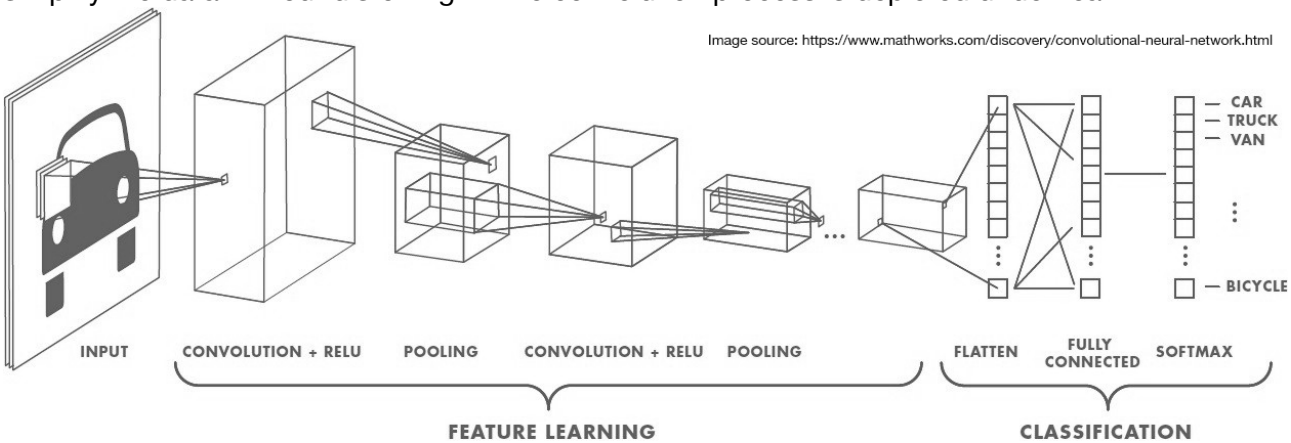
March 4th 2018
Jonas Søren Bjerg
jonassorenbjerg@gmail.com

Neural Networks potential impact on the film industry

Nowadays almost every motion picture utilises green screen technologies. Masking (removing) pixels with a certain colour limits the production process. It hinders the use of green colours elsewhere, renders certain camera angles problematic and confines production to locations with optimal lighting. Could Artificial Intelligence and Neural Networks improve this process?

Convolutional Neural Networks (CNN) have proven exceptional at handling image classification and object detection problems. The network type was the most popular approach among contestants of the "Carvana Image Masking Challenge" held by Kaggle[1] in September 2017. Every team in top 493 had an accuracy score above 0.99 (above 99% accuracy on images similar to the dataset). The popular open source website GreenScreen.ai[2] utilises the same network type for automated image background removal. But how does this type of neural network function?

In essence a CNN consists of neurons stacked on top of each other in different layers. Using a standard multi-layer perceptron[3] with all the layers fully connected to each other, it would be almost computationally impossible to assign a neuron to every pixel. It is therefore necessary to simplify the data without distorting it. The convolution process is depicted underneath.

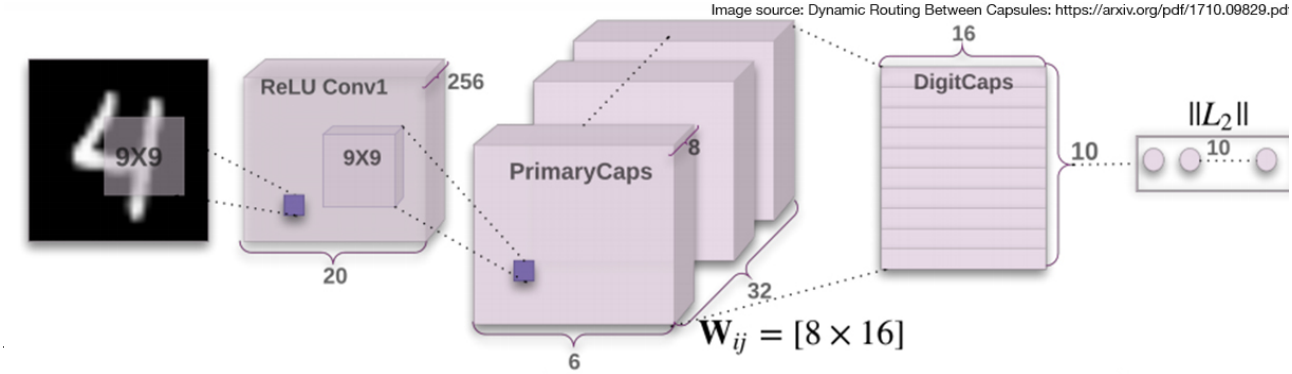


Convolution[4] uses matrix multiplication[5] to sort pixels into groups and surmise the results. This process is initially done by filters (or kernels) which scan the input and create feature maps. The features are then passed through an activation layer which induce a sense of non linearity to the network (e.g. ReLU and The Sigmoid Function). The features are then passed through a pooling layer (e.g. MaxPool), a function that progressively reduces its spatial size, necessary parameters

1 https://www.kaggle.com/c/carvana-image-masking-challenge#description
2 https://greenscreen-ai.boorgle.com/
3 An MLP is a feedforward artificial neural network consisting of at least three layers
4 A mathematical operation used to combine two functions into one joined function
5 http://mathworld.wolfram.com/MatrixMultiplication.html

and computations by summarising subsections of the data. MaxPool has often been described as the messenger between two layers, it helps the network from overfitting[6] the data and provides it with positional and translational invariance in object recognition. These steps are then repeated as many times as necessary before finally passing the features through a classifier (e.g. Softmax) which assigns the image a class. When training the network "Backward propagation of error" is used. It is an algorithm for supervised learning[7] that utilises gradient descent to test results against a pre-labelled data set.

