Information Preserving Frame-based Image Interpretation

Vasudeva Kilaru
Lovely Professional University
(Jalandhar)
Venus.kilaru@gmail.com

Abstract— Interpreting image data to neural networks is challenging. Deep convolutional neural network methods have shown promising results in interpreting image data to neural networks, with convolution and pooling operations over the traditional fully connected dense layers. The performance of these deep convolutional methods, however, is often compromised by the constraint that the convolution and pooling operations interpret the image data to neural networks by compressing the data into lower dimensions that lead to information bottleneck. To mitigate this, neural networks tend to be larger with more parameters (filters) thus increasing the computational cost (GPU Resources). In this paper, I present an information-preserving way to interpret image data to a convolutional neural net without any information bottleneck and with relatively fewer parameters, that also ensures no loss in information due to convolution or pooling operations. Uniquely my method adds two additional operations, one over the first regular convolution operation and the other operation next to the deeper convolution operation in the neural net. The first operation is to divide the image into individual frames by frame-based crop operation and then apply regular convolution, pooling operations to interpret individual frames of the image into low dimensional tensors that preserve information from being bottlenecked since operations use frames of the image instead of using total image data. The second operation is a convolution operation applied after joining all low dimensional tenors of individual frames to extract information that later passes through the rest of the layers. This idea of using frames can better help model tasks like image generation and completion as well. Using frames of data instead of the total image helps in parallelizing computations thereby drastically decreasing the computational cost and depth of a neural network. Experiments conducted on inception networks worked better with relatively small network architecture on vision tasks.

Keywords—image interpretation, compute cost, network complexity, feature detection, network parameters, task parallelization.

I. INTRODUCTION

Interpreting unstructured data, especially image data to artificial neural networks is challenging. Existing methods like convolutional neural networks interpret image data more efficiently than earlier methods of fully connected layers that lack **parameter sharing** and **sparsity of data**. Convolutional neural networks have shown promising results in the field of computer vision from image recognition to self-driving cars however; the static, rigid operations like convolution, pooling leads to information bottleneck that compresses data and even force the network to be large with increased parameters to perform well on tasks. In convolutional neural networks, static set of weights/filters are applied in common on total image to detect different features of different regions on the image through convolution or pool operations which lead to the higher number of less insight full convolution/pool operations on the image thereby increasing the number of operations/computations on an image to detect all features. Convolutional neural networks often encounter information bottlenecks due to higher static convolution and pooling operations so, to perform better on vision applications these networks require more hardware resources for computations.

In this paper, I present a frame-based crop operation, a method applied before and after the extreme convolution operations in a Convolutional Neural Network (CNN) architecture. The frame-based crop operation parallelizes operations and addresses issues like information bottleneck, higher static convolution operations, and high computations (convolution operations).

This method of using additional frame-based crop operation is capable of effectively interpreting the data to neural nets without any information bottleneck so the network is relatively small. This method is also capable of reducing the number of overall filters needed to detect any feature of the image by dynamic filters which are specific to specific regions on the image without any extra or insightful filters on frames that don't detect any feature thereby decreasing the number of convolution operations (computations).

The two advantages of additional operation i.e., frame-based crop operation, are: (1) Better interpreting the data in a relatively smaller network can reduce computational cost, (2) Ability to parallelize the operations on individual frames reduces the computational cost.

II. RELATED WORK

A. Foundational works on image interpretation

In the early days of the deep learning era, the vision applications like image classification or semantic segmentation were carried out by neural networks with fully connected dense layers [1][2], however, these fully connected layers were not so good at interpreting unstructured data, especially image data which lead to poor performance on vision tasks. Around the mid-'90s in the paper "Convolutional Neural Network for Image, Speech, Time series" the idea of novel convolution operation was proposed which improved the performance of machine vision tasks like image classification [3], facial recognition and so on, by spatial interpretation of image data

to neural networks. The Convolutional operations were widely adapted in all the vision applications even now in self-driving cars, image-related tasks because of two main advantages of CNNs over the traditional fully connected layers, (1) **parameter sharing** that drastically reduced the number of parameters in the network and (2) **sparsity of data**. CNN's have shown very promising results by extending their intuition applicability from computer vision to NLP [4] (Natural Language Processing). Sometime later other methods were added to CNN's architecture to reduce the computations like Pooling [5], Batch Normalization. With these advancements, CNNs have shown better results on computer vision applications from traditional image classification, image semantic segmentation in (Figure 1) to more advanced tasks like facial recognition [6] [7] [Figure 2], check detection [8] [Figure 3], object detection [9] [Figure 4], and autonomous driving [10] [Figure 5].

