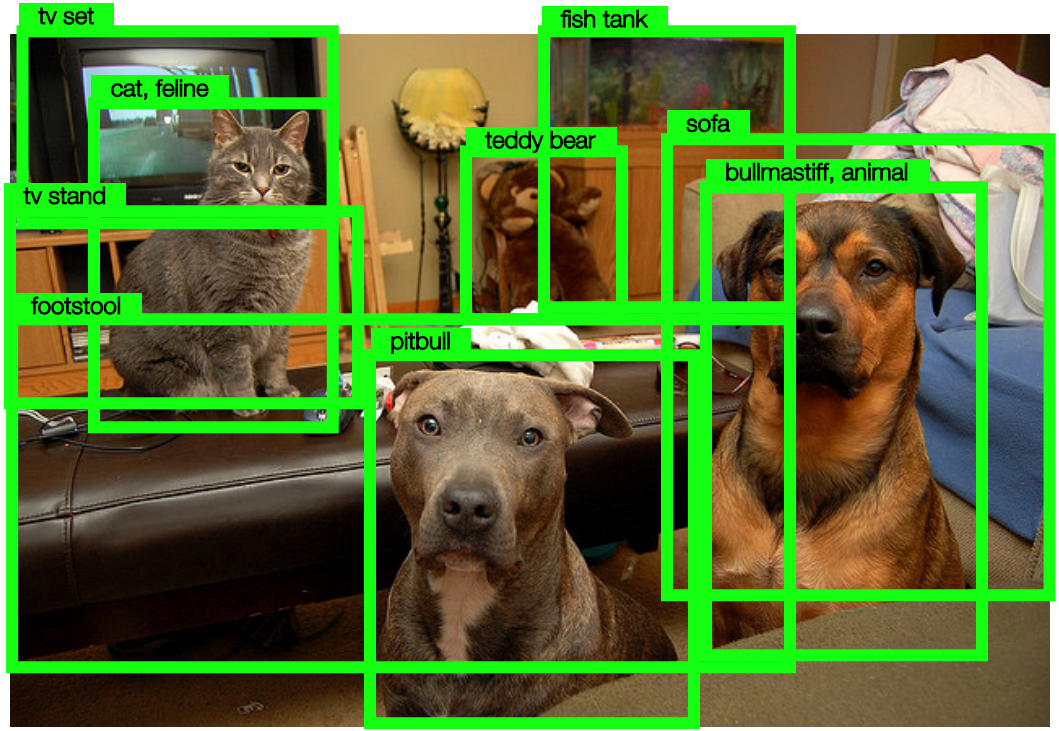
- Referring Expressions
  - Referring Expressions vs Image Captions
  - Generating Referring Expressions
  - Referring Expression Comprehension

Computer Vision

feline
tv set
teddy bear
pitbull
bullmastiff
cat
tv stand
group of dogs
fish tank
room
indoor
man-made
footstool
furniture

Image tagging / Image classification

Computer Vision



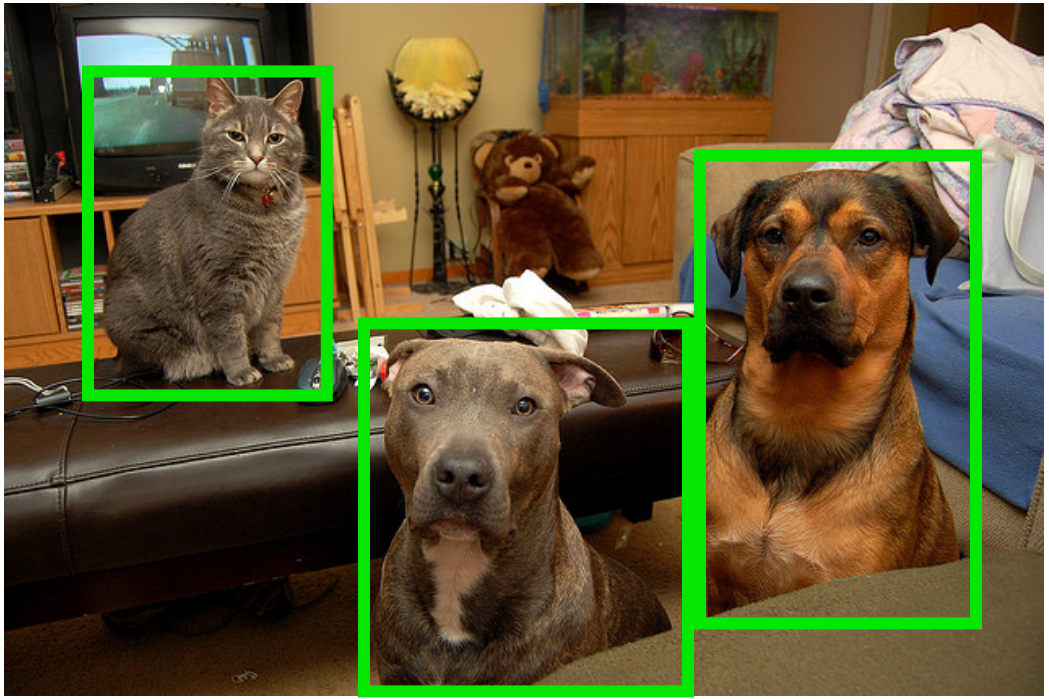
Object Detection

feline
tv set
teddy bear
pitbull
bullmastiff
cat
tv stand
group of dogs
fish tank
room
indoor
man-made
footstool
furniture

Computer Vision

Image Parsing / Image Segmentation

How do we describe images?



