Solar spots classification using pre-processing and deep learning image techniques[⋆]

Thiago O. Camargo[1 [0000−0002−6917−8339]], Sthefanie Monica Premebida[1 [0000−0002−8964−2119]], Denise Pechebovicz[1 [0000−0002−1953−0533]], Vinicios R. Soares[1 [0000−0002−3565−6646]], Marcella Martins[1 [0000−0002−5716−4968]], Virginia Baroncini[1 [0000−0001−6512−3958]], Hugo Siqueira[1 [0000−0002−1278−4602]], and Diego Oliva[2 [0000−0001−8781−7993]]

[1] Federal University of Technology - Paraná - Ponta Grossa - (UTFPR-PG)
[2] Universidad de Guadalajara,CUCEI Guadalajara, Mexico
marcella@utfpr.edu.br

Abstract. Machine learning techniques and image processing have been successfully applied in many research fields. Astronomy and Astrophysics are some of these areas. In this work, we apply machine learning techniques in a new approach to classify and characterize solar spots which appear on the solar photosphere which express intense magnetic fields, and these magnetic fields present significant effects on Earth. In our experiments we consider images from Helioseismic and Magnetic Imager(HMI) in IntensitygramFlat format. We apply pre-processing techniques to recognize and count the groups of sunspots for further classification. Besides, we investigate the performance of the CNN AlexNet layer input in comparison with the Radial Basis Function Network (RBF) using different levels and combining both networks approaches. The results show that when the CNN uses the RBF to identify and classify sunspots from image processing, its performance is higher than when only CNN is used.

Keywords: Image Processing · Astronomy and Astrophysics · Neural Network.

1 Introduction

Eruptions on solar surface are correlated with solar spots and there is a probability of an event be evaluated based on the area, class and the lifetime of the spot [1]. In this sense, Coronal Mass Ejections (CME) present significant effects on Earth civilization [2]. However, the prediction of Flares and these effects are still difficult to be performed [3]. Coronal Mass Ejections is a release of huge quantity of plasma enclosed with magnetic field of the Solar Corona, Flares consist in a light flash, occurs near a solar spot and often are followed by a CME.

The sun presents areas of shear or interfacial layers forming the dynamo, which generates a main magnetic field whose movement stretches and twists the

[⋆]

existed magnetic field lines through the solar poles, forcing the poles to writhe and create bulbs, due to differential sun rotation. When this action creates bulbs, they behave as local magnetic fields in the photosphere, with their own north and south poles. These bulbs appear in the photosphere in the forms of ”loops”, lumps, filaments (which are not more than one protrusion view with solar bottom surface) and the sunspots.

These patches are colder regions with temperature around $4100K$, and darker than the photosphere. It is formed by a core, Umbra, which is the darkest part of the spot size of 300 to $2500km$, and greater strength magnetic field around 2000 to 2500 $\mathrm{Mx.cm}^{-2}$, being more vertical. The Penumbra appears that around 50% of the spots, being a surrounding area of Umbra, about the size of 2.5 times the size of it, which has a gray scale and field strength of about 500 and 2000 $\mathrm{Mx.cm}^{-2}$ more horizontal [4, 5]. Currently, most benchmarks to measure solar activity consider the number of sunspots present on the sun at any given time. We highlight that it is very important the noting, counting and the classification of them. However, there is a few studies on different image processing for easy sorting and counting sunspots.

After the solar cycle of 11.2 years, the entire sun reverses its overall magnetic polarity: the north magnetic pole becomes the south pole, and vice versa [6]. Thus, a complete magnetic solar cycle lasts on average about 22 years, being known as Hale cycle, but the behavior varies with the variation of the activity. According to Hathaway [7], observations of sunspots and solar activity from the middle of XVII showed that the number of sunspots and the area they cover grow rapidly from a minimum (close to zero) to the maximum (3 to 4 years after reaching the minimum). However, the maximum decline to minimum is slower. This asymmetric growth and decline exhibit substantial variations from one cycle to another. This non-linear and chaotic behavior suggest that the dynamo is not only a oscillating phenomenon due to is possible to observe the solar Hale cycle [5].

Since 1981 the analysis of images provided by satellites and observatories from sun, for automated monitoring, have been done by the Solar Influences Data Analysis Center (SIDC), which has been producing monthly the International Relative Sunspot Number, Ri, calculated statistically from all contributors and available observations, using the Wolf number [8]. In order to equalize the data to find a consistent Ri, it is used the personal reduction coefficient k, which is the factor scale between the individual station and the overall network average [8].