CNN's have difficulties with tilted or rotated images. The networks distortion tolerance can be improved by including images with aspect variations in the training set. This will make the network invariant to distortions, but it will not make it understand the distortions. For true understanding translational equivariance[8] is necessary. This is where Capsule Networks (CapsNet) flourish. Sara Sabour, Nicholas Frost and Geoffrey Hinton[9], published a paper called "Dynamic Routing Between Capsules"[10] in October 2017 revealing CapsNet and explaining it as a three dimensional neural network. Arguing that a human brains have modules the paper calls "capsules" and that humans learn and analyse visual information hierarchically starting with boundaries and colours before moving on to shapes and entities. The network outperformed CNNs on the labelled handwritten numbers dataset MNIST[11]. MNIST has almost become the industry baseline of which handwriting recognition is evaluated. CapsNet especially showed improvements with images of overlapping numbers. Proving that mimicking the human brain by implementing capsules (a nested set of neural layers within a neural layer) significantly improved the networks translational equivariance and understanding. The basic CapsNet architecture is depicted underneath.



[6] Overfitting is when the network's analysis corresponds too closely or exactly to the dataset
[7] https://brilliant.org/wiki/backpropagation/
[8] The ability to identify the position of one object relative to another
[9] Theres a second paper currently under review for ICLR 2018 called "Matrix Capsules with EM Routing"
[10] Link to full paper: https://arxiv.org/abs/1710.09829
[11] MNIST is a database of tagged handwritten digit images

March 4th 2018
Jonas Søren Bjerg
jonassorenbjerg@gmail.com

The neurons in a capsule captures the properties of one entity in an image and outputs it as a vector. The vector is then sent to the parents in the capsule who calculates a vector prediction by multiplying its own weight with the weight matrix. The parent with the largest scalar vector prediction product increases the capsule bond and influences the end result. This method is called “routing by agreement” and is superior to MaxPool, according to Hinton, because MaxPool only routes based on the strongest feature detected in the previous layer. Hinton introduced a new squashing function ($v_j = ||s_j||^2/(1+||s_j||^2)*(s_j/||s_j||)$) that gets applied to the vector output of each capsule. Its purpose is to squash the small vectors to 0 and limit the output vector to ≤1.

This becomes interesting for camera production because of Hinton’s inverse graphics approach. Hinton states that brains looking at a 2D image of a 3D object estimates the cameras viewpoint, rendering it able to distinguish background from foreground, which is the most crucial step in background removal. CNNs are still the primary network used publicly for background removal. This will assumably shift to CapsNets in the future because of its superior approach to translational equivariance, which is especially impactful on pixel based data. Pixels can change dramatically when distorted or the viewpoint is changed, but the pose matrix, which CapsNet utilises, changes linearly. Meaning data (e.g. locational data for subsection within an image) is kept better intact with the CapsNet method than with the convolutional method. In both approaches the training of the networks include brute forcing the learning process by supplying vast amounts of data. This process could be further improved by providing the network with an initial algorithm for which pixels make out the background. This algorithm has existed for decades, yet nobody is including it in the learning process. It is called difference keying and it is best explained with an example. Imagine a short video of moving objects recorded with a stationary camera. The algorithm generates a mask by taking the average of every pixel in frame. This mask will be a still image of every object that has not moved in the video. Meaning people, trees and cars have all been removed. This mask can then be used to identify which pixels are part of the background. Supplying the network with this additional dataset can improve the training process significantly. It provides meaningful data which can help the networks understand issues like transparent surfaces and motion blur both of which are issues current neural networks struggle to understand.

March 4th 2018
Jonas Søren Bjerg
jonassorenbjerg@gmail.com

Conclusion:

CapsNet is a new concept that has not yet been thoroughly tested. MNIST is a decent dataset, but the concept needs to be tried on bigger and more diverse datasets before anything can be said with certainty. There are a few concerns already being raised, my main concern being the use of cosine to measure agreement between two vectors. Cosine does not leave room for interpretation between “good” and “very good” agreement, which can make a big difference when training a network. That being said, CapsNet still looks immensely capable especially when working with pixel based data. It utilises the same aspects that gave CNNs an upper hand in image based problems while at the same time taking a new approach. Adding the “difference key”-mask to the training set could potentially reduce training time significantly while also enabling the network to move away from stationary recordings. CapsNet understands objects position in the frame and estimates the viewpoint. Adding the mask of the stationary recordings can potentially enable the network to put two and two together and create background removal in real time. The training process takes up most of the necessary computing power. Once the network is trained it uses so little computing power it could potentially be run inside the camera itself enabling alpha based recording without green screens.

References:

https://arxiv.org/pdf/1703.03872.pdf
https://arxiv.org/pdf/1702.01731.pdf
https://www.kaggle.com/c/carvana-image-masking-challenge/leaderboard
http://cs231n.github.io/convolutional-networks/#pool
http://ufldl.stanford.edu/tutorial/supervised/ExerciseConvolutionAndPooling/
https://openreview.net/pdf?id=HJWLfGWRb
https://kndrck.co/posts/capsule_networks_explained/
https://hackernoon.com/what-is-a-capsnet-or-capsule-network-2bfbe48769cc
https://saoyan.github.io/DL-background-subtraction-1/
https://hackernoon.com/uncovering-the-intuition-behind-capsule-networks-and-inverse-graphics-part-i-7412d121798d
https://forums.creativecow.net/thread/244/861773