B. Recent works on image interpretation

Around 2011, the research team at Microsoft has introduced the concept of residual blocks or skip connections in the RESNET paper that enabled to better interpret the data into deeper layers and train without problems like vanishing gradient/ exploding gradient [11]. The idea of 1x1 convolution operations i.e., network to network layers is adapted into inception network from RESNET paper.

The idea of Frame-based crop operation is inspired by the YOLO (you only look once) algorithm [12]. The idea of dividing the full-size image into smaller grids proposed in the paper is effective on tough tasks like real-time object detection in autonomous cars, however, the FBCO operation reduces the compute costs like 1x1 convolutions or bottleneck layers and also reduce network size drastically.

The idea of Frame-based crop operation is applied on the inception network of inception blocks which outputs the inception modules formed by stacking multiple convolution/ pooling operations (as shown in Figure 1) and large computational costs are managed by adding a bottleneck layer [13].

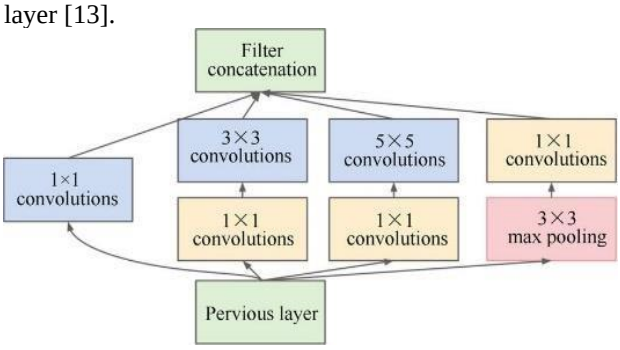


Figure 1. Inception block [13].

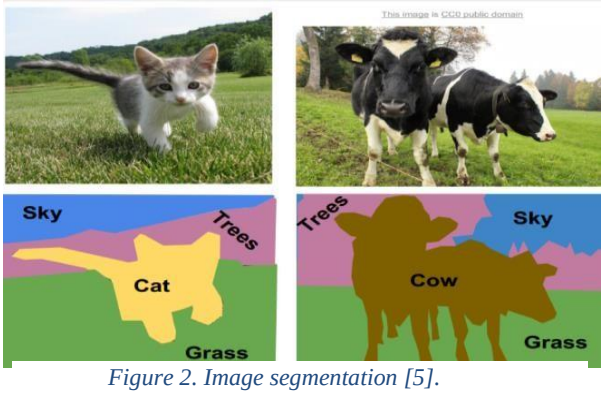


Figure 2. Image segmentation [5].

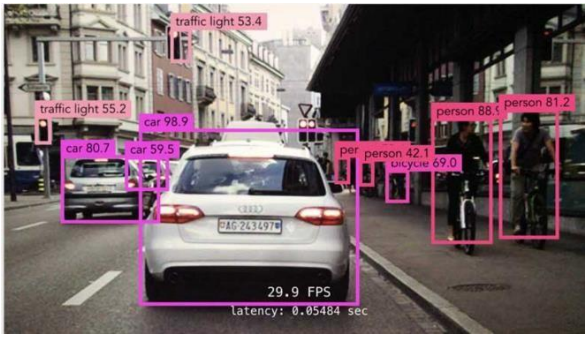


Figure 3. Object detection [8].

Figure 4. Facial detection and recognition [6]

Figure 5. Autonomous Car [9]

Despite having many proven applications, the neural network often needs a large network with high parameters (specifically filter weights that detect different features on image) to train well and perform well. The novel method proposed in the paper, which is inspired by the methods of tokenization of input in NLP, YOLO and Inception network can significantly reduce the size of the network along with the number of parameters in the neural network that enables us to train fast with low computational cost without compromising on the performance of on the vision tasks.

III. INFORMATION PRESERVING FRAME-BASED OPERATION

We first review how the conventional neural networks are applied to interpret image data for image classification problems (say).