Object Importance

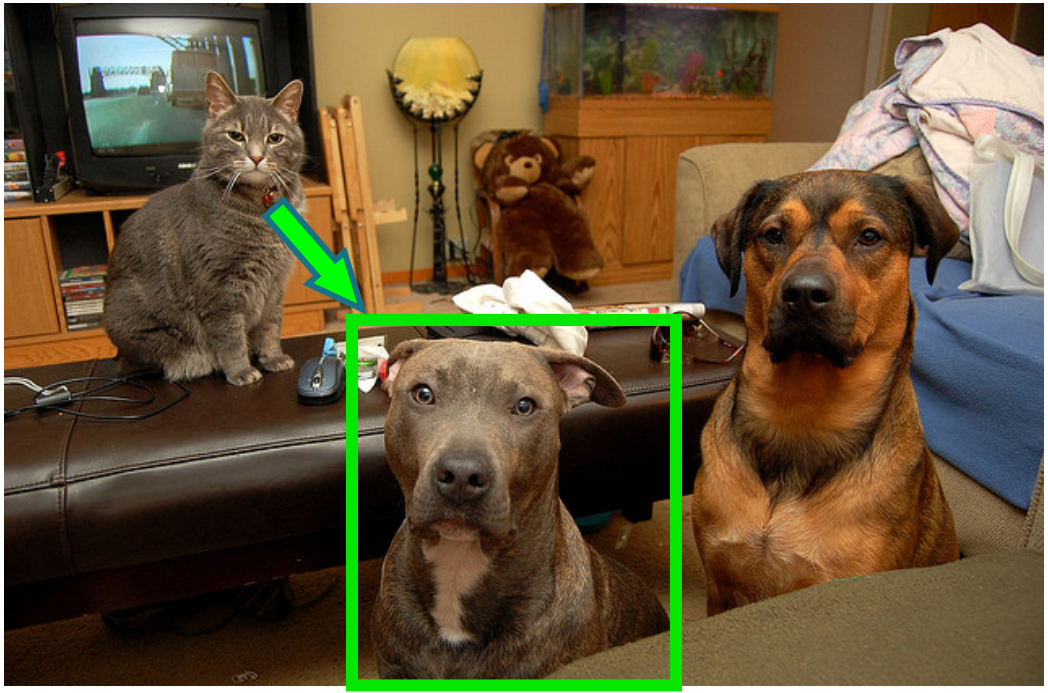
Attribute Importance

Action Importance

World knowledge

A cat and two big dogs staring at the camera

Referring to objects



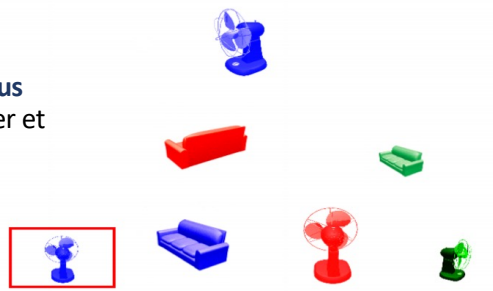
The dog in the middle

The gray dog in the middle

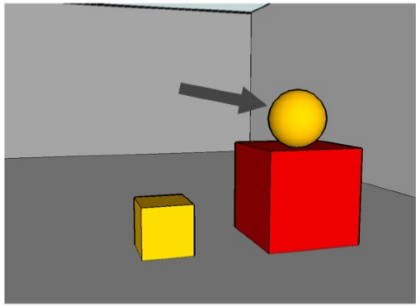
The gray dog

Work on Referring Expression

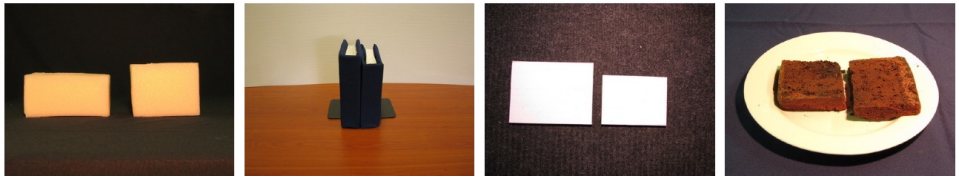
TUNA Corpus
van Deemter et al 2006



GRE3D3 Corpus
Viethen and Dale 2008
[20 scenes]

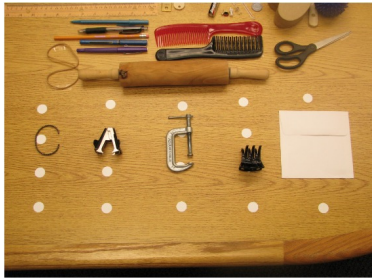


Size Corpus
Mitchell et al 2011
[96 scenes]



GenX Corpus
FitzGerald et al 2013
[269 scenes]

Typicality Corpus
Mitchell et al 2013
[35 scenes]



ReferItGame: Referring to Objects in Photographs of Natural Scenes
Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, Tamara L. Berg.
Empirical Methods on Natural Language Processing. **EMNLP 2014**. Doha, Qatar. October 2014.

Referit Game

Player 1

Player 2

Referring Expressions for Natural Scenes

Diverse

Many real world objects

Complex

Many object instances

Big

IAPR TC-12 Segmented and Annotated Dataset. Escalante et. al. 2009

Referit Game Dataset

Blue shirt man

Blue guy

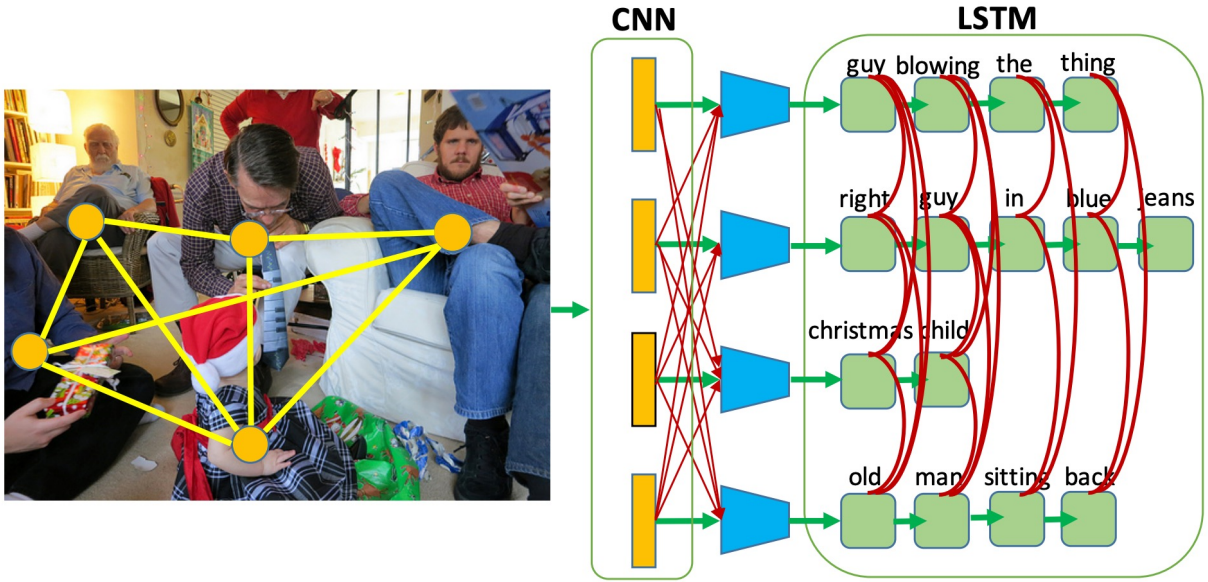
Second guy from left

ReferItGame Dataset

130k Referring expressions for **90k** Objects in **19k** images

ReferItGame: Referring to Objects in Photographs of Natural Scenes
Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, Tamara L. Berg.
Empirical Methods on Natural Language Processing. **EMNLP 2014**.

