DittoNet: Pokemon Visual Question Answering

Satish Palaniappan, Ayush Agarwal, Sonakshi Grover

Code: https://github.com/tpsatish95/pokemon-vqa

Problem Statement

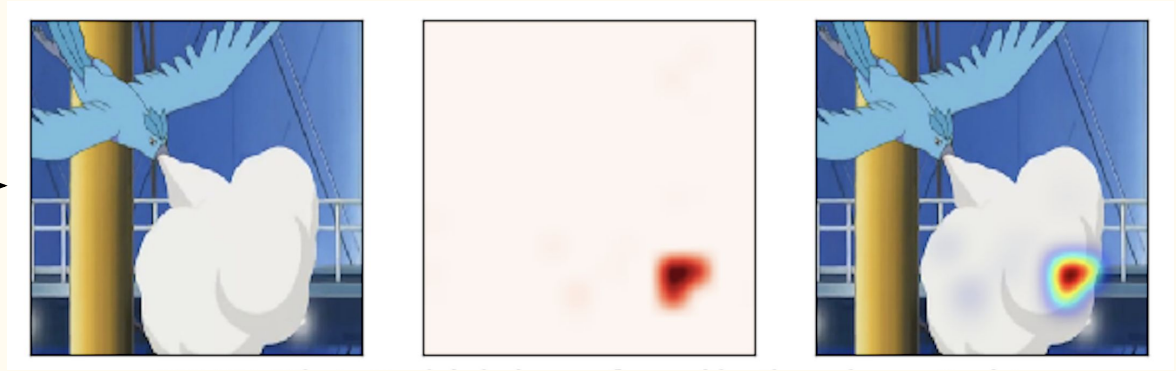
We will be solving the VQA problem in the domain of the popular anime series, Pokémon. Our system considers the first 150 Pokémon (Kanto Region) and some typical questions our Pokémon VQA model will try to answer are:

Q: What Pokémon is there in the image? **A:** Pikachu.
Q: What attack is being performed by Pikachu? **A:** Thunderbolt
Q: What type of Pokémon is Pikachu? **A:** Electric

The problem statement involves the development of an algorithm which takes an image and a textual question as input and outputs an answer to the that question based on the visual features of the input image.

