Object Recognition and Image Understanding Project Proposal

Shivali Dubey, Sascha Stelling

June 14, 2018

Question 1

- **Team:**
  Shivali:
  - Defining PReLU function
  - Training of datasets (backpropagation)
  - testing (result from Euclidean loss objective function)
  - Reducing overfitting

  Sascha:
  - Convolution filters
  - Training of dataset (feed forward)
  - Testing (results from multinomial logistic regression)

- **Problem Definition:** *Object detection and image classification* has various applications, for instance, in self-driving cars to detect and classify pedestrians, motorcycles, trees, bicycles etc; classification of features on the Earth such as roads, rivers, agricultural fields etc using satellite images. With the advancements in deep learning, every year, new algorithms/ models keep on outperforming the previous ones, to achieve the best possible accuracies for image classification. One of the most popular dataset used is the ImageNet dataset. In our project we propose to implement the deep learning algorithms based on a few selected studies[1][2][3][4] with the aim to attain the best possible classification accuracy.

[1] Girshick, R., Donahue, J., Darrell, T., Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. Tech Report (V5). UC Berkeley. October, 2014.

[2] He, K., Zhang, X., Ren, S., Sun, J. Delving Deep into Rectifiers: Surpassing Human-level Performance on ImageNet Classification. Microsoft Research. February, 2015.

[3] Krizhevsky, A., Sutskever, I., and Hinton, G. E. ImageNet classification with deep convolutional neural networks. In NIPS, pp. 1106–1114, 2012.

[4] Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556, 2014.

- **Dataset:** Tiny ImageNet[5]. Tiny Imagenet has 200 classes. Each class has 500 training images, 50 validation images, and 50 test images.
- **Approach:** With such a large dataset, one of the main challenges of classification is diversity of the images. Our model/algorithm must be able to handle fine-grained and specific classes even when they are hard to distinguish. In other words, we need to maximize inter-class variability, while minimize intra-class variability. At the same time, attaining the best possible classification accuracy is always a challenge for any given algorithm. The predictions go wrong when you have too many false positives and false negatives.

  Image Classification:
  Architecture:

  - Convolution: The main purpose of using multiple convolution layers is feature extraction. We use Scale-Invariant Feature Transform (SIFT)[6] descriptors which computes the Difference of Gaussians (DoG)[7]. DoG is used to detect blobs by subtracting two blurred images from another with different Gaussian kernels. The maxima and minima of this operation are taken by SIFT as key feature locations for the next neurons. To classify feature more accurately we make use of densely sampled SIFT, Extended Opponent SIFT and RGB-SIFT detector as described in [8] in three different convolution layers. We use Parametric ReLU (PReLU)[9] for the activation of a neuron and Adam[10] for optimization. PReLU is being used instead of ReLU because it improves model fitting with nearly zero computational cost and little overfitting risk[11]. Also, PReLU (and also ReLU) brings non-linearity into the system which allows learning complex functions. The weight initialization can be performed using Xavier's initialization[12]. Furthermore, the weight optimization is also controlled by PReLU.
  - Downsampling: Though the activation function passes only relevant pixels to the next layer, the array could still be big. To reduce the size of the array, we downsample it using an algorithm called max pooling where the output is separated into equally sized squares and only the maximum pixel in each square is taken for the next layer.

[5] https://tiny-imagenet.herokuapp.com/

[6] https://pdfs.semanticscholar.org/presentation/e903/196678c93315f2bf6f0235b3bab59c157b04.pdf

[7] http://micro.magnet.fsu.edu/primer/java/digitalimaging/processing/diffgaussians/index.html

[8] See 1

[9] https://www.cv-foundation.org/openaccess/content_iccv_2015/papers/He_Delving_Deep_into_ICCV_2015_paper.pdf?spm=5176.100239.blogcont55892.28.pm8zm1&file=He_Delving_Deep_into_ICCV_2015_paper.pdf

[10] https://arxiv.org/pdf/1412.6980.pdf

[11] See 2

[12] http://proceedings.mlr.press/v9/glorot10a/glorot10a.pdf

- Fully-connected Neural Network: We construct the last layer as fully connected neural network with hidden layer and logistic regression, and set all feature maps produced from previous layers as inputs. Softmax logistic regression can be used to represent categorical distribution i.e. probability distribution over different outcomes.
- Backpropagation: We use gradient descent method for optimization algorithm which is thus used for learning and training.

Reducing Overfitting: As studied in previous research works (Imagenet class), we perform data augmentation and dropout to reduce overfitting. Data augmentation refers to artificially enlarging image size using augmentation. Dropout refers to dropping out the output of each hidden neuron with probability 0.5 and less, such that the respective neuron can't participate in backpropagation[13].
Training: The image is first cropped representing an object part or a small object. The training is carried out by optimising the multinomial Euclidian loss (for bounding box representation, otherwise regression is more common) objective function using mini-batch gradient-descent (based on back-propagation)[14][15].
Testing: At test time, we have a trained dataset and an input image. The fully connected layers are converted to convolutional layers resulting into a Fully Convolutional (FC) network. During training, the images are cropped, however during testing the FC network is applied on the entire image. The result is a label prediction[16]. Note: We could also train our dataset by optimising the multinomial logistic regression objective function. In this case, the testing would result in a class score map with the number of channels equal to the number of classes[17].

- **Evaluation** & **Expected Results:** We expect our output images to be classified by the correct label. The results can be quantitatively evaluated by calculating the top 1 and top 5 test set error rates as done in previous studied[18]. The qualitative evaluation can be performed by selecting 5 most probable class for a given object in an image and computing their probabilities.

[13] See 4
[14] See 3 4
[15] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. Backpropagation applied to handwritten zip code recognition. Neural Computation, 1(4):541–551, 1989.
[16] See 3 4
[17] See 15 4
[18] See 1 2 3 4