Earlier images were interpreted in the form of a single-dimensional array that stores the pixel values (as shown in Figure 6) of the image [14], but this way of interpreting the image to neural networks had issues like large parameters (weights) and data sparsity. These issues lead to the need for large network models to be trained well on tasks and huge compute costs to perform computations of huge parameters associated with pixel values at each layer.

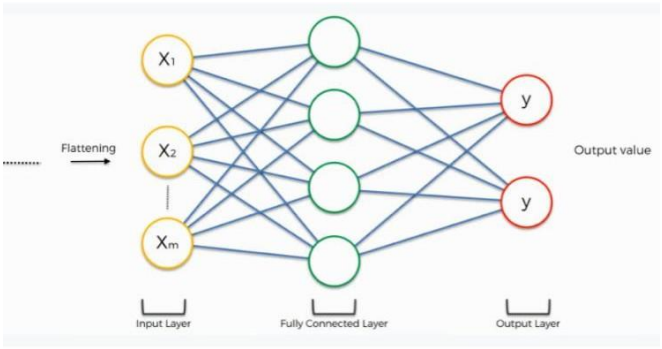


Figure 6. Fully connected layers for Image

In the LeNet paper-1989 by Yann Lecun, Convolution operations were proposed to do better image interpretation to neural networks than regular fully connected (dense) layers that need a relatively very small number of weights (parameters). Convolution operations (as shown in [Figure 7] [15]), which were inspired from the feature detection mechanism of human vision, worked out well due to the ability of parameters sharing and data sparsity that reduced the number of trainable parameters with concept convolution filters, used to detect features on the image.

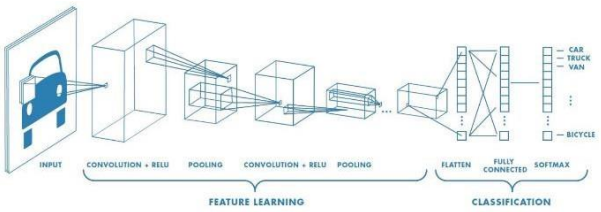


Figure 7. Convolutional layers for Image

Frame-Based Crop operation (FBCO)

The vision tasks are often tough for neural networks to generalize and be robust with fine variations because of their complex mapping function from input images to classes (say). Many advancements made neural networks work better in complex mapping vision tasks with larger networks. The computational costs are managed by bottleneck layers and backpropagation gradient issues are managed by skip connections or residual blocks. All the methods could significantly improve the neural networks to map complex functions but drastically increase the computational cost, the idea of FBCO is to decrease the computational cost without compromising on the performance on different vision tasks.

To avoid the information bottleneck of image data by conventional convolution, pooling operations and enable better information interpretation during training and testing with relatively low computational cost, I employ the frame-based crop operation. Below, I first describe the frame-based crop operation and then illustrate how it incorporates in the conventional deep convolutional neural network to improve image interpretation.

Frame-based Crop operation. As discussed in related work, existing methods of convolutional neural networks compress data to lower dimensions by bottlenecking the information (As a part of interpreting the image data to neural network). The main reason behind this is due to hardcoded convolution, pooling operations over the total image at stretch thereby increasing the number of parameters/filters, the computational cost to detect complete features. The frame-based crop operation, inspired from the method of tokenizing data, YOLO and Inception network with 1x1 convolution operations and bottleneck layers [13], proposed in this paper divides the single image data into individual dependent frames on which the conventional CNN methods are applied parallelly. As illustrated below and in Figure 8, the frame-based crop operation (FBCO) takes the image input and divides it accordingly, into individual image frames. These frames are later on passed through the regular convolutional neural network separately.

The actual execution of FBCO operation.