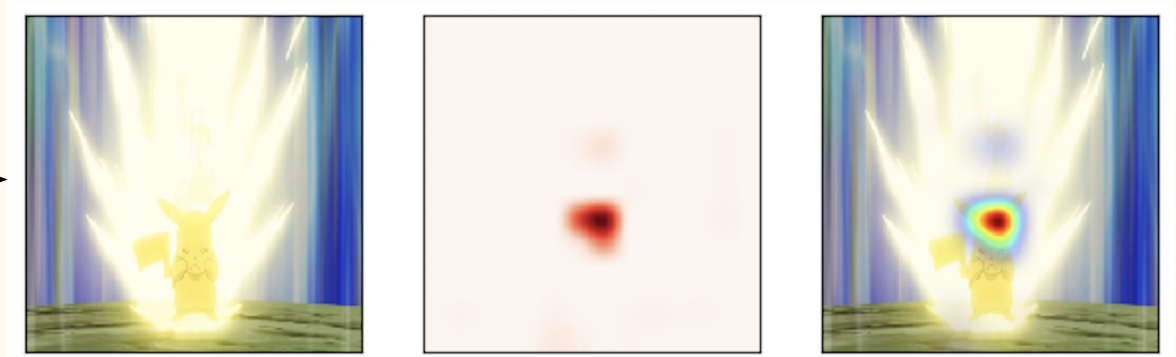
Expected Output

Q: *Which attack is being performed by the Pokemon?*



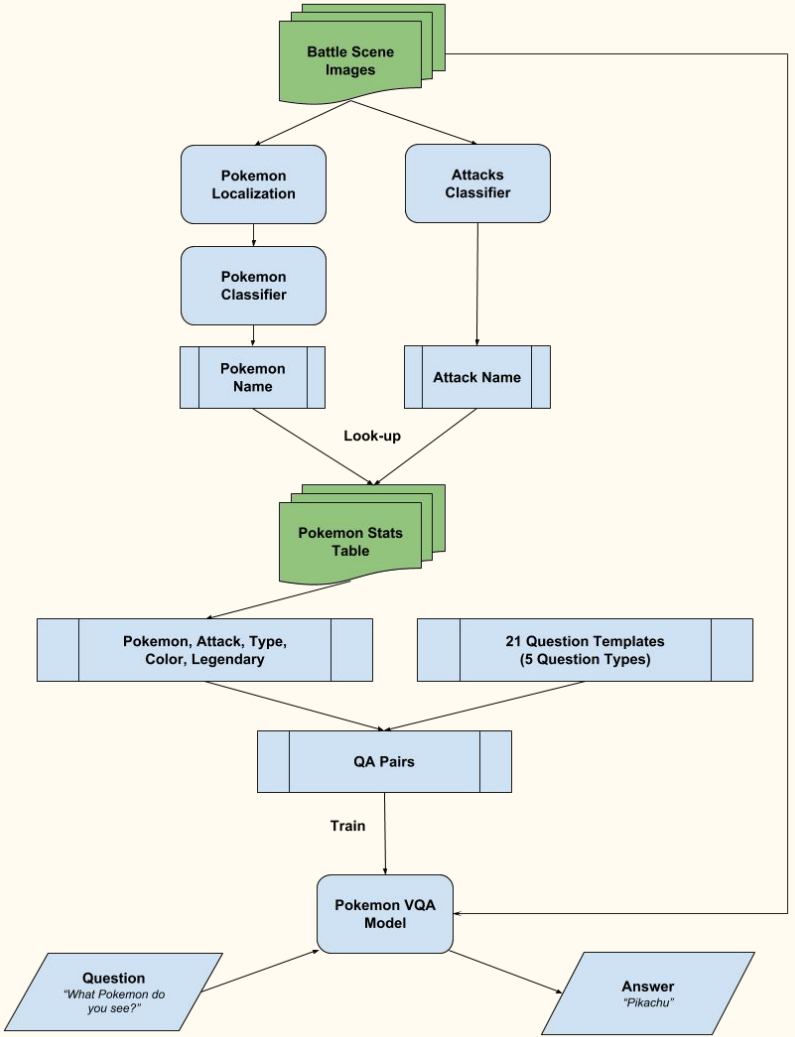
A: *Mist*

Q: *What Pokemon do you see in this picture?*



A: *Pikachu*

Project Pipeline



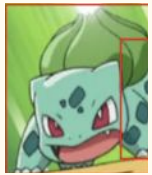
Challenges

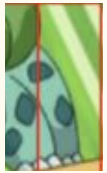
- Dataset’s dimensionality is less as compared to MSCOCO dataset.

- Unlike MSCOCO, our images are very bright, filled with fire/water/grass type Pokemon which are scattered across the image and are not localized.

- We faced an additional challenge of manually constructing questions, which was tougher in our case due to a lesser number of discriminative objects in our dataset.

Creating the VQA Dataset: Localize Pokemon

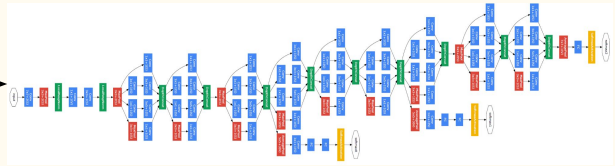




Pokemon Other Part

Selective Search
***Scale**: 350, 450, 500*
***Sigma**: 0.8*
***Min. Size**: 30, 60, 120*
***Min Area:** 2000*
(+ basic region filtering)

