The Good, the Bad, and the Ugly: Neural Networks Straight from JPEG

S. F. dos Santos[1], N. Sebe[2], and J. Almeida[1]

[1]Instituto de Ciência e Tecnologia,
Universidade Federal de São Paulo – UNIFESP
{felipe.samuel, jurandy.almeida}@unifesp.br

[2]Dept. of Information Engineering and Computer Science
University of Trento – UniTn
niculae.sebe@unitn.it

ICIP'20 – Abu Dhabi, United Arab Emirates – October 25-28 – 2020

Session: ARS-14 – Machine Learning for Image and Video Classification IV

Outline

Introduction

Convolutional neural networks (CNNs) have achieved state-of-the-art performance in many computer vision tasks.

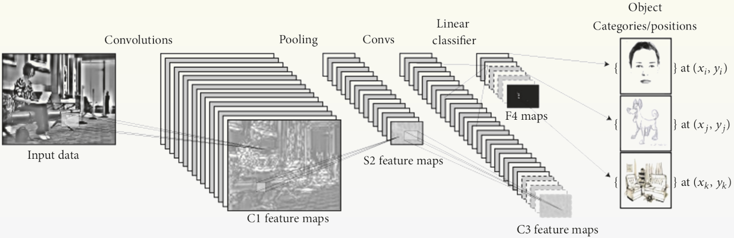


Figure: Example of a CNN used in a computer vision task[1].

[1] Voulodimos et al. 2018.

Introduction

CNNs usually process **RGB pixels**, but **image data** are often stored in a **compressed format**, like JPEG, PNG and GIF.

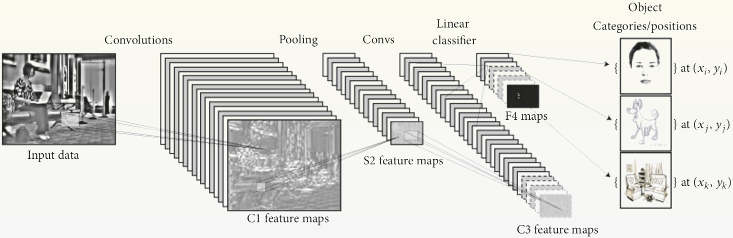


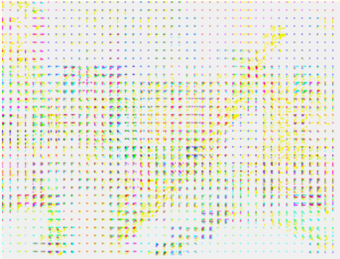
Figure: Example of a CNN used in a computer vision task[2].

[2] Voulodimos et al. 2018.

Introduction

A **costly decoding process** is required for obtaining **RGB images**.

RGB Image



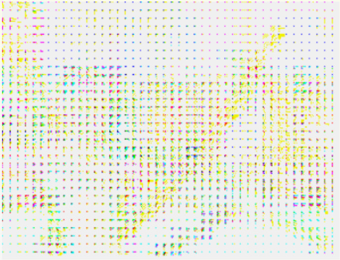
DCT Coefficients

Figure: A RGB image and its DCT coefficients from compressed data.

Introduction

A **costly decoding process** is required for obtaining **RGB images**.

RGB Image



DCT Coefficients

Figure: A RGB image and its DCT coefficients from compressed data.

What if CNNs are designed to process JPEG compressed data?

JPEG Compression

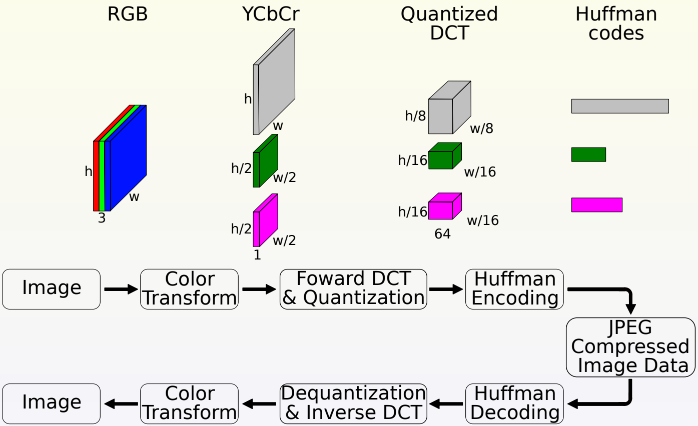


Figure: JPEG compression and decompression process[3].

