Negative results in computer vision: A perspective

Ali Borji
Center for Research in Computer Vision, University of Central Florida
aborji@crcv.ucf.edu

Abstract

Before delving into details and discuss the value of negative and inconclusive results, first we need to look at computer vision research in a broader perspective with respect to scientific practices and methodologies conducted in other fields such as social or biological sciences. Here, I discuss what is a negative result, how can we tell whether a result is positive, negative or inconclusive, and how negative results should be incentivized and disseminated. Hopefully, this writing will spark further conversations in the community regarding dealing with negative results [1].

1. Introduction

Computer vision research involves of a mixture of theoretical and experimental research. A small fraction of publications introduce principled theories for vision tasks (e.g., optical flow [6]). A large number of publications report models and algorithms (e.g., for solving the object detection problem) that are more powerful than contending models. Thus, compared to other fields, computer vision is less hypothesis-driven and more practical. The emphasis has traditionally been placed on improving existing models in terms of performance over benchmark datasets. While some papers conduct statistical tests, it is not the common practice. As in some other fields, there is a high tendency among computer vision researchers to submit positive results as such results are often considered to be more novel by the reviewers. A large of experiments also end up with negative or conclusive results which are not shared with the community. A discussion seems to be necessary to properly address this issue.

2. What is a negative result?

The term “publication bias” was first coined by Theodore Sterling in 1959 [10]. It points to the situation where “publication of research results depends not just on the quality of the research but also on the hypothesis tested, and the significance and direction of effects detected” [7]. It is sometimes referred as the “file drawer effect,” [9]. As the name suggests, results not supporting the original hypotheses (i.e., negative results or Null hypotheses) often end up buried in researchers’ file drawers. It is also called the “Positive-results bias” where positive (or successful) results are more likely to be submitted, or accepted than negative or inconclusive results. Therefore, what is published is not the true representative of all results. Perhaps the main reason behind the tendency towards publishing positive results is the intense competition among scientists. The unwritten “publish or perish” rule drives academics to publish interesting high quality papers in large volumes to get more citations and to secure funds to do their research.

Notice that negative result is different than no result. No result is a situation where nothing is complete or a work has been done incompletely or incorrectly thus leading to inconclusive or unreliable findings. Let me bring an example in the context of computer vision. Consider training a CNN to do a certain task. Often a lot of trickery is involved to properly train a CNN. Now, should one write a paper on the basis that choosing a certain parameter (e.g., training the network for 10 epochs instead of 100) does not lead to a convergence? At the end of the day, what matters is how much a study adds to what is already known, regardless of the sign of the outcome.

Negative results, with a slightly more liberal definition, highlight limitations, failures, or flaws of computer vision models, datasets, or scores. For instance, adversarial images demonstrate situations where CNNs can be easily fooled [8] while humans have no trouble recognizing them. In image captioning literature, it has been shown that a nearest neighbor classifier that simply chooses the caption of the most similar image to the test image, outperforms state of the art methods (in 2015 [4]). In saliency modeling [2], naive baselines such as the average fixation map or a central Gaussian blob outperform several fixation prediction models [11]. In object detection, some works have identified the cases where histogram of oriented gradients [3] features fail on certain detection problems [12]. Smart baselines (e.g., a

[1] See https://arxiv.org/pdf/1705.04402.pdf, for a longer version of this paper.

classifier that picks the label of the most frequent class) define the lower bounds while human performance gives an upper-bound on performance and helps identify the weak links in models. Thus, these are complementary to negative results and help assess the progress and move the field forward.

A related problem here is the issue of replicability. Replicability of findings is believed to be at the heart of empirical sciences [1]. Computer vision researchers tend not to replicate other people's works for two major reasons. Either it is possible to replicate someone else's work or it is not. In the former case, a reviewer may find your results boring or predictable. In the latter case, you may be accused of not following the right procedure. So, in both case there is a risk factor involved. To mitigate the risks, the community needs to advocate for well-documented solid negative results and educate the reviewers.

3. Statistical hypothesis testing

Negative results either go completely unpublished or are somehow turned into positive results through adjustments (e.g., selective reporting, methods alteration, different data analyses, increasing the number of observations). A generic term coined to describe these post-hoc choices is HARKing ("Hypothesizing After the Results are Known").

Data dredging, data fishing, data snooping, p-hacking, and HARKing are tricks and ways to tweak data, consciously or unconsciously, such that statistically significant results can be obtained. When talking about this, people often quote Ronald Coase's famous saying "If you torture the data long enough, it will confess". One major flaw is analyzing the data without first devising a specific hypothesis as to the underlying causality. There is a clear distinction between exploratory versus confirmatory analyses. While searching for patterns in data is legitimate, applying a statistical hypothesis tests on the same data is wrong. A simple way to avoid this problem is to form a hypothesis before carrying out significance tests. Notice that the p-value is valid only if you stick to exactly what you had planned to do in advance. Another way is to conduct randomized out-of-sample tests. Here, a data set is randomly partitioned into two subsets. One subset is used for formulating a hypothesis and the other is used for testing the hypothesis. Fortunately, this is routinely done in computer vision research. Another flaw in statistical testing is multiple comparisons. If you try large numbers of hypotheses, the chance that one of them may be positive increases. One solution to overcome this is to simply divide the significance criterion ("alpha") by the number of all significance tests conducted during the study. This is known as the Bonferroni correction [5]. Notice that this is a very conservative test. An alpha of 0.05, divided in this way by 100 to account for 100 comparisons, yields a very stringent per-hypothesis alpha of 0.0005.