Deep Generation of Referring Expressions



Modeling Context in Referring Expressions

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, Tamara L. Berg

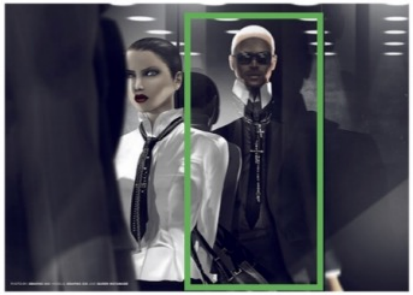
Department of Computer Science,
University of North Carolina at Chapel Hill
{licheng,poirson,alexyang,aberg,tlberg}@cs.unc.edu

2016

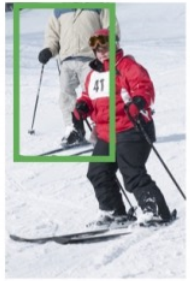
RefCOCO+ testA

Baseline: blue shirt
MMI: black shirt
visdif: person in stripped shirt
visdif+tie: arm with stripped shirt

Baseline: tennis player
MMI: girl
visdif: woman in white
visdif+tie: tennis player

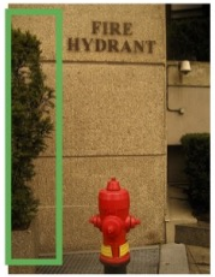


Baseline: man
MMI: man
visdif: man with glasses
visdif+tie: man with glasses

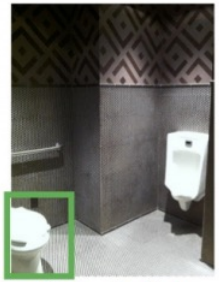


Baseline: red jacket
MMI: red jacket
visdif: skier in white
visdif+tie: man in white

RefCOCO+ testB

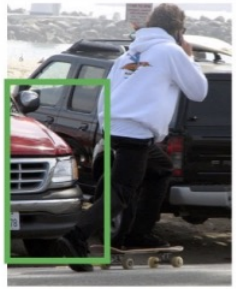


Baseline: plant
MMI: plant that is cut off
visdif: tall plant
visdif+tie: plant on screen side



Baseline: toilet
MMI: toilet
visdif: toilet with lid
visdif+tie: toilet with lid

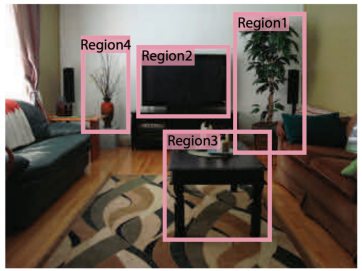
Baseline: donut at 3
MMI: glazed donut
visdif: donut with hole
visdif+tie: donut with hole



Baseline: car with red roof
MMI: car
visdif: car with headlights
visdif+tie: car with headlights

Referring Expression Comprehension

The plant on the
right side of the TV

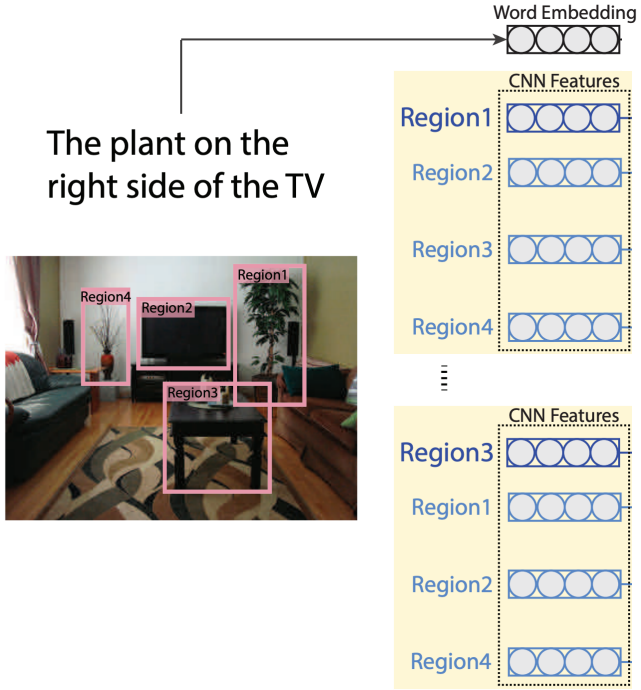


Modeling Context Between Objects for Referring Expression Understanding

Varun K. Nagaraja Vlad I. Morariu Larry S. Davis

University of Maryland, College Park, MD, USA.
{varun,morariu,lsd}@umiacs.umd.edu

