Viterbi Internship - Final Work Report

Arka Sadhu
Supervised by: Prof. Ram Nevatia

July 11, 2017

Contents

1 Abstract

Media forensics in general involves detection of the tampered media, identification of the tampered portion as well as trying to recover the original media.

2 Introduction

The work is done as a part of the MediFor Project. The MediFor project aims at pushing the state of the art research in the field of media forensics which in broad sense deals with the tampering of the media (image, video or audio) and its detection. This work only deals with image forensics. For each manipulated image the MediFor project demands the actual image on which manipulation is done (this is called the baseline image), the kind of manipulation, and in case of splice manipulation where one image is spliced onto another image it also demands the donor image. This work focuses only on the first part, where the aim is to find the baseline image. It is assumed that the world set contains the true baseline image. All experiments are done on Nimble Dataset which is publicly available for use.

3 Theory

3.1 Basic Definitions

- Probe Image : This is the given image. It may or may not be manipulated.
- Probe folder : Folder containing the probe images.
- Base Image : This the actual image corresponding to a probe image with no manipulations exists.
- Donor Image : In the case where the manipulation is such that a part of image A is pasted onto image B, then image A is called the Donor Image and B is the base image. The resulting image would be the manipulated image which would exist in the probe folder.
- World folder : Folder contaning all the images. This includes base, donor as well as the probe images.
- World set : The collection of images in the world folder. It is used interchangeably with world images.
- Provenance : Provenance in simple sense means the origin, so it defines the original image of a particular probe image.
- Provenance Graph : A relational graph which depicts all the transformations a particular baseline image would've undergone to reach the probe image. It is assumed that all the intermediate images are also a part of the world dataset.
- Base detection : Detection of the base image from a given probe image and the entire world set.
- Donor detection : Detection of the donor image from a given probe image and the entire world set.

3.2 MediFor Project

The MediFor project broadly has two main categories Video and Image. For any kind of media, MediFor Project wants automated assessment of the integrity of the media. If successful, the MediFor platform will automatically detect manipulations, provide detailed information about how these manipulations were performed, and reason about the overall integrity of visual media to facilitate decisions regarding the use of any questionable image or video.[1]

Table 1: Places365 Validation

Correct Matches	Total Images	Accuracy
2975	3650	81.5068493151
2969	3650	81.3424657534
2952	3650	80.8767123288
2993	3650	82
2977	3650	81.5616438356
3036	3650	83.1780821918
2941	3650	80.5753424658
2976	3650	81.5342465753
2941	3650	80.5753424658
2938	3650	80.4931506849

There are three technical areas of interest for integrity analytics. [2]
—-May need to add a few more lines here———

- Digital Integrity : This is related to the noise modelling and statistics and its consistency.
- Physical Integrity : This is related to shadow consistency.
- Semantic Integrity : This is related to semantic consistency

In this work we are concerned only with semantic integrity.

3.3 Contributions of this Work

3.4 Base Detection

Base detection problem is essentially finding the underlying base image given a probe image. Here we make the assumption that the base image exists in the world set. The next problem is to get all the manipulated images derived from the base image. And beyond this is to create a provenance graph of the collected manipulated images. The last problem is not addressed in this work.

3.4.1 Neural Networks used

We use two pre-trained caffe [3] models in this work. AlexNet[4] trained on Places365[5] and AlexNet trained on ImageNet. The reason for using AlexNet instead of VGG16 or any other models is that we wanted to work with a simplest model and test our performances without compromising memory and time. Places365 is a scene-centric dataset while ImageNet is object centric dataset. And as such we expect there should be a difference in their base detection capability.

In this work we use the AlexNet trained on Places-365 everywhere unless explicitly mentioned that the AlexNet trained on ImageNet is used.

3.4.2 Speeding up the Test Time

3.4.3 Which layer and metric to choose?

To find the baseline image, we employ the following method. We use the Nx1 dimensional vector produced by the network. As we go deeper into the layers, we expect more semantic features to be captured. The features are represented in the form of a vector and is known as a feature vector. In the AlexNet architecture we specifically compare three layers fc7, fc8, and prob layer which is the output after the operation of softmax function.

So we intend to find a way such that given the feature vectors from the probe image, we want to be find the base image. We use a simple approach for this. We find the feature vectors of the base image as well, and then compare the feature vectors using different metrics. For a metric to be good we would ideally want for a probe base pair it should give a high value and for unrelated images it should return a very low value. Also we would prefer a substantial difference between related and unrelated images. For this work we tried the following metrics :

- SSD : Sum of Squared Distances
- SAD : Sum of Absolute Distances
- NCC : Pearson's correlation coefficient

It turned out that NCC gave the most desirable results.

Figure 1: Probe Base Pair taken from Nimble Dataset 2017 Dev 1 Beta 4

For example in the image pair 1 the metrics for the prob layers were as follows :

Table 2: Prob Layer Metrics

	Prob layer
SSD	0.072518922
SAD	0.29026049
NCC	0.98303729249860383

Clearly SAD is not desirable since it gives a medium score to a matching pair. Both SSD and NCC give good results in this case, but emperically it was found that NCC is not only easier for comparison (need not invert the high and low score), but is also more robust, that is gives high score even in cases where the images have been manipulated to larger degree and SSD isn't able to caputre the similarity. As a result, NCC has been the prime candidate for the rest of the work.

Another important part was to choose a layer. It was found that for many cases that using the last layer (prob) gave a very low score for a matching pair.

Figure 2: Probe Base Pair taken from Nimble Dataset 2017 Dev 1 Beta 4

For example the image pair 2 returns the following NCC scores for the three layers.

Table 3: NCC scores for different layers

	NCC
fc7	0.75791334934860821
fc8	0.87673938938958917
prob	0.35012885670990512

Clearly fc8 gives the most desirable result, but it is interesting to theorize the reason why prob gives such a low score. We hypothesize that the SoftMax layer in some sense disturbs the features because it gives the probability of closeness to a particular scene. So if a scene is not present in the Places365 database, this would give a weird output. Also going by the emperical knowledge that the deeper layers tend to extract out more semantic features, fc8 should give the best result and this intuition follows our finding. fc8 gives consistently higher score than fc7 and fc6 for probe base pair and lower score for unrelated images.

Here is a small table comparing the different methods

Table 4: Layers and Correlation threshold

Tot Images = 320	Prob		fc8		fc7	
Threshold	correct	%correct	correct	%correct	correct	%correct
0.95	203	0.634375	271	0.846875	175	0.546875
0.9	243	0.759375	288	0.9	243	0.759375
0.8	264	0.825	312	0.975	287	0.896875
0.5	295	0.921875	316	0.9875	313	0.978125
0.4	302	0.94375	320	1	316	0.9875

3.4.4 Image Slicing

A general observation in the datasets was that the manipulation existed in only a part of the probe image. For this reason, we use the method of image slicing, that is cutting the image into two halves horizontally or vertically, even getting four quadrants as well. This gives a very easy boost to the accuracy but at the same time demands more computational resources or time.

3.5 Donor Detection

4 Troublesome Cases

Here are some cases, for which we are clueless as to why the correlation is low.

References

[1] DARPA, "Medifor project description." `http://www.darpa.mil/program/media-forensics`.

[2] "Medifor project description 2." `https://researchfunding.duke.edu/media-forensics-medifor`.

[3] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25* (F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, eds.), pp. 1097–1105, Curran Associates, Inc., 2012.

[5] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.