This paper aims to extend the work presented in [COLCACI PAPER] by investigating a comparison with some networks, presenting these main characteristics and providing quality metrics for both approach not explored in [COLCACI PAPER]. The main objectives here are to identify and classify sunspots from image processing for further being explored to measure the solar activity. For this purpose, we use the numerical communication software analysis and data visualization, MatLab, besides a pretrained convolutional neural network (AlexNet) and a Radial Basis Function (RBF) network. Similar works have addressed sunspots on a image processing context, especially using computational vision, but here, in our proposal, we aim to apply machine learning techniques which can be further explored within the graphics processing approaches.

This work is organized as follow: Section 2 presents the background and the related investigations of images processing techniques. The Section 3 discusses the proposed approach to achieve the goal, while Section 4 shows the experiments and the computational results. Section 5 presents the conclusions from the presents results.

2 Background

Currently, sunspots are the main references to determine the level of solar activity. Besides, the captured images of the sun are the basis of several studies to develop theories and better understanding of the star. The quality, the good use and ease of observation from images is essential and, to improve these characteristics, pre-processing techniques can be addressed. This section presents a background for pre-processing and machine learning techniques usually applied on image manipulation in a general context.

2.1 Pre-processing techniques

When manipulating images, some techniques should be applied before their complete processing. These techniques can help to make a data optimization: filling nulls, treating noise, identifying or removing outliers, and removing inconsistencies; integrating data; processing and data reduction with particular value for numeric data; normalization and aggregation; discretization of the data [9] . We list some methods to perform this processing as follows:

- Image adjust (stretching)
- RGB to gray
- Image to black and white
- Image open (growth)
- Image complement
- Region proprieties
- Image crop
- Image write

2.2 Machine Learning

The Machine Learning techniques can be used in several automated situations because they can produce quickly and automatically models able to analyze larger and more complex data, and deliver faster and accurate results, even in a large scale[10]. Machine Learning is a part of Artificial Intelligence (AI), and present 4 major groups of approaches: supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning, as seen in Figure 1 .

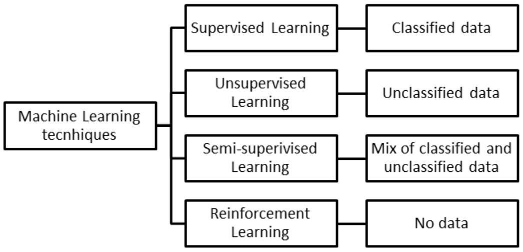


Fig. 1: Machine Learning techniques.

- Supervised Learning: in these cases, the machine learns from a set of labeled data. In this case, the inputs and the respective outputs are known and the reference signal acts as a teacher, guiding the adjust of the parameters in the training phase. It is of two types, classification and regression. In the first, we have the input and tune the model to predict the output. In a regression problem, we have the output and try to find the possible inputs. Neural Network (NN) can be applied as supervised learning, and convolutional neural network (CNN) is a kind of deep neural network which presents architecture based in a multi-stage processing of an input image, generating a high-level and hierarchical features [11]. This process is fully automated, discarding external and manual classification data.
- Unsupervised Learning: in this case, there is no teacher to guide the classification procedure. The main methods of this kind is the clustering algorithms, which is a set of data mining models, that automatic grouping of data according to their similarities, like human facial expressions [12], without prior training as in the last case.
- Semi-Supervised Learning: this type of method is like a mixture of the firsts, using supervised and unsupervised procedures, being the closest thing to our way of learning [13]. Here, program learns from its own mistakes and improves its performance, that is, in addition to working to obtain the results, it the model make an analysis of future test data much better than the model previously generated.
- Reinforcement learning: these methods use observations obtained from interaction with the environment to perform actions that maximize reward or minimize risk [14]. The Markov Decision Process (MDP) [15] is an example of this typo, which is a stochastic process in which the next state depends on the current state.

  The MDP present 5 values:

  1) Finite set of states (S), such as a door being open or closed;

  2) Finite set of actions (A), what possible actions can be taken;

  3) Probability model (P), the probability of a current action taking the problem to an action in the future state;

  4) Reward (R) is a value that depends on which state you are in. Rewards can be defined for states, even without actions taken;

  5) Discount factor (Y), usually between 0 and 1. This factor influences the total future reward that the Agent will receive, ie if there is a discount factor of 0.9 we know that the more advanced the Agent is, the greater your reward.