Referring Expression Comprehension

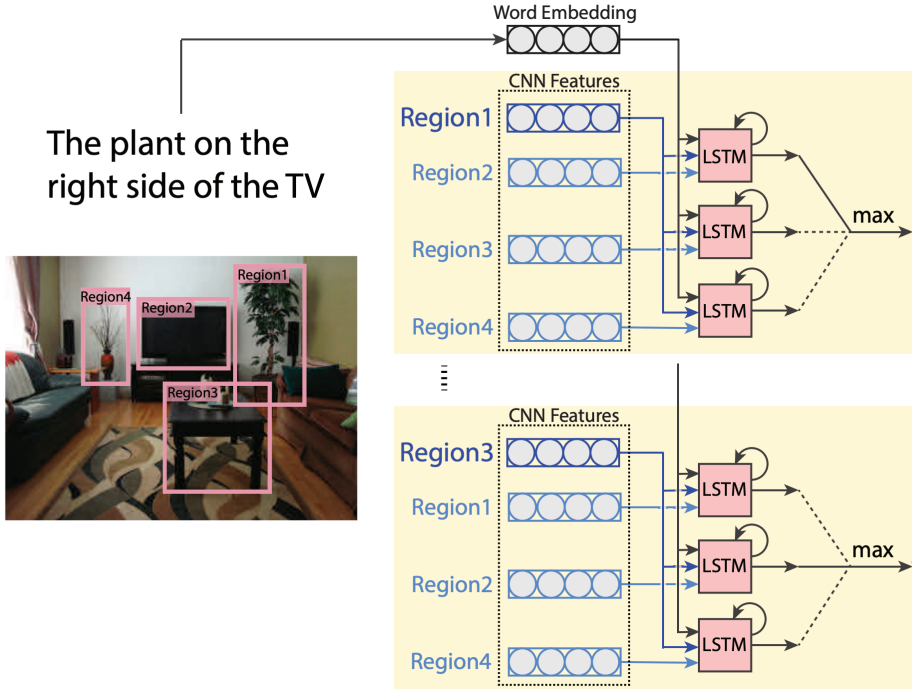


Modeling Context Between Objects for Referring Expression Understanding

Varun K. Nagaraja Vlad I. Morariu Larry S. Davis

University of Maryland, College Park, MD, USA.
{varun,morariu,lsd}@umiacs.umd.edu

Referring Expression Comprehension

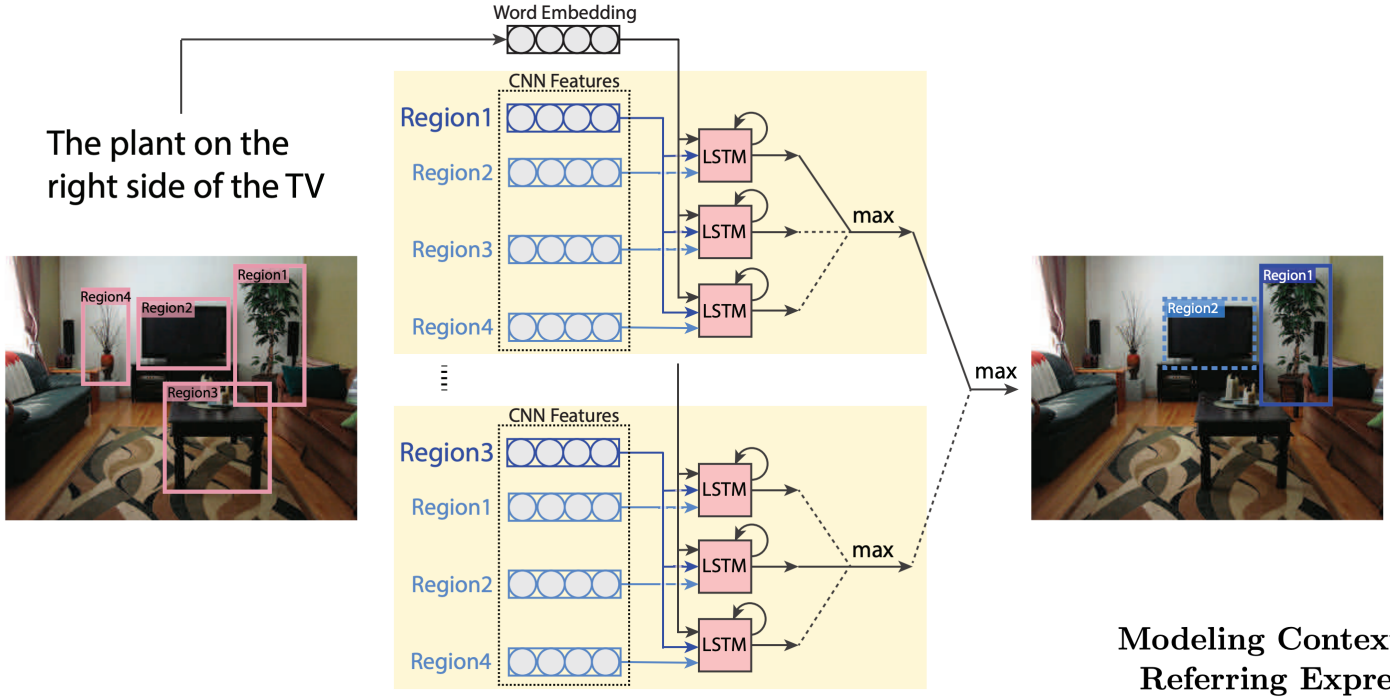


jects for anding

Varun K. Nagaraja Vlad I. Morariu Larry S. Davis

University of Maryland, College Park, MD, USA.
{varun,morariu,lsd}@umiacs.umd.edu

Referring Expression Comprehension



Modeling Context Between Objects for Referring Expression Understanding

Varun K. Nagaraja Vlad I. Morariu Larry S. Davis

University of Maryland, College Park, MD, USA.
{varun,morariu,lsd}@umiacs.umd.edu

2016

Other important work

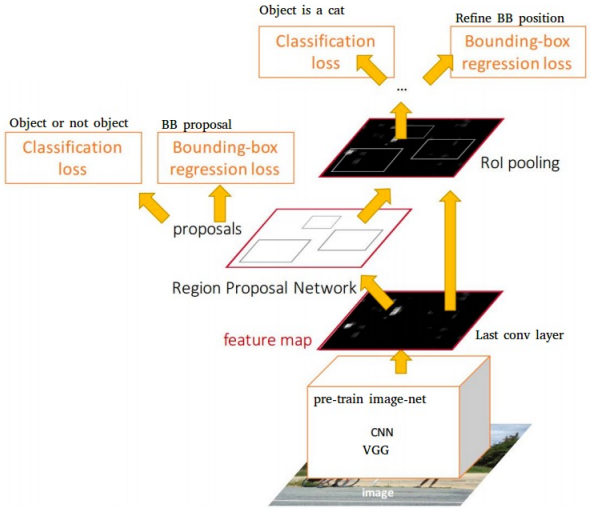
MattNet: Yu et al. https://arxiv.org/abs/1801.08186