One major challenge when designing experiments is dealing with confounding factors (a.k.a confounders or confounding variables). Not controlling the confounding factors can lead to misleading and useless results. Let me clarify this by an example. Assume you aim to investigate the effect of exercise (independent variable) on weight loss (dependent variable). Let's say you collect data from 2n subjects (n male and n female) and conclude that indeed exercise leads to weight loss. Is this a reliable finding? Maybe not, due to several concerns: a) some subjects might have been using drugs so the weight loss could be attributed to that, b) female subjects might have eaten less than male subjects, so gender might be a confounding factor, c) some subjects might have spent less time exercising than others, and so on. In this regard, it is extremely important to understand the difference between correlation and causation. For example, shoe size correlates with reading level in children but it not the true reason of better reading ability (the true reason might be age or education).

Let me bring in an example related to computer vision. Let's say you have designed a system that tells whether a scene is captured in China or in the United States. Let's assume you test your model on a dataset that accidentally has people visible in images taken in China while none of the images taken in US contain people. Can we say for sure this model is able to do the task? Not definitively. The reason is that the model might have discovered that the existence of a person determines the location where it was taken. The model may fail when presented with images with no people in them. In this example, randomly sampling the data and scaling up the size of the dataset might mitigate the problem and reduces the bias.

4. Dissemination of negative results

Computer vision has a unique model of publication. While there are several prestigious journals (e.g., IEEE PAMI, IJCV) to publish the results, top-tier conferences are where the real action happens (e.g., CVPR, ICCV, ECCV). A large number of papers are submitted to these conferences and get reviewed in a short period of time (around 3 months with the net reviewing period varying from 1 to 2 months). These conferences are very competitive (acceptance rate of around 25% to 30%) thus leaving place only for novel, interesting and often positive results. Although, once in a while interesting negative results appear in these conferences, researchers usually do not risk conducting such studies. Some conferences (ICLR and NIPS; publishing some vision papers) have recently adopted an open review system where the communications between the reviewers and the authors are made available to the public. While this does not directly address publishing negative results, it is an effective way of disclosing the hidden chunk of knowledge to the scientific community. Unfortunately, vision conferences have not yet adopted this platform. The reason might be protecting ideas and ongoing efforts.

An important concern in publishing negative results is giving a fair chance to the original authors (especially in cases where published results are questioned) to respond to the counter arguments. Journals seem to be a better venue for such conversations and open debates. Some fields have already devised effective strategies for dealing with this concern. For example, the Journal of Behavioral and Brain Sciences (BBS) invites other scientists to comment on an accepted paper. The paper and the corresponding comments then get published in the same issue of the journal. Journal of Vision and Journal of Vision Research publish commentary and re-analysis papers (sometimes discussing the negative results) as the "letters to the editor". In all of these journals, all material has to go through the peer review process. These practices enrich the scholarly work.

ArViv and blogs are rising venues for publication. Both, however, suffer from a lack of peer review. One advantage of arXiv is rapid distribution of findings. One drawback is that sometimes papers are early half-baked progress reports often published to claim an idea. Blogs allow personal opinions and discussions in an informal setting (i.e., conversations). Although very interesting, such a venue includes sporadic, noisy thoughts. Nevertheless, occasionally people exploit these venues for communication or settling a matter.

One of the most effective habits in computer vision is sharing code and data which has contributed tremendously to the progresses of the field and has been rightfully incentivized by high number of references to such works (similar to benchmark papers). Not only has this habit proven to be extremely useful to deal with replicability issues and speeding up contributions, it also serves as a good model for incentivizing negative results.

5. Conclusion

Conclusive, important, well-documented, and peer-reviewed negative results that come from rigorous investigations should certainly be welcome. Such results can save a lot of efforts by preventing redundant efforts, add to the intellectual richness of the community, promote scholarly culture, and give tremendous insights regarding limits of models, datasets, and scores.

Negative results should be properly and effectively disseminated, incentivized, shared, encouraged and discussed. A culture needs to be built to recognize and embrace such efforts. Negative results as well as smart baselines can be as important as algorithm development or dataset collection and should be given a fair chance to be presented in conferences and journals. To this end, we may need to change the mindset on a larger scale (e.g., funding agencies). Also, negative results should be disseminated in such a way that the original authors can get a chance to respond (in case of replication failure).

Statistical testing has been undermined in computer vision and should be taken into account in the future. Several factors need to be carefully taken into account in conducting statistical testing including selection of the appropriate tests, controlling for confounding factors, compensating for multiple comparisons, etc. Statistical testing should be also exploited in model comparisons.

Overall, negativity towards negative results is counterproductive and such results should be published given that they follow appropriate and sound scientific methodologies.

References

[1] J. B. Asendorpf, M. Conner, F