Pokemon/Part/Other Classifier



Classes: 3
Images: 3639
Train:Test :: 80:20

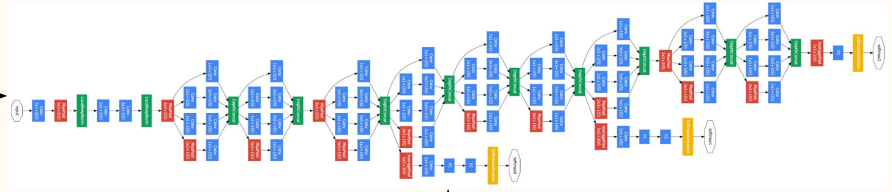
GoogLeNet Inception v1
Accuracy: 74.9%

Creating the VQA Dataset: Classify Pokemon

Pokemon Classifier

Pokemon Dataset
Classes: 150
Images: 134508 (BG+Aug)
Train:Test :: 70:30

GoogLeNet Inception v1
Accuracy: 91.16% (Top 1)

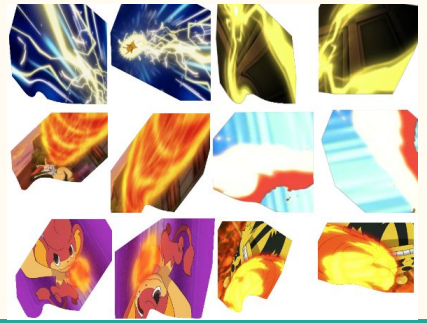


“ImageNet” Pre-trained Weights

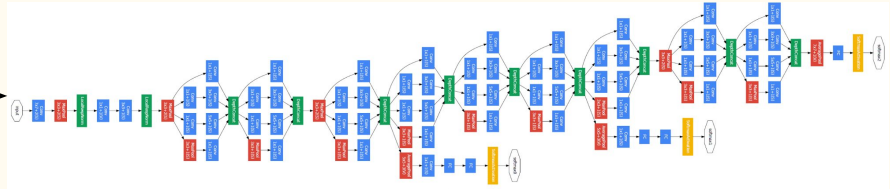
Creating the VQA Dataset: Classify Attacks

Attacks Classifier

Attacks Dataset
Classes: 144
Images: 21188 (Aug)
Train:Test :: 70:30



GoogLeNet Inception v1
Accuracy: 55% (Top 1), 88% (Top 5)



"MIT Places 205" Pre-trained Weights

Creating the VQA Dataset: Get Misc. Info

Pokemon: Bulbasaur
Attack: Leech Seed

Pokemon Stats Table

Number	Name	Type_1	isLegendary	Color
1	Bulbasaur	Grass	False	Green
2	Ivysaur	Grass	False	Green
3	Venusaur	Grass	False	Green
4	Charmander	Fire	False	Red
5	Charmeleon	Fire	False	Red
6	Charizard	Fire	False	Red

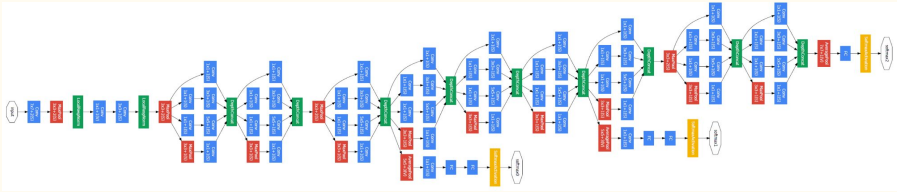
Pokemon: Bulbasaur
Attack: Leech Seed
Type: Grass
Legendary: True
Color: Green

Creating the VQA Dataset: QA Pairs

- We have a battle scenes dataset of 1511 images.
- We generate [Battle Scene Image, Pokemon, Attack,Color, Type, Is Legendary] sets for each of the battle scenes.
- Question Types
  - What pokemon is there in the image?
  - What attack is being performed by the pokemon?
  - What type of pokemon is it?
  - Is the pokemon legendary?
  - What is the color of the pokemon in the image?
- We rephrased the above 5 questions and generated 21 questions in total, per battle scene image.
- Thus we get: **31,731 QA pairs** and we use a **70:30 train:test** split.