In this paper, the classification of the spots groups has been done with CNN and with Radial basis function network. Convolutional Neural Networks is one of the most know models deep learning, and also one of the most utilized. The network is composed of convolutional layers that process the inputs, normally images, because in your convolution we can filter pictures considerating their spatial structure. Most layers of a CNN are pre-trained, but the last ones are trained on a image store according the user needs. This feature turns this deep learn model very accurate.

3 Proposed Approach

This works presents an approach to classify the sun images according to the spots and some special features. We address a dataset taken from the Helioseismic and Magnetic Imager (HMI), which processes images of the Solar Dynamics Observatory (SDO).

The classification is performed according to two stages: i) a pre-processing and ii) a training phase, the second stage is applied to the CNN and the Radial Basis Function Network.

First, the images have been submitted to pre-processing techniques, such as stretching, threshold, object properties analysis and cropping. An image example can be shown on Figure 2.

After the preprocessing, the training phase uses two spots groups: one with positive for O type and the another one with negative. The O type is arbitrary, defined to simplify the analysis: it is a class basically characterized by almost only penumbra. A comparison is presented in Figure 3. The criterion used here is based on the amount of black pixels for each subfigure. All of those methods are presented in Algorithm 1.

The method utilized to reformulate AlexNet to ours purpose is transfer learning.

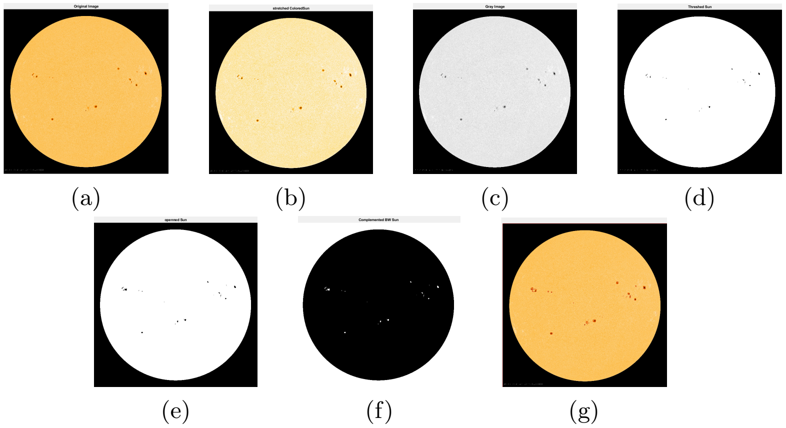
3.1 Preprocessing Stage

A folder with HMI images, such as the example on Figure 2a, is created in Step 1. All these images are loaded in an array and each element is applied to algorithm 1. In Step 3 the image is *stretched*, as shown in Figure 2b, therefore, the contrast between the spots and the solar surface increases, and the *gray scale*, in Step 4, process a color change from RGB to a gray scale, (see Figure 2c). This can facilitate the *thresholding* in Step 5.

The thresholding turns black the region near the group centroid (Figure 2d), but in many cases the black object created shown itself divided in many small centroids. For this reason the *growth* of these objects is necessary, Step 6, (Figure 2e). This action decreases the number of objects, which can be showed as an advantage, however, some frames with different views of the same group make the neural network learning more accurate.

Algorithm 1 A simplified pseudo-code presenting the main components of pre-processing and sorting

```
INPUT: $I$: HMI Intensitygram Flat images
OUTPUT: $O_\text{p}$: image dataset with positive to $O$ type
        $O_\text{n}$: image dataset with negative to $O$ type
  {Initialization}
  $F \leftarrow$ Load All HMI Images
  {Main loop}
  for each image $\in F$ do
    {Treatment and Filters}
    $I_\text{sc} \leftarrow$ stretching $(F_\text{g})$
    $I_\text{gr} \leftarrow$ gray scale$(I_\text{sc})$
    $I_\text{bw} \leftarrow$ thresholding$(I_\text{gr})$
    $I_\text{op} \leftarrow$ oppening$(I_\text{bw})$
    $I_\text{cbw} \leftarrow$ complementing$(I_\text{bw})$
    {Spot Detection}
    $P_\text{sp} \leftarrow$ coordinates of spots groups$(I_\text{cbw})$
    $N_\text{ob} \leftarrow$ number of objects$(I_\text{cbw})$
    {Image Cropping}
    $S_\text{ian} \leftarrow$ size of the cropped images for AlexNet
    $z \leftarrow 1$
    for each coordinate $\in N_\text{ob}$ do
      $I^z_\text{cr} \leftarrow$Cut Image( $S_\text{ian}$,$F_\text{g}$, $N^z_\text{ob}$)
      $I_\text{testgr} \leftarrow$ gray scale$(I^z_\text{cr})$
      $I_\text{testbw} \leftarrow$ thresholding$(I_\text{testgr})$
      {Image Testing}
      if $I_\text{testbw} = BlackImage$ then
        discard $I^z_\text{cr}$
      else if $I_\text{testbw} = WhiteImage$ then
        discard $I^z_\text{cr}$
      else
        {Image Saving}
        $N_\text{bp} \leftarrow$ number of black pixels$(I_\text{testbw})$
        if  $N_\text{bp} \in OTypeParameter$ then
          $O_\text{p} \leftarrow I^z_\text{cr}$
        else
          $O_\text{n} \leftarrow I^z_\text{cr}$
        end if
      end if
      $z \leftarrow z + 1$
    end for
  end for
```



