Using Attention for explanation

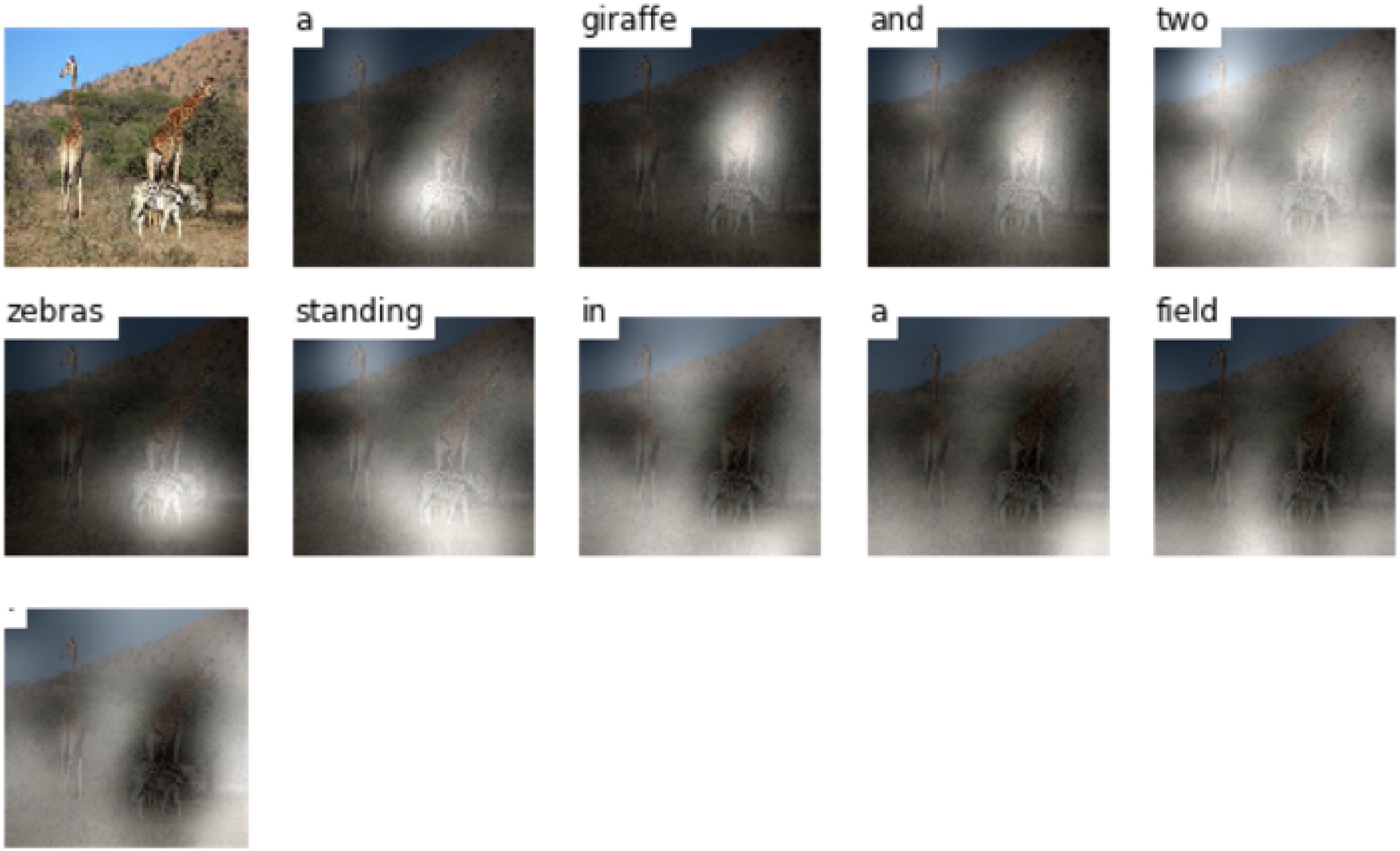
We briefly introduced the Attention mechanism in a previous lecture (Intro_to_Attention.ipynb).

When the output of a task is a sequence and the input is large

- Either a long sequence, or high spatial dimension
- It is usually the case that the predicted value for each time step
- Is a function of a *restricted* subset of the input
- That changes with the step

We gave the example of Image Captioning

- The input is an Image
- The output is a sequence of words describing the image
- We are attending over the spatial dimension
    - Which part of the image do we focus on when generating the next word in the caption



Attribution: [https://arxiv.org/abs/1502.03044 (https://arxiv.org/abs/1502.03044)](https://arxiv.org/abs/1502.03044)

The Show, Attend, and Tell paper (https://arxiv.org/pdf/1502.03044.pdf) paper is a good reference.

In that paper, they suggest using Attention in order to diagnose errors.

- What part of the image is being attended to
- When an incorrect word is generated

Image captioning example: diagnose bad caption

- Source: Image
- Target: Caption: "A large white **bird** standing a forest."
- Attending over spatial dimension (*pixels*) **not** temporal sequence

Visual attention

A large white **bird** standing a forest.

Attribution: [https://arxiv.org/pdf/1502.03044.pdf (https://arxiv.org/pdf/1502.03044.pdf)](https://arxiv.org/pdf/1502.03044.pdf)

Image captioning example: diagnose bad caption

- Source: Image
- Target: Caption: "A man is talking on his cell **phone** while another man watches."
- Attending over *pixels* **not** sequence

Visual attention

A man is talking on his cell **phone** while another man watches.

Attribution: https://arxiv.org/pdf/1502.03044.pdf (https://arxiv.org/pdf/1502.03044.pdf)

Perhaps this technique will prove useful in improving the understanding of how the Neural Network works.

Conclusion

Attention is a mechanism to "narrow the focus"

- From a large input
- To a particular region that is most important at a given moment

We speculate that this mechanism might be valuable for understand and interpreting the behavior of Neural Networks.

```
In [4]: print("Done")
```

```
Done
```