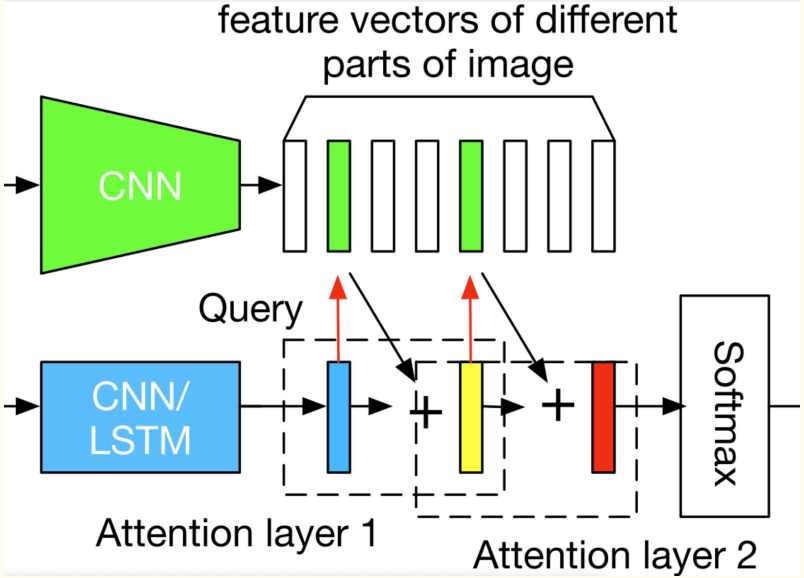
Model Architecture and Parameters

Pokemon and Attack Detection



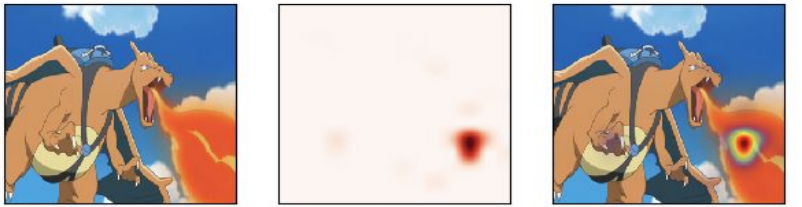
```
net: "googlenet/train_val.prototxt"
test_iter: 1407
test_interval: 4000
test_initialization: false
display: 40
base_lr: 0.001
lr_policy: "step"
stepsize: 32000
gamma: 0.1
max_iter: 10000000
momentum: 0.9
weight_decay: 0.0002
snapshot: 4000
snapshot_prefix: "bvlc_googlenet_pokenet"
solver_mode: GPU
```

VQA

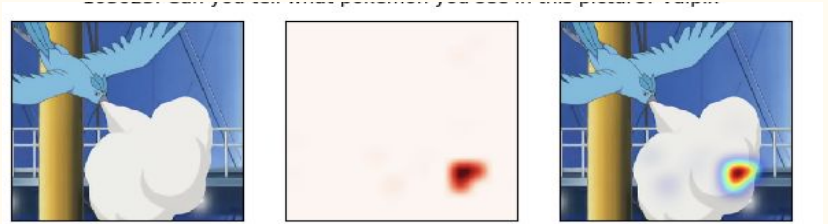


Batch Size	256
Learning Rate	0.001
Number of Iterations	800
Number of Attention Layers	2
Word Embedding Size	200