Original Image

11	21	32	22	25	76	87	99
22	62	19	49	61	65	43	76
10	08	20	47	55	77	21	43
54	06	57	12	32	57	13	25
21	66	40	15	54	42	44	54
19	54	21	36	87	41	43	33
23	23	71	31	37	98	53	21
73	99	66	23	34	76	73	35

Frame 1

00	**00**	**00**	**00**	**00**	**00**	**00**	**00**
00	**00**	**00**	**00**	**00**	**00**	**00**	**00**
00	**00**	20	47	55	77	**00**	**00**
00	**00**	57	12	32	57	**00**	**00**
00	**00**	40	15	54	42	**00**	**00**
00	**00**	21	36	87	41	**00**	**00**
00	**00**	**00**	**00**	**00**	**00**	**00**	**00**
00	**00**	**00**	**00**	**00**	**00**	**00**	**00**

Frame 2

00	**00**	**00**	**00**	**00**	**00**	**00**	**00**
00	62	19	49	61	65	43	**00**
00	08	20	47	55	77	21	**00**
00	06	57	**00**	**00**	57	13	**00**
00	66	40	**00**	**00**	42	44	**00**
00	54	21	36	87	41	43	**00**
00	23	71	31	37	98	53	**00**
00	**00**	**00**	**00**	**00**	**00**	**00**	**00**

Frame 3

11	21	32	22	25	76	87	99
22	62	19	49	61	65	43	76
10	08	**00**	**00**	**00**	**00**	21	43
54	06	**00**	**00**	**00**	**00**	13	25
21	66	**00**	**00**	**00**	**00**	44	54
19	54	**00**	**00**	**00**	**00**	43	33
23	23	71	31	37	98	53	21
73	99	66	23	34	76	73	35

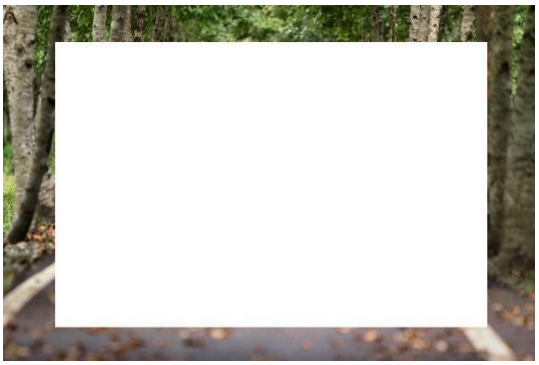
High-level intuition of FBCO operation.

Figure 8. Original Image (frame {0, 1, 2, 3})

Frame 1

Frame 2

Frame 3



A. FBCO-Parallelizing the operations.

The frame-based crop operation which divides the original image into individual frames as shown in Figure 8, enables us to parallelize the computations. Earlier the total image is passed into the CNN, but here the image is divided into frames that can pass through the CNN simultaneously resulting in a corresponding low dimensional representation. These representations are combined back to represent the total image in low dimensions. At least one convolution operation is performed before flattening the image data as it improves the quality of representing the image (as shown in Figure 9) in low dimensions. Thus, the final output can be used to perform many tasks related to computer vision applications.

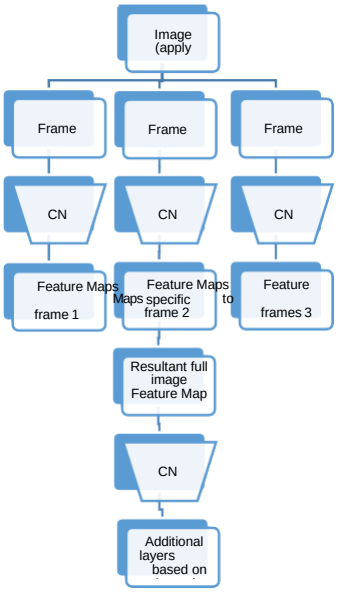


Figure 9. FBCO introduced CNN architecture

B. FBCO-Dynamic parameters on frames

The additional method of frame-based crop operation can significantly reduce the number of overall parameters thereby reducing the computational cost. The advantage of tokenizing the data i.e., dividing the total image into frames enables the model to learn specific filter weights to detect features on different regions or frames of the image that contribute to decreasing the number of filters to detect total features. For example, To detect all features of an image, let's suppose 256 unique filters are needed but practically 256 filters don't detect features in every region because, there may specific features like eyes or eyebrows that exists only in a particular region of the image which means the eye feature detecting filters are not useful to apply on the other parts of the image so, if we divide the image into frames, the model learns specific features on every region i.e., the model learns eye detecting filters only to that particular region/frame that has eyes thereby reducing the number of convolution operations of various filters throughout the image that decreases the computational cost.