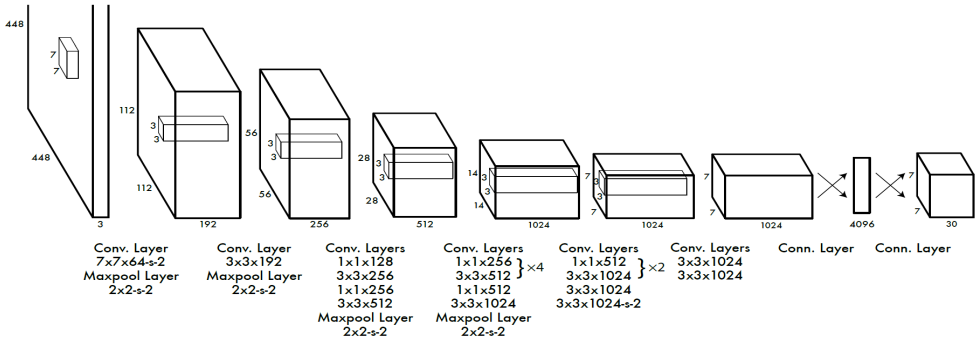
Mao et al. https://arxiv.org/abs/1511.02283

Rohrbach et al. https://arxiv.org/abs/1511.03745

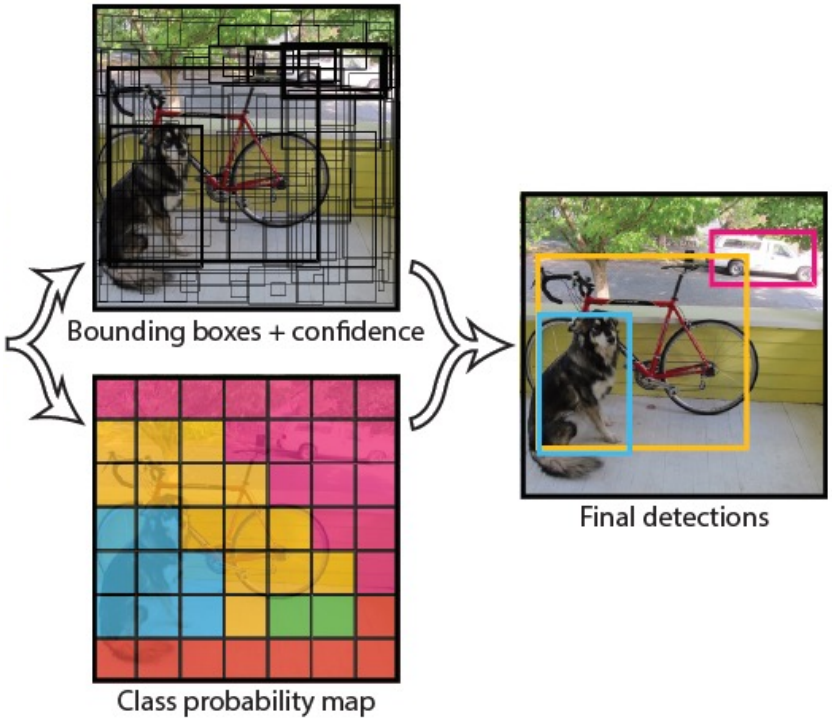
Detour: Recap on Object Detection

- Two-stage: Faster-RCNN, Mask-RCNN
- Single-stage: YOLO (You Only Look Once), SSD (Single Shot Detector)

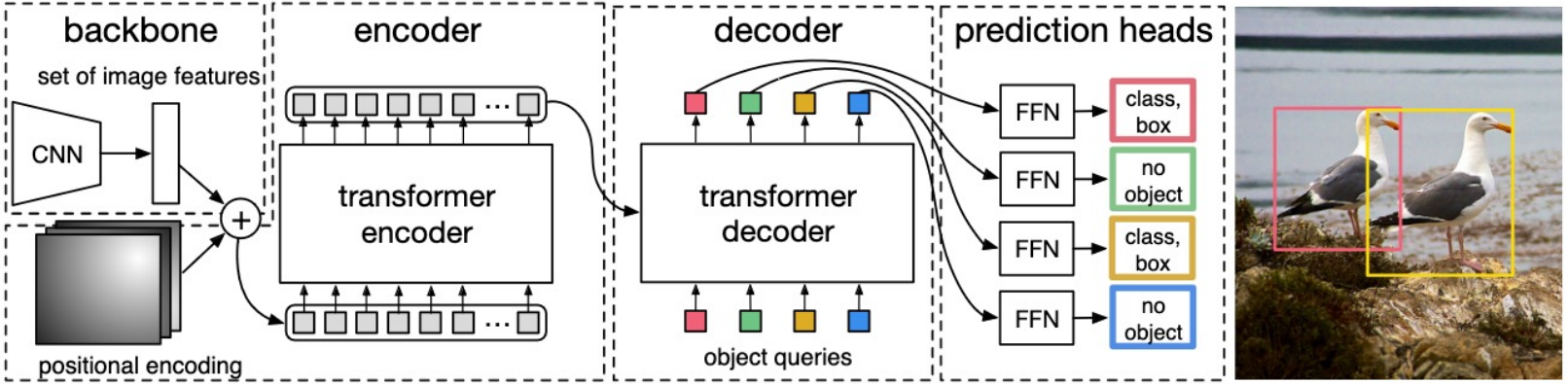




Post-processing: Non-Max Suppression



End-to-end Object Detection with Transfromers (DETR) (2020)