[3]Gueguen et al. 2018.

JPEG Compression

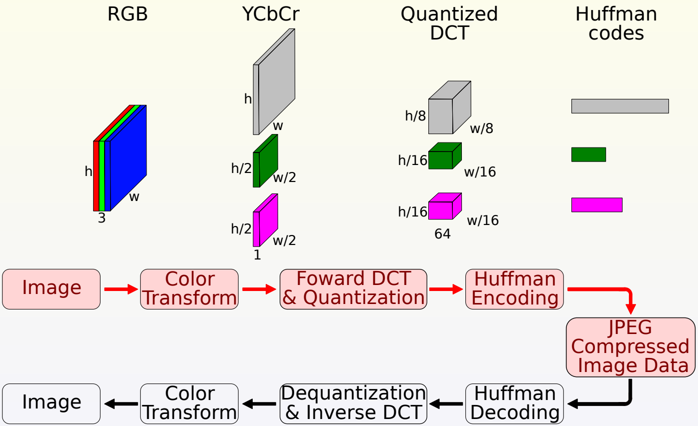


Figure: JPEG **compression** and decompression process[3].

[3]Gueguen et al. 2018.

JPEG Compression

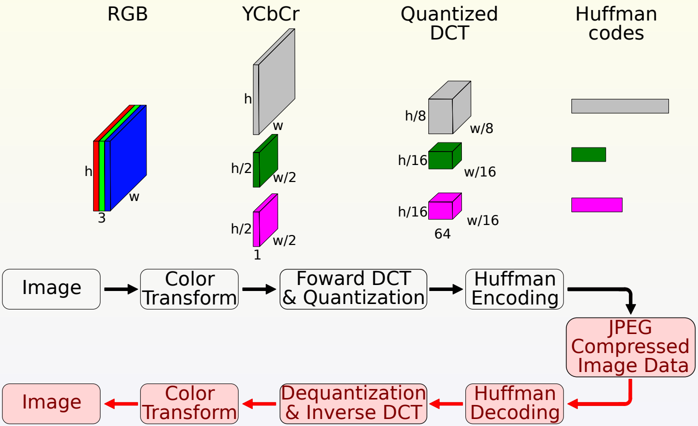


Figure: JPEG compression and **decompression** process[3].

[3]Gueguen et al. 2018.

Related Work

The potential of the **JPEG compressed domain** has been widely explored by **many conventional image processing techniques** ...

Related Work

The potential of the **JPEG compressed domain** has been widely explored by **many conventional image processing techniques** but exploited by **only a handful** of **deep learning methods**:

- **Gueguen et al. 2018:** architectural modifications to the ResNet-50 network to accommodate DCT coefficients from JPEG images;
- **Deguerre, Chatelain, and Gasso 2019:** adaptations to the Single Shot MultiBox Detector (SSD)[4] to accommodate block-wise DCT coefficients as input;
- **Ehrlich and Davis 2019:** reformulation of the ResNet architecture to perform its operations directly on the JPEG compressed domain.

[4] Liu et al. 2016

Our starting point is the work of Gueguen et al.[5]

Modifications on ResNet-50 to accommodate DCT inputs:

1. the **first stage** is **skipped**;
2. the amount of **input** channels of the **second and third stages** are **changed** to ensure that their number of output channels are equal to the original ResNet-50;
3. the **strides** of **early blocks** from the **second stage** are **decreased** in order to mimic the increase in size of the receptive fields in the original ResNet-50.

[5] L. Gueguen et al. "Faster Neural Networks Straight from JPEG". In: *NIPS*. 2018, pp. 3937–3948.

Before and After

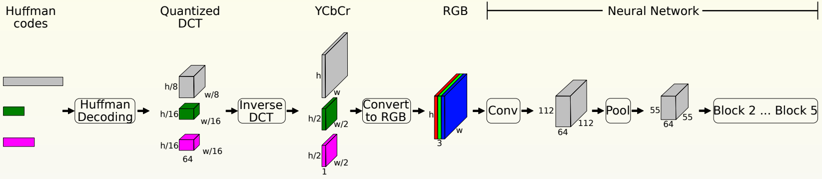


Figure: Original ResNet-50 network.

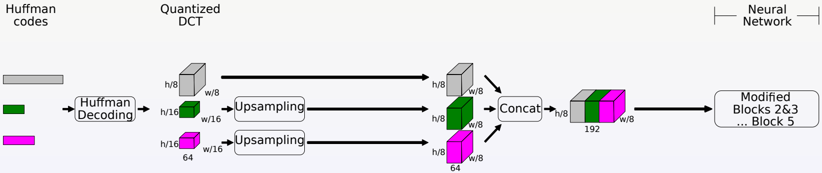


Figure: ResNet-50 using DCT as input.

