**Describe the ImageNet dataset.**The imagenet dataset is a dataset of 15 million images with over 22000 categories that is used to benchmark visual recognition models. Each year they hold a competition with images and see which models can do the best. The original imagenet dataset consisted of 1,000,000 images from a 1000 different objects, with 1000 in each class.

**Explain what a ‘top-k error rate’ means. What error rates did AlexNet achieve in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2010 and 2012?**

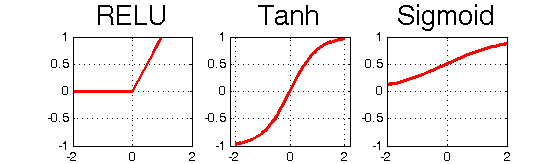
To understand top k error rate one must first understand what the model outputs. The model is outputting a 1000 class probability distribution. Top k error rate is asking the question: is the correct label in the top k highest probabilities from the outputted distribution of the model. Quoting the paper: “… the top-5 error rate is the fraction of test images for which the correct label is not among the five labels considered the most probable by the model”. Top 1 error rate is simply how many times the model predicts the correct label as not the highest probability.

**What is the total number of model parameters for AlexNet? Comment on the effect of the number of parameters on overfitting.**   
  
The neural network architecture used in the paper has 60 million parameters. Since there are 1000 classes, “each training example imposes 10 bits of constraint on the mapping form image to label.” Since there are 60 million parameters 10 bits of constraint is insufficient to learn all the parameters without overfitting.

**What strategies did the authors employ for mitigating the overfitting problem? Mention the specific hyperparameter values used, if any.**   
There were multiple techniques used for combatting overfitting. One technique used was data augmentation. They would take 224x224 patches of the image and horizontally reflect them. The second form of data augmentation used was altering the intensities of the RGB channels in the training images. They do this by applying PCA to the images in the ImageNet training set. They then added the found principal components to each sample with some gaussian noise added to it. Dropout was also used. This was used in the fully connected layers and was done at a rate of 0.5. They also used overlapping pooling, which is setting the stride to lower than the pooling kernel in the pooling layers, which reportedly also decreased overfitting slightly.

**Describe the architecture of the AlexNet model. List the types of layers and the corresponding number for each type.**  
Alexnet consists of 5 convolutional layers followed by three fully connected dense layers. (Would these be called Dropout in keras since they were trained with Dropout?)

**What activation function does AlexNet use? Elaborate on the effect of the choice of this function on the training time. Be specific.**  
  
AlexNet uses ReLu, which stands for rectified linear units. They use this, because it trains faster than the typical neuron modeling function tanh. They show that it can reach 25% training error on the CIFAR-10 dataset, 6 times faster than its tanh counterpart. Part of the reason why it trains faster is because it does not require input normalization to avoid saturation (getting stuck at extreme values). Other activation functions can sometimes run into the vanishing gradient problem, which happens when a neuron becomes saturated, which is not the case for ReLu.



**Personal Reflection:**  
AlexNet was trained with a vast amount of labeled images, however in many real world cases it is not possible to gather large amounts of labeled images. I wanted to look into alternatives one can use when large amounts of labeled images are not available. An example of this is when it comes to medical data, especially new and rare diseases or labeling data requires an expert. This is where transfer learning comes in.   
Table

Description automatically generated

Since 2012, there have been numerous predecessors to AlexNet. Since these models were all trained on ImageNet they all have a vast amount of knowledge encoded into their layers. We can then use this knowledge (the weights) in other models to save on training time and produce better results. Some general guidelines that I have read for transfer learning for image classification are: deeper networks tend to lead to better performance. When the source and target dataset are very alike, transferring more layers leads to better results, however if the source and target dataset are less alike using less layers is better. Fine tuning on the target dataset almost always leads to better performance, especially when the target dataset is more different and larger. We can use these benchmark models layers, in new image classification tasks to achieve state of the art results.

ChatGPT use:  
ChatGPT was used very lightly to clarify certain terms and sentences in the paper. For example why does ReLu train faster than tanh.

Links:  
[(PDF) Deep Learning and Transfer Learning Approaches for Image Classification (researchgate.net)](https://www.researchgate.net/publication/333666150_Deep_Learning_and_Transfer_Learning_Approaches_for_Image_Classification)  
[2205.09904.pdf (arxiv.org)](https://arxiv.org/pdf/2205.09904.pdf) (Deep transfer learning for image classification: a survey)