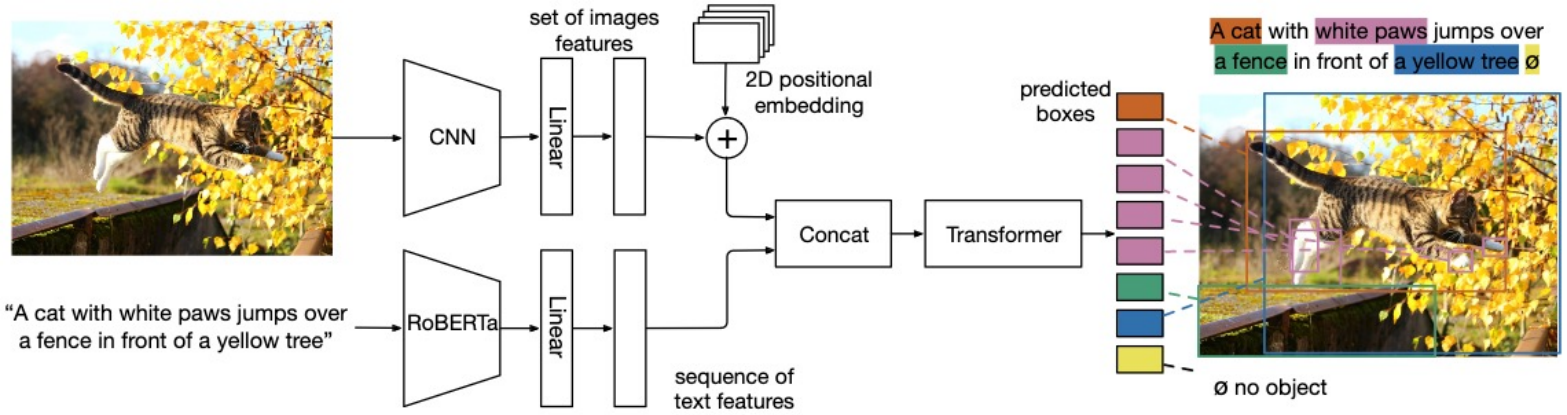
$$\mathcal{L}_{\text{Hungarian}}(y, \hat{y}) = \sum_{i=1}^{N} \left[-\log \hat{p}_{\hat{\sigma}(i)}(c_i) + \mathbb{1}_{\{c_i \neq \varnothing\}} \mathcal{L}_{\text{box}}(b_i, \hat{b}_{\hat{\sigma}}(i)) \right]$$

where $$\hat{\sigma} = \arg\min_{\sigma \in \mathfrak{S}_N} \sum_{i}^{N} \mathcal{L}_{\text{match}}(y_i, \hat{y}_{\sigma(i)})$$

https://arxiv.org/abs/2005.12872

https://github.com/facebookresearch/detr

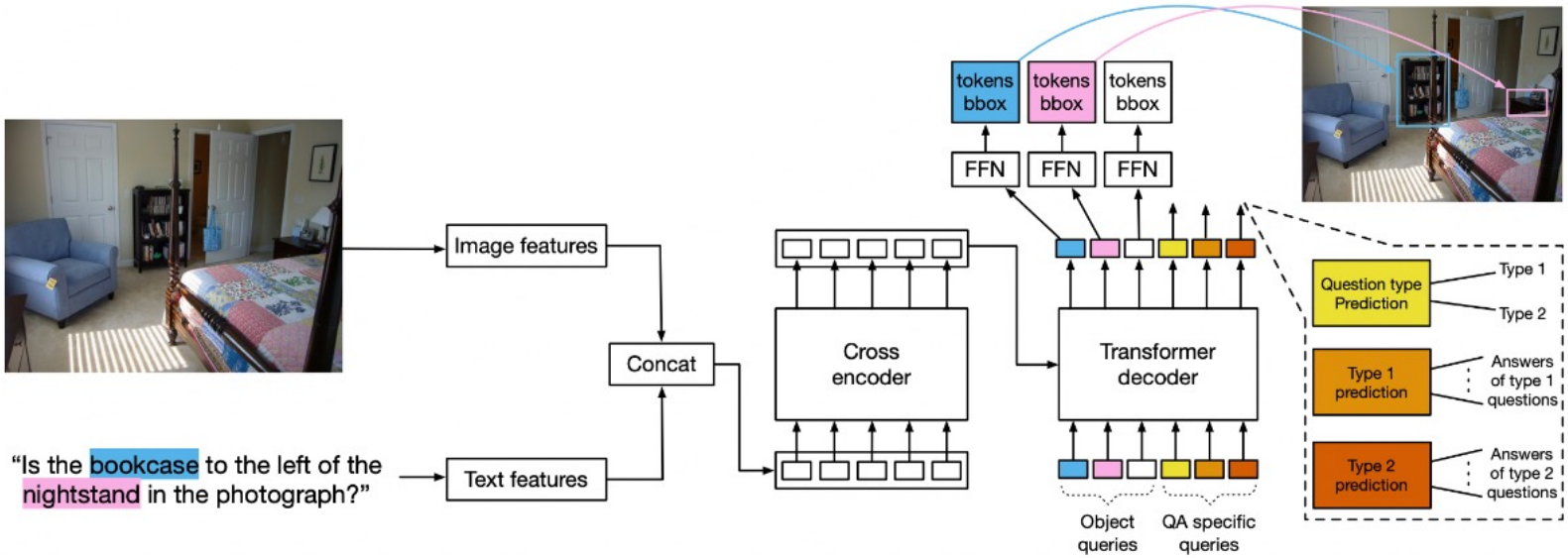
MDETR: Modulated Detection for Multimodal Understanding (2021)



https://arxiv.org/pdf/2104.12763.pdf

https://github.com/facebookresearch/detr

MDETR: For Question Answering



Visually Grounded Question Answering

How many slices of pizza are there?
Is this a vegetarian pizza?

https://arxiv.org/pdf/1505.00468.pdf

Visually Grounded Question Answering

How many slices of pizza are there?
Is this a vegetarian pizza?