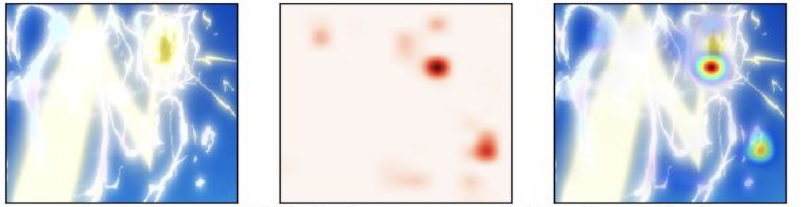
Results



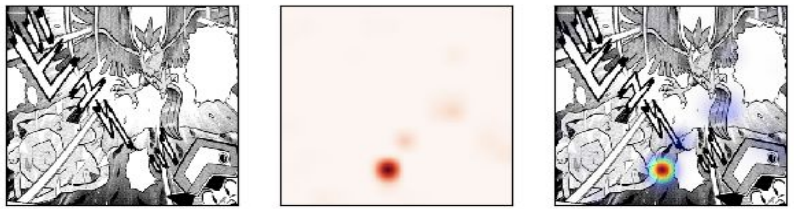
108887: What attack is being performed by the pokemon? flamethrower



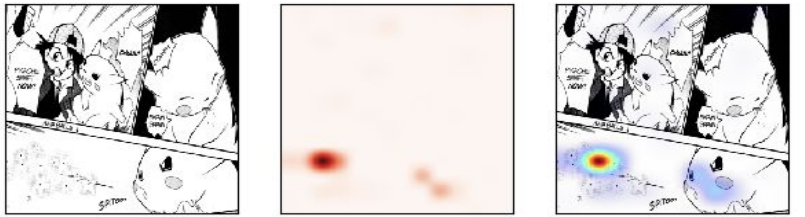
113427: What attack is being performed by the pokemon? mist



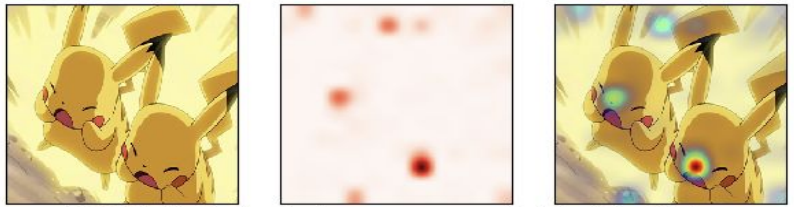
102143: Can you tell what pokemon you see in this picture? pikachu