Less steps but more miles! And now?

Drawbacks ...

The **changes** introduced by Gueguen et al.[6] in ResNet-50 **raised** its **computation complexity** and **number of parameters**.

... and Opportunities

To **alleviate** the **network complexity**, we use a **Frequency Band Selection (FBS)** to select the most relevant DCT coefficients.

[6] L. Gueguen et al. "Faster Neural Networks Straight from JPEG". In: *NIPS*. 2018, pp. 3937–3948.

After and Now

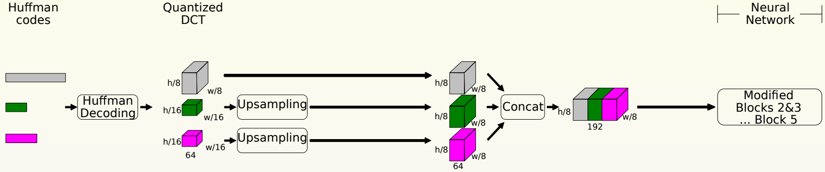


Figure: ResNet-50 using DCT as input.

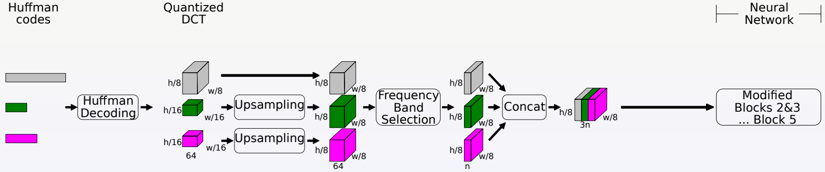


Figure: ResNet-50 using DCT and FBS.

Our Approach

Frequency Band Selection (FBS)

- **High frequency** data have **little visual effect** on the image.
- Only the n **lowest frequency** coefficients are **retained**.
- The **second stage** is **changed** to have $3n$ **input channels**.

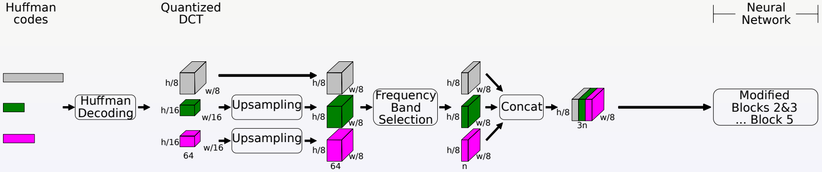


Figure: ResNet-50 using DCT and FBS.

Dataset

Subset of the ImageNet[a] dataset:

- 268,773 images from 211 classes;
- 215,018 (80%) of training images;
- 53,755 (20%) of test images;
- Different difficulty levels:
  - fine-grained: 211 of the 1000 classes from ImageNet;
  - coarse-grained: 211 classes grouped into 12 categories;
- Smallest side resized to 256 pixels;
- Crop size of 224x224 pixels.

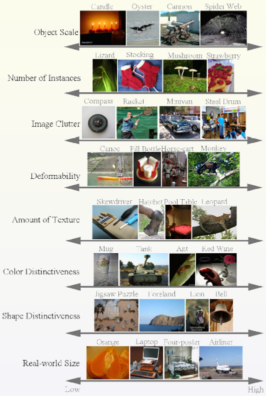


Figure: The diversity of data in ImageNet dataset.

[a] Russakovsky et al. 2015.

Implementation Details

Table: The hyperparameters used for training all the networks.

Parameter	**Options**
Batch size	128
Initial learning rate	0.05
Total number of epochs	120
Step-decay scheduler setting	LR divided by 10 every 30 epochs
Data augmentation operations	random crops and horizontal flips

Network Complexity

Table: Computational complexity (GFLOPS) and number of parameters for the original ResNet-50 with RGB inputs and networks using DCT.

Approach	Input Channels	GFLOPs	Params
ResNet-50 + RGB[7]	3x1	3.86	25.6M
ResNet-50 + DCT[8]	3x64	5.40	28.4M
ResNet-50 + DCT + FBS	3x32	3.68	26.2M
ResNet-50 + DCT + FBS	3x16	3.18	25.6M

[7] He et al. 2016.
[8] Gueguen et al. 2018.