https://arxiv.org/pdf/1505.00468.pdf

VQA: Visual Question Answering

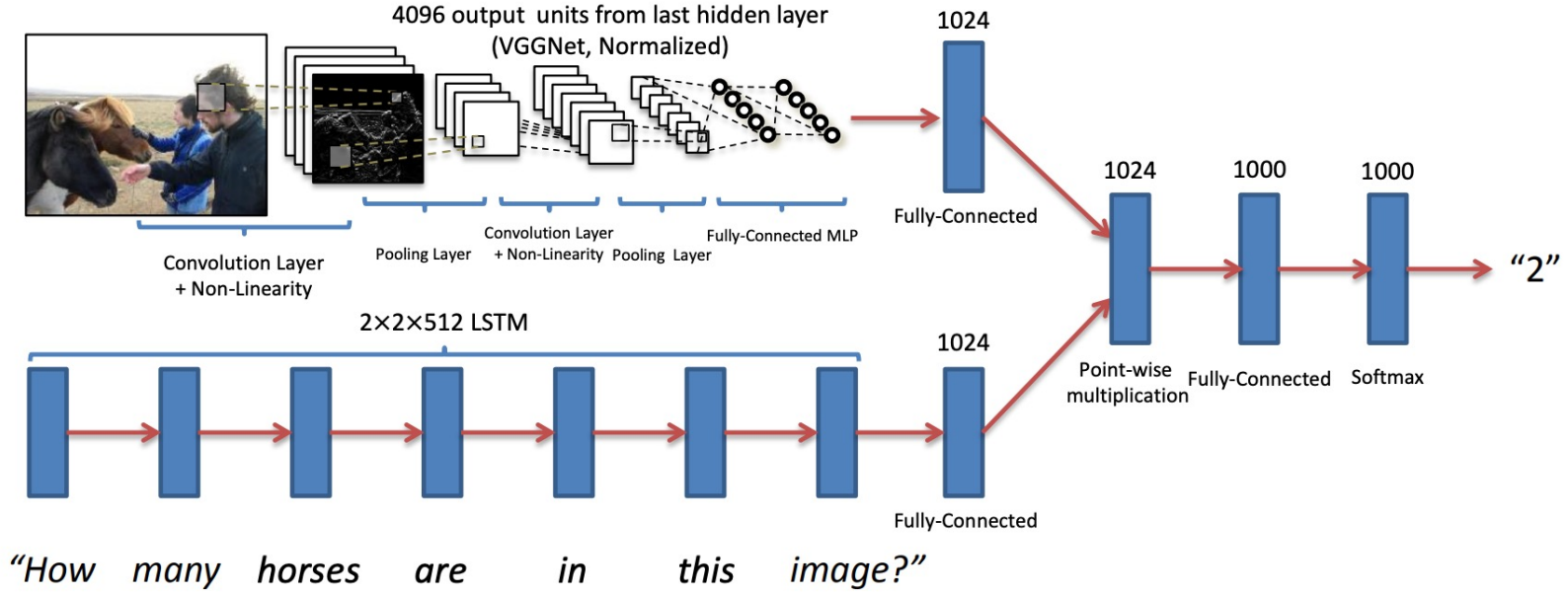
www.visualqa.org

Aishwarya Agrawal*, Jiasen Lu*, Stanislaw Antol*,
Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, Devi Parikh

Is this person trying to hit a ball?	yes yes yes	yes yes yes
What is the person hitting the ball with?	frisbie racket round paddle	bat bat racket

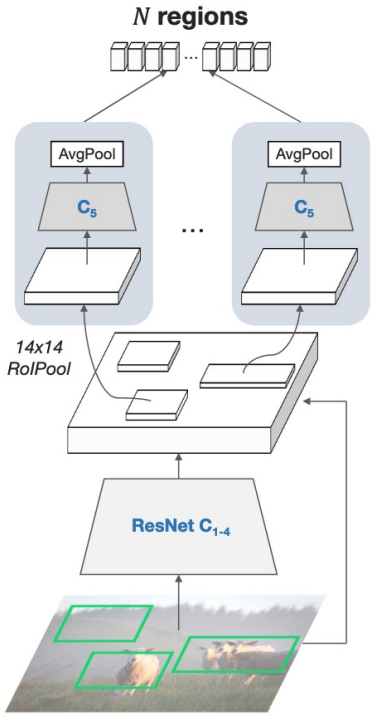
What is the guy doing as he sits on the bench?	phone taking picture taking picture with phone	reading reading smokes
What color are his shoes?	blue blue blue	black black brown

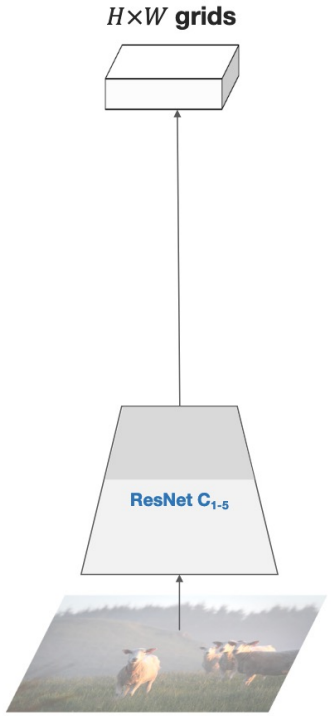
Visually Grounded Question Answering



https://arxiv.org/pdf/1505.00468.pdf

What Features to use as input visual features?

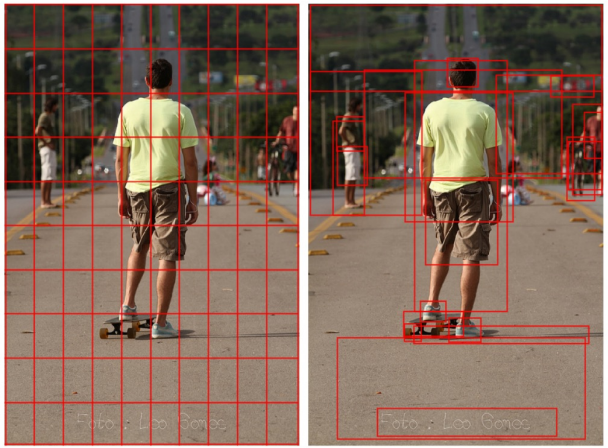




CVPR 2017

Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering

Peter Anderson[1]* Xiaodong He[2] Chris Buehler[3] Damien Teney[4]
Mark Johnson[5] Stephen Gould[1] Lei Zhang[3]
[1]Australian National University [2]JD AI Research [3]Microsoft Research [4]University of Adelaide [5]Macquarie University
[1]firstname.lastname@anu.edu.au, [2]xiaodong.he@jd.com, [3]{chris.buehler,leizhang}@microsoft.com
[4]damien.teney@adelaide.edu.au, [5]mark.johnson@mq.edu.au



Question: What room are they in? Answer: kitchen

CVPR 2020

In Defense of Grid Features for Visual Question Answering

Huaizu Jiang[1,2*], Ishan Misra[2], Marcus Rohrbach[2], Erik Learned-Miller[1], and Xinlei Chen[2]
[1]UMass Amherst, [2]Facebook AI Research (FAIR)
{hzjiang,elm}@cs.umass.edu, {imisra,mrf,xinleic}@fb.com

VQA Solution today?