C. FBCO-Feature Detection

Feature detection with FBCO remains as appropriate as it is in conventional CNN, however, FBCO can make feature detection much more effective with the concept of frames. In the conventional methods, feature detection takes place on the full image but with the FBCO, the full image is divided into individual frames over which the feature detection is done. As discussed in point C, it helps to reduce the number of parameters

by allotting only necessary and dynamic filters for detecting accurate features on each frame. For example:

01	01	01	01	01	01
01	01	01	01	01	01
01	01	01	01	01	01
00	**00**	**00**	**00**	**00**	**00**
00	**00**	**00**	**00**	**00**	**00**
00	**00**	**00**	**00**	**00**	**00**

The above tensor is corresponding to a horizontal edge image and applies conventional convolution operation to detect edge over the image of the horizontal edge using the horizontal edge detecting filter, the output would be a low dimensional feature map same as the resultant feature map shown below.

Horizontal edge filter

01	01	01
00	00	00
-1	-1	-1

Resultant Feature map

00	00	00	00
03	01	-1	03
03	01	-1	03
00	00	00	00

and applying the same horizontal edge detecting filter on the image by the FBCO method also results in the same feature map along with an option to parallelize the computations faster on combining.

FBCO operation on horizontal edge image into frames

Frame 1

00	**00**	**00**	**00**	**00**	**00**
00	**00**	**00**	**00**	**00**	**00**
00	**00**	01	01	**00**	**00**
00	**00**	**00**	**00**	**00**	**00**
00	**00**	**00**	**00**	**00**	**00**
00	**00**	**00**	**00**	**00**	**00**

Frame 2

00	**00**	**00**	**00**	**00**	**00**
00	01	01	01	01	**00**
00	01	**00**	**00**	01	**00**
00	**00**	**00**	**00**	**00**	**00**
00	**00**	**00**	**00**	**00**	**00**
00	**00**	**00**	**00**	**00**	**00**

Frame 3

01	01	01	01	01	01
01	**00**	**00**	**00**	**00**	01
01	**00**	**00**	**00**	**00**	01
00	**00**	**00**	**00**	**00**	**00**
00	**00**	**00**	**00**	**00**	**00**
00	**00**	**00**	**00**	**00**	**00**

Resultant Feature Map

00	00	00	00
03	01	-1	03
03	01	-1	03
00	00	00	00

Therefore, by dividing the image into frames, operations like convolution and pooling operations can better interpret image data while without compromising on the efficient feature detection task that helps in performing better in a long run.

IV. IMPLEMENTATION DETAILS

Step by step approaches to implement FBCO operations over the conventional convolutional neural networks are as follows:

1. The numerical tensor of the image is to be divided into frames via an operation called frame-based crop operation. The operation/function takes in the image tensor along with any two hyperparameters among the following:

   1. Stride-skip: The numerical value that specifies the method of cropping i.e., it tells, how many pixels need to be skipped while dividing the image into frames.
   2.Frames: The numerical value that specifies the total number of frames need to be made out of a full-size image.
   3. Shared pixels: The numerical value that specifies the number of pixels to be shared between every consecutive frame ensures the spatial connections between the frames.

The relation between all the three parameters with the full size of an image i.e., the number of pixels of the image is given by

(1) (Image is expected to be of shape nxn where n ∈ positive even integers).

$$f = [n/((s\_s - s\_p)+1)] \quad (1)$$

the f-number of frames, n-shape of the image, s_s-number of strides to skip ∈ positive even integers, s_p number of shared parameters ∈ positive even integers.

Note-a. The FBCO operation, relations, equations are only applicable on images of shape nxn where n ∈ positive even integers.

b. The FBCO operation can be applied on images of any shape by applying a composition preserving multi-level spatial pooling layers to bring down images to shape nxn without losing any aesthetic value of image before FBCO [16].

2. Each frame is passed through a state of art convolutional neural network separately like VGG16 [17], Resnet [11], etc. The output of the neural network of each frame (low dimensional representation of each frame on applying operations like convolution, pooling) is then combined in reverse order to form an overall low dimensional tensor representing the full image. The combined tensor is passed at least through one convolution operation before flattening to ensure the data sparsity issue as shown in Figure 9.

3. The tensors are flattened into single-dimensional arrays to pass through fully connected layers to perform any task

related to vision applications from image classification, semantic segmentation to object detection, path detection.