1135520: What color of pokemon can you spot in this picture? Blue



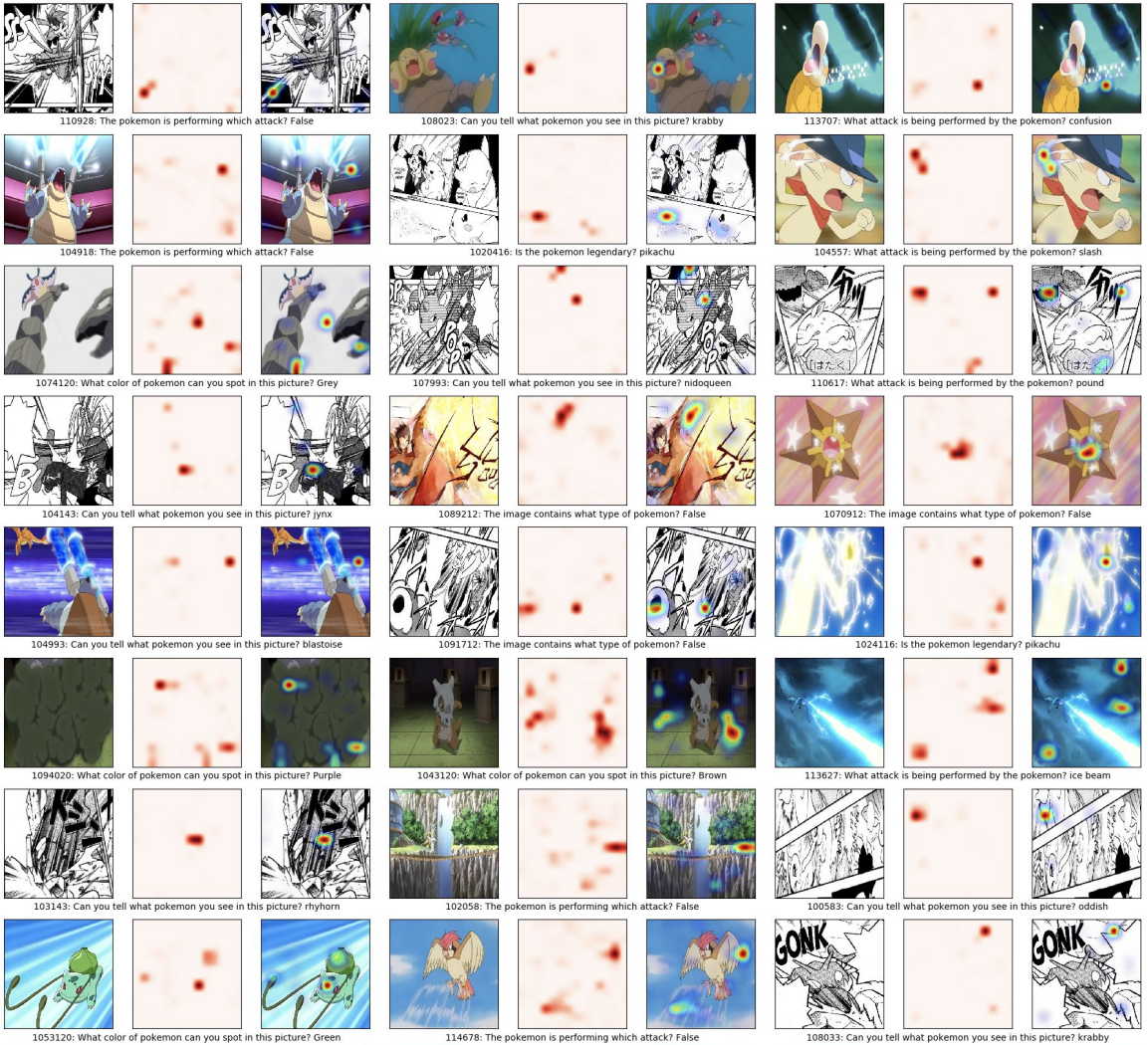
1020416: Is the pokemon legendary? pikachu



1022320: What color of pokemon can you spot in this picture? Yellow

Results

Accuracy: 65.9%



Demo!

Limitations and Future Work

- We had to constrain our number of questions due to less visual content in the image. One of the advanced topics we look forward to working on involves predicting the outcome of a battle by looking at the battle stage in the image and some additional prior about the Pokemon weaknesses and strengths.

- We could only form questions having a fixed answer, though in real world (in the pokemon world too!) there can be multiple correct answers to a single question.

References

[1]Jiang, Yu, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. "Pythia v0. 1: The winning entry to the vqa challenge 2018." arXiv preprint arXiv:1807.09956 (2018).

[2]Anderson, Peter, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. "Bottom-up and top-down attention for image captioning and visual question answering." In CVPR, vol. 3, no. 5, p. 6. 2018.

[3]Uijlings, J.R.R., van de Sande, K.E.A., Gevers, T. et al. Int J Comput Vis (2013) 104: 154. https://doi.org/10.1007/s11263-013-0620-5

[4]Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich; The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1-9

[5]Danna Gurari and Kristen Grauman. 2017. CrowdVerge: Predicting If People Will Agree on the Answer to a Visual Question. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17). ACM, New York, NY, USA, 3511-3522. DOI: https://doi.org/10.1145/3025453.3025781

[6]https://github.com/abhshkdz/neural-vqa-attention

[7]https://github.com/iamaaditya/VQA_Keras

[8]https://github.com/anantzoid/VQA-Keras-Visual-Question-Answering