Impact of the Difficulty Level of Classification Tasks

Table: Accuracy (%) of the original ResNet-50 network and its modified versions for image classification tasks with different difficulty levels.

Approach	**Classification Task**	
	Fine (211 Classes)	*Coarse (12 Classes)*
ResNet-50 + RGB (3x1)[9]	76.28	96.49
ResNet-50 + DCT (3x64)[10]	70.28	94.15
ResNet-50 + DCT + FBS (3x32)	69.79	94.53
ResNet-50 + DCT + FBS (3x16)	68.12	93.92

[9] He et al. 2016.
[10] Gueguen et al. 2018.

Impact of the Image Resolution

Table: Accuracy (%) for the original ResNet-50 with RGB inputs and networks using DCT as input for images with different resolutions.

Approach	Image Resolution			
	32	*64*	*128*	*256*
ResNet-50 + RGB (3x1)[11]	81.82	90.39	94.56	96.49
ResNet-50 + DCT (3x64)[12]	72.72	82.06	90.32	94.15
ResNet-50 + DCT + FBS (3x32)	71.83	82.22	90.78	94.53
ResNet-50 + DCT + FBS (3x16)	70.35	81.35	90.16	93.92

[11] He et al. 2016.

[12] Gueguen et al. 2018.

Impact of the JPEG Quality Level

Table: Accuracy (%) for the original ResNet-50 with RGB inputs and networks using DCT as input for images with different JPEG qualities.

Approach	JPEG Quality			
	25	*50*	*75*	*100*
ResNet-50 + RGB (3x1)[13]	95.78	95.98	96.09	96.49
ResNet-50 + DCT (3x64)[14]	93.84	94.02	94.50	94.15
ResNet-50 + DCT + FBS (3x32)	93.63	93.97	94.20	94.53
ResNet-50 + DCT + FBS (3x16)	92.69	93.26	93.66	93.92

[13] He et al. 2016.

[14] Gueguen et al. 2018.

Conclusions

Remarks

- Evaluation of the potential of CNNs designed for JPEG data.
- Several aspects of the work of Gueguen et al.[15] were studied.
- Frequency Band Selection (FBS) to alleviate complexity.
- Experiments were conducted on a subset of the ImageNet.
  - Classification tasks with different difficulty levels.
  - Different spatial resolutions and JPEG quality settings.
- Networks were robust to changes in the JPEG quality but susceptible to variations in the spatial resolution.
- FBS proved to be effective in reducing network complexity.

[15] Gueguen et al. 2018.

Conclusions

Future Work

- Evaluation of other CNNs designed for JPEG images.
- Evaluation of our network on the whole ImageNet dataset.
- Evaluation of smarter strategies for selecting DCT coefficients.
- Extension of our ideas to networks devised for MPEG videos.

References I

- B. Deguerre, C. Chatelain, and G. Gasso. “Fast object detection in compressed JPEG Images”. In: *IEEE Intelligent Transportation Systems Conference (ITSC’19)*. 2019, pp. 333–338.
- M. Ehrlich and L. S. Davis. “Deep Residual Learning in the JPEG Transform Domain”. In: *IEEE International Conference on Computer Vision (ICCV’19)*. 2019, pp. 3484–3493.
- L. Gueguen et al. “Faster Neural Networks Straight from JPEG”. In: *NIPS*. 2018, pp. 3937–3948.
- K. He et al. “Deep Residual Learning for Image Recognition”. In: *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR’16)*. 2016, pp. 770–778.

References II

- W. Liu et al. "SSD: Single Shot MultiBox Detector". In: *European Conference on Computer Vision (ECCV'16)*. 2016, pp. 21–37.
- O. Russakovsky et al. "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision* 115.3 (2015), pp. 211–252.
- Athanasios Voulodimos et al. "Deep learning for computer vision: A brief review". In: *Computational intelligence and neuroscience* 2018 (2018).

Acknowledgments

The authors are grateful to:

- CAPES
- CNPq (grants 423228/2016-1 and 313122/2017-2)
- FAPESP (grants 2017/25908-6 and 2018/21837-0)
- Caritro Deep Learning Lab of the ProM facility at Rovereto

Obrigado!!!

Thank you for your attention!!!

If you have any questions, do not hesitate to contact us:

- Samuel Felipe dos Santos (felipe.samuel@unifesp.br)
- Nicu Sebe (niculae.sebe@unitn.it)
- Jurandy Almeida (jurandy.almeida@unifesp.br)