A. Experiments

The addition of the FBCO method drastically reduces the computational cost and accomplishes any task with comparatively a smaller network, without compromising the performance. The method is applied to an existing state of art vision model inception. The experiments conducted have shown better performance on image classification tasks with a smaller network of inception models. The regular inception model has got an accuracy of 91.813% on dog breed classification at the lowest SoftMax. The FBCO applied inception managed to get the accuracy of 91.809% at third side branch SoftMax from the lowest and even more accuracy ranging 92%-95% at later SoftMax side branches in the inception network and with relatively reduced computations. The FBCO applied inception could do the task in a relatively lesser time (0.7x) than the regular inception on image classification with the reduced number of effective weights and by parallelizing the execution.

V. Conclusion

This paper presents a frame-based crop operation over the extreme convolution operations in any state of art CNN model to preserve essential information/features in an image. Interpreting the image data in low dimensional tensors by compressing it may lead to information bottleneck issues but with FBCO, the model ensures the best quality of image data interpretation which significantly reduces the network size, parameters and operations required to perform on any vision task without compromising on any factor. First, the frame-based crop operation is implemented on full image to break it down to smaller frames which enables to parallelize the computations and decrease the parameters thereby decreasing the computational cost. Finally, the resultant low dimensional tensors of each frame are combined and passed to layers to carry out machine vision tasks like object detection etc. The positive aspect of the method is that it significantly reduces the computational cost without any tradeoff with the performance of the model.

Acknowledgment

This research was supported by Naustone Inc, Lovely Professional University. I am grateful to the CEO (chief executive officer) of Naustone Mr Sreekar Nayani for all support both morally and intellectually to pursue research and bring out great insights on the subfield of vision i.e., unstructured data interpretation to deep neural networks.

Thanks to all the open-source research papers, lectures especially for IEEE CVPR papers and Coursera.org on various concepts that supported me in pursuing this research on data (image) interpretation.

References

[1] J. Schmidhuber. Deep learning in neural networks: An overview. Neural Networks, 2015. Published online 2014; based on TR arXiv:1404.7828 [cs.NE]. 2

[2] Ken-yuh hsu, Hsin-yu li, Demtri. Implementation of a Fully Connected Neural Network. IEEE Xplore. 2

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

[4] Aditya Mogadala, Marimuthu Kalimuthu, Dietrich Klakow. Trends in integration of Vision and Language Research. ArXiv.org. 2, 3

[5] Scherer, Dominik & Müller, Andreas & Behnke, Sven. (2010). Evaluation of Pooling Operations in Convolutional Architectures for Object Recognition. 92-101. 10.1007/978-3-642-15825-4_10. 2

[6] Image is accessed on 02/12/2021 from https://tariq-hasan.github.io/concepts/computer-vision-semantic-segmentation/. Image of semantic segmentation. 2, 3

[7] Image is accessed on 02/12/2021 from https://sciforce.solutions/. Image of facial recognition and detection. 2, 3

[8] The Concept has been accessed from the below Wikipedia page https://en.wikipedia.org/wiki/Signature_recognition on 02/12/2021. 2

[9] Image is accessed on 02/12/2021 from https://www.analyticsvidhya.com/ . image of object detection in self-driving cars. 2, 3

[10] Image is accessed on 02/12/2021 from https://www.analyticsvidhya.com/ . image of object detection in self-driving cars. 2, 3

[11] Kaiming He, Xiangyu Zhang, Jain Sun. Deep residual learning for image recognition. ArXiv.org. 2

[12] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi. You Only Look Once. CVPR. 6

[13] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich. Going Deeper with convolutions (Inception network). 1409.4842, September 2014. 2, 3, 6

[14] Image is accessed on 02/12/2021 from https://www.superdatascience.com/blogs/convolutional-neural-networks-CNN-step-4-full-connection . Fully connected layer representing image data.

[15] Image is accessed on 02/12/2021 from www.towardsdatascience.com. Convolutional Neural Network.

[16] Long Mai, Hailin Jin, Feng Liu. Composition Preserving Deep photo aesthetic assessment. CVPR 2016. 5

[17] Karen simonyan, Andrew Z. Very Deep Convolutional Network for Large Scale Image Recognition (VGG16). arXiv:1409.1556v6 **[cs.CV]** 5