How does a Deep Learning Classifier work ?

We are able to construct Neural Networks that are quite successful at many tasks.

But it is still somewhat of a mystery as to how they are able to achieve this success.

In this lecture (which we previewed in an earlier lecture) we will try to motivate the search for Interpretability.

We will illustrate the issues using a state-of-the-art model for Image Classification.

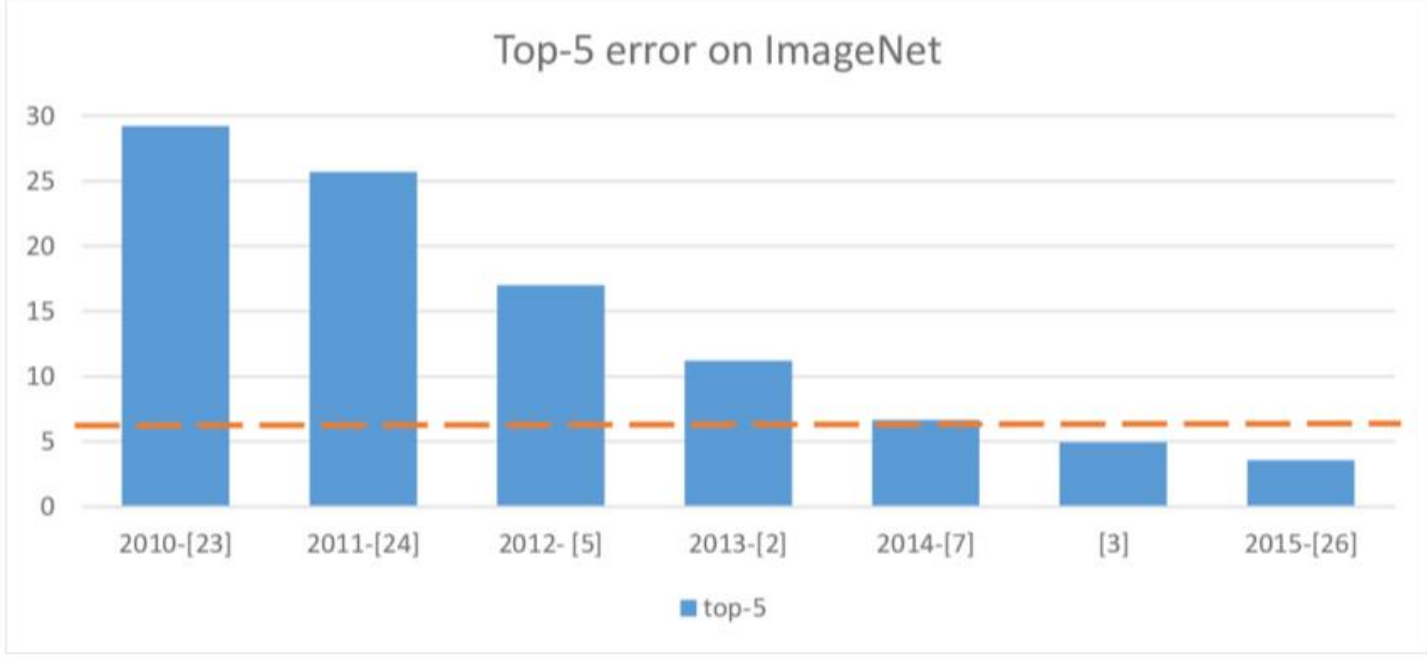
This was a winning model in an earlier competition called ImageNet that contributed greatly to the advance of Deep Learning.

ImageNet was a contest held annually

- Objective: Correctly classify images
- Training data: hand-labeled images
    - 1.2 million images over 1,000 classes
        - 200 classes of dogs and cats !
    - Subset of a larger set of 14 million images, from 22,000 classes
- Predated the Deep Learning revolution

Here is the Percent Error Rate of the winning entry over time.

Deep Learning Revolution



Chennupati: https://www.researchgate.net/figure/Shows-the-progress-of-classification-performance-top-5-error-on-Imagenet-dataset-over_fig27_312935261

After several years of small decline in error rates

- There was an unexpectedly large drop in 2012
- So large that the judges thought they made an error in evaluation !

This was the first year that a Deep Learning model was submitted

- It transformed Image Classification into an "easy" problem
- It catalyzed the current Deep Learning Revolution

As you can see, the winners in subsequent years (also Deep Learning models) continued to improve.

The dashed horizontal line is *human performance* on the task.

Classifiers with error rates belong this line are said to exhibit *super-human* performance.

Models behaving badly (or at least, unexpectedly)

Let's test some theories as to how this highly accurate model classifies images.

Here is the classifier's response to a cat image:

High confidence.

How does the classifier "recognize" this as a "tiger cat" ?

Maybe: by it's parts ?

How does it work: Parts ?

Perhaps by the parts, but certainly **not** by the parts in correct spatial order !

This may be due to the use of Convolutional Neural Networks (CNN) in many models

- The filter is narrow so only local spatial relationships are captured
- Some layer types (e.g., Global Pooling) discard spatial relationships
- Only preserve "Is present/absent" property for a feature
- **Not** if two features (like eyes) are adjacent

Assuming that the parts are important

Which parts of the cat are the ones that most contribute to the classification ?

Perhaps the ears or tail ?

How does it work: Parts ?

Probably not. Covering up (occluding) various parts still results in correct classification.

Maybe it's the *shape* thats important ?

How does it work: Shape ?

Probably not.

Maybe: texture ?

How does it work: Texture ?

Perhaps it's the texture.

Conclusion

We want to know how a highly accurate classifier is able to work its magic.

By conducting a couple of simple experiments, we hope to have peaked your interest into the problem of interpretability of Deep Learning models.

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

```
Done
```