(a) (b) (c) (d)

(e) (f) (g)

Fig. 2: Example of preprocessing stage using techniques for a) Original Image, b) Stretched, c) Gray Scale, d) Threshold, e) Grown, f) Complemented and g) Detected. This image is from May 15th, 2014, at 00:00 hours.

For the object detection the background must be black and the object white. This condition is limited by the function *regionprops* from MatLab, (Figure 2f). Therefore, Step 7, the *complementing* of the image is gotten. figure 2f.

The centroid coordinates gives the parameters to crop the original image in subfigures with $227x227$, Step 9, size to fit as an AlexNet input. An example of spot groups detection is presented in Figure 2g. Thereafter, in Steps 12 to 20, a staining test is performed to ensure the validity of the figure, black and white images are rejected.

Another test is done to separate O type positive and negative in Step 20 to 27. The test is applied using the total of black pixels on the image, and a range is defined to classify the O type positive. Subfigures which do not fit in these range are classified as O type negative.

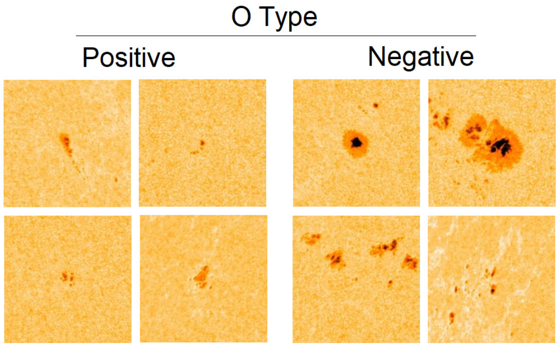


Fig. 3: Comparison between O type positive and negative.

3.2 Training and Testing

An image dataset has been create with the two folders of the groups of O type separated by Algorithm 1 The image dataset is randomly divided into two arrays, one to train both networks, containing 70% percent of the sub-figures, 15% to validation and the other containing 15% to test.

Convolutional Neural Network The pretrained CNN AlexNet was loaded and the three lasts layers have been discarded in our proposal.

The fully connected layer contains the number of classes, weight learn rate factor and bias learn rate factor. Softmax layer, originally is responsible to gives the percent of classification of 1000 classes [11], but in this paper is reformulated to process only two classes. The last layer is the classification output layer.

The AlexNet layers parameters has been set to keep the features in the predefined layers with a small value of initial learn rate. In the fully connected layer the learning rate has been increased. The function responsible to train the CNN validates the network according to a frequency during training, and automatically stops training when no more improvement is achieved.

After the configurations the CNN is trained, then used to classify the validation/test data.

The investigation has been made computing the precision, recall and accuracy coefficients. [16]. Control variables has been created to count the quantity of true positives, T_{positive}, true negatives, $T_{\text{negatives}}$, false positives, F_{positive}, and the false negatives, F_{negative}, the counting process consist in decision structures comparing the labels gave by the trained network and the labels of the test part of the image data store.

Radial Basis Function Network The input data for the RBF network was got from three different layers of AlexNet, data, which gives the normalization of the images, fully connected layer 6 and 8. These different inputs creates a different kind of neural network, with AlexNet fully connected layers as the RBF input the neural network can be described as a new CNN, because the SoftMax layer is no longer used, the method was replace with hyperbolic tangent method and also a different kind of SoftMax method. For the data layer as input the RBF works entirely with their method.