What is the color of the jacket of the man on this picture?

?

?

?

Cross Entroy Loss Across 5000 possible answers

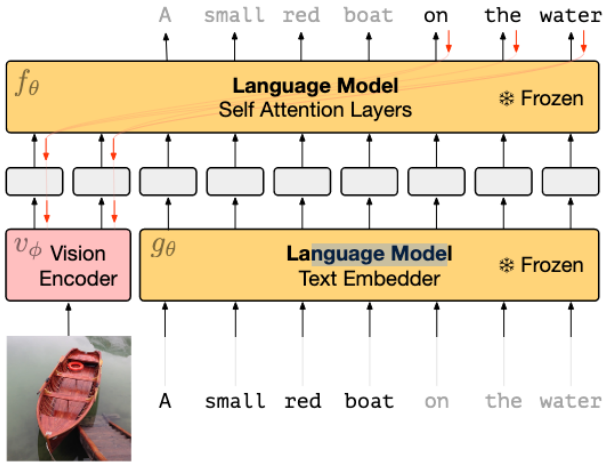
VQA Solution today?

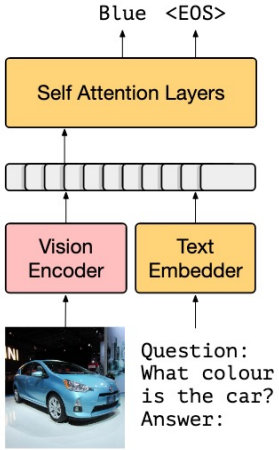
What is the color of the jacket of the man on this picture?

?

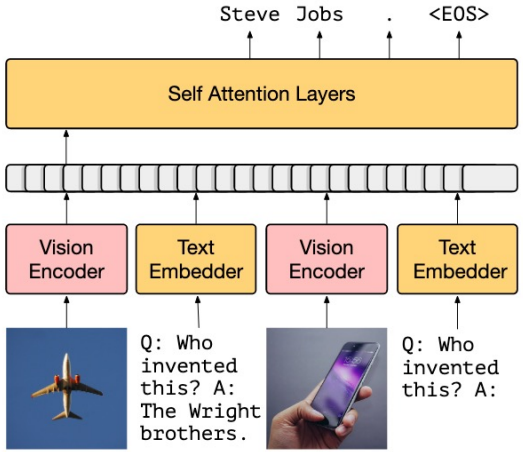
Cross Entroy Loss Across 5000 possible answers

Training:

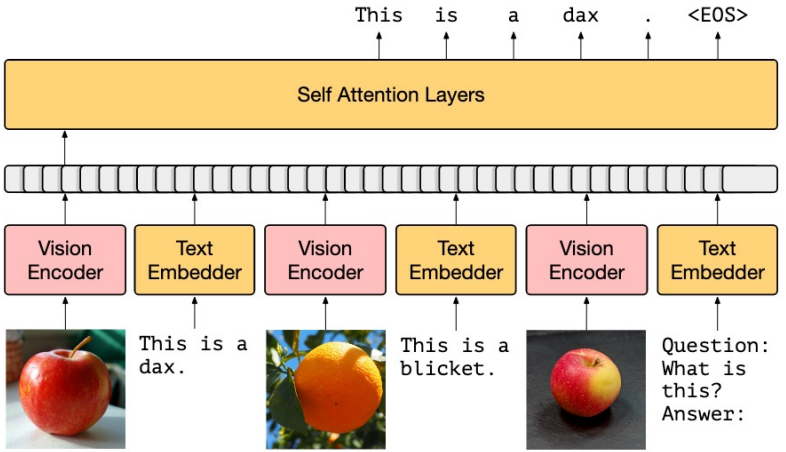




(a) **0-shot VQA**



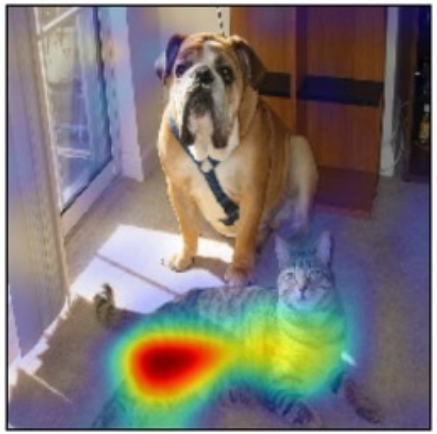
(b) **1-shot outside-knowledge VQA**



(c) **Few-shot image classification**

Explainability: GradCAM

(a) Original Image

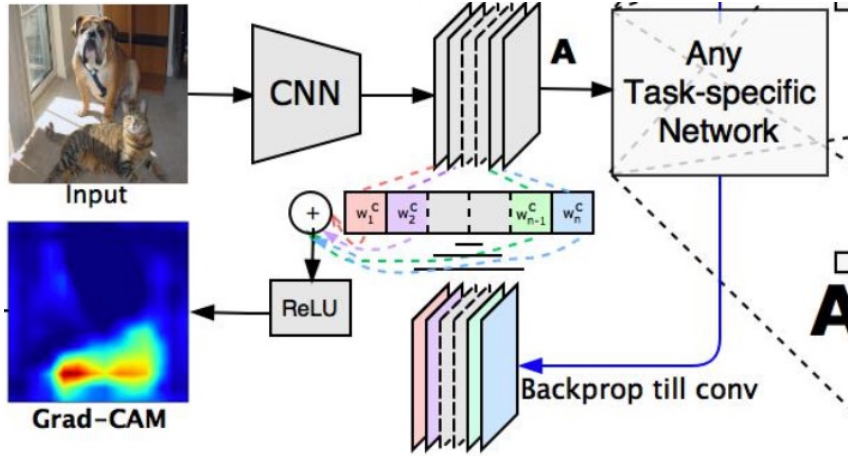


(c) Grad-CAM ‘Cat’

(i) Grad-CAM ‘Dog’

https://arxiv.org/abs/1610.02391

Explainability: GradCAM



$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}}$$

$$L^c_{\text{Grad-CAM}} = ReLU\underbrace{\left(\sum_k \alpha_k^c A^k\right)}_{\text{linear combination}}$$

https://arxiv.org/abs/1610.02391

Questions?