The Radial Basis Function network parameters was formulated to be similar to AlexNet parameters, like the network momentum, the both have 0.9, learning rate, the both have 0.0001, basically all the parameters that the two networks have in common are the same, except for the neuron quantity and the max training epochs. The neuron quantity is different according with RBF input, for the fully connected layers the neurons quantity vary between 5 and 100, for the AlexNet data layer the network have only 100 neuron, this fact is limited by computational time. The max training epochs in RBF is 500, but this value was never reached.

4 Experiments and Results

In total, 96 images in the HMI Intensitygram Flat format were analyzed, of which 920 subfigures were generated, 138 to validation, another 138 to testing and 644 to the training phase. Except for AlexNet data layer, the quantity of images had to be less than the total, this reduction was necessary because our processing machine could not handle, therefore, the total images was reduced to 400. We use Intensitygram Flat Orange 4K images [3]. This format allows the easy manipulation for counting, identifying and classifying sunspots from image processing.

We have investigated the neural networks performance calculating three coefficients, precision, recall and accuracy. All of these depends of some components, results of the networks prediction. True positive(TP), this component is given when there is O type positive on the subfigure and the NN detect it. False positive(FP), when there is O type negative on the subfigure and the NN classify as O type positive. True negative(TN), is given by the detection of O type negative correctly. False negative(FN), this component is given when the NN classify a subfigure containing O type positive as O type negative.

The Recall coefficient can be described as the capacity of the NN classify correctly:

$$\frac{\sum TP}{\sum TP + \sum FN} \tag{1}$$

Precision coefficient gives the proportion of correctly classification of a true positive:

$$\frac{\sum TP}{\sum TP + \sum FP} \tag{2}$$

Accuracy coefficient determine the fraction of right classifications:

$$\frac{\sum TP + \sum TN}{\sum TP + \sum FP + \sum TN + \sum FN} \tag{3}$$

The Table 1 presents all the components and the calculated coefficients for AlexNet performance.

Table 1: Performance analysis.

True Positives	26
True Negatives	96
False Positives	9
False Negatives	5
Recall	0.8710
Precision	0.7714
Accuracy	0.9130

The Table 2 shows the best and worst calculated coefficients for RBF with AlexNet layers as input.

[3] The images can be downloaded from http://hmi.stanford.edu/.

Table 2: Performance comparison.

	Recall		Precision		Accuracy	
	Best	Worst	Best	Worst	Best	Worst
FC6-SoftMax	1.0000	0.4909	0.939	0.3925	0.9333	0.9333
FC6- Hyperbolic Tangent	1.0000	0.9692	0.9482	0.7393	0.9619	0.9619
FC8 - SoftMax	0.9732	0.4722	0.8925	0.4305	0.9167	0.9167
FC8- Hyperbolic Tangent	1.0000	0.9833	0.9638	0.8376	0.9381	0.9381
Data - SoftMax		0.0000		0.0000		0.7452
Data - Hyperbolic Tangent		1.0000		0.6063		0.4929

AlexNet confusionchart The confusion chart is present in Figure 4, where the rows represent the assumed classes, in this case 0 to O type negative and 1 to positive. The columns represent the target class assigned by the neural network.

The diagonal in green shows the correctly classified cases, and the diagonal in red presents the incorrectly classified. For example, 91 images were correctly classified in 0 class, representing 71.7% from the 138 testing images. On the other hand, 8 images, or 5.8% were not classified correctly for the same 0 class. The right column in gray represents the precision highlighted in green, and the false discovery rate highlighted in red, for each class. The gray bottom row represents the recall highlighted in green and the false negative rate highlighted in red for each class. The blue cell shows the overall accuracy highlighted in green.

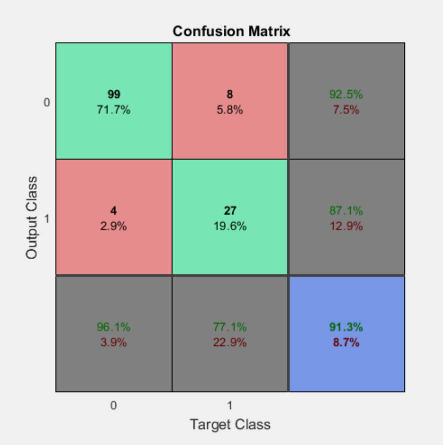


Fig. 4: Confusion Chart.

Network comparison Comparing the RBF performance according with the AlexNet layer input it's noticed that the best result achieved is with the fully connected layer 8, with hyperbolic tangent method, and the worst result is given by the fully connected layer with softmax method.

The reached results are higher than AlexNet, this can be explained by the combination of the two networks, AlexNet process the images and the Radial

Basis Function Network classify the processed data. The solo RBF performance is far from solo AlexNet, Table 2 'data' row, result not pertinent, compared with 1, the RBF development way doesn't allows the processing of a large database, but with a pre-processed database, like the fully connected 6 and 8 layers, the RBF is able to perform a reasonable classification.

5 Conclusions

This work investigated a method to identify and classify sunspots using image processing techniques. This method consists of two steps: image pre-processing and training phase using both convolutional neural network (CNN) and the Radial Basis Function Network (RBF).

The addressed images were taken from the Helioseismic and Magnetic Imager (HMI), of the Solar Dynamics Observatory (SDO). A total of 96 images were analyzed, of which 920 subfigures were generated in the pre-processing stage, separated in 138 for testing, 138 for validation and 644 for training.

We analyzed the network performance according to the precision, recall and accuracy components. The obtained accuracy, precision and recall showed competitive results for the both considered networks. This means the proposed approach is a competitive classifier for the sunspots groups, making possible the exploration and extension for other related images. In the future we expect to investigate more techniques to identify other sun relevant features, improving the research in this spatial area.

Bibliography

[1] R. G. Giovanelli, "The relations between eruptions and sunspots." *The Astrophysical Journal*, vol. 89, p. 555, 1939.

[2] G. Siscoe, "The space-weather enterprise: past, present, and future," *Journal of Atmospheric and Solar-Terrestrial Physics*, vol. 62, no. 14, pp. 1223–1232, 2000.

[3] R. Schwenn, A. Dal Lago, E. Huttunen, and W. D. Gonzalez, "The association of coronal mass ejections with their effects near the earth," in *Annales Geophysicae*, vol. 23, no. 3, 2005, pp. 1033–1059.

[4] J. T. Hoeksema, Y. Liu, K. Hayashi, X. Sun, J. Schou, S. Couvidat, A. Norton, M. Bobra, R. Centeno, K. Leka *et al.*, "The helioseismic and magnetic imager (hmi) vector magnetic field pipeline: overview and performance," *Solar Physics*, vol. 289, no. 9, pp. 3483–3530, 2014.

[5] G. Damião, "Estudo da atividade solar no passado em função da radiação cósmica," 2014.

[6] P. P. P. MALUF, "O numero de manchas solares, indice da atividade do sol medido nos ultimos 50 anos." *Revista Brasileira de Ensino de Fsica*, vol. 25, pp. 157–163, 2003.

[7] D. H. Hathaway, "The solar dynamo," *NASA Technical Report NASA-TM-111102, NAS 1.15:111102*, 1994.

[8] F. Clette, D. Berghmans, P. Vanlommel, R. A. Van der Linden, A. Koeckelenbergh, and L. Wauters, "From the wolf number to the international sunspot index: 25 years of sidc," *Advances in Space Research*, vol. 40, no. 7, pp. 919–928, 2007.

[9] K. M. HAN, J, "Data mining: concepts and techniques," in *Data mining: concepts and techniques.* Elsevier, 2000, pp. 83–120.

[10] A. Smola and S. Vishwanathan, "Introduction to machine learning," *Cambridge University, UK*, vol. 32, p. 34, 2008.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[12] L. Wang, K. Wang, and R. Li, "Unsupervised feature selection based on spectral regression from manifold learning for facial expression recognition," *IET Computer Vision*, vol. 9, no. 5, pp. 655–662, 2015.

[13] O. Chapelle, B. Scholkopf, and A. Zien, "Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]," *IEEE Transactions on Neural Networks*, vol. 20, no. 3, pp. 542–542, 2009.

[14] R. S. Sutton, A. G. Barto *et al.*, *Introduction to reinforcement learning.* MIT press Cambridge, 1998, vol. 135.

[15] R. A. Howard, "Dynamic programming and markov processes." 1960.

[16] F. Sarwar, A. Griffin, P. Periasamy, K. Portas, and J. Law, "Detecting and counting sheep with a convolutional neural network," in *2018 15th IEEE*

International Conference on Advanced Video and Signal Based Surveillance (AVSS), Nov 2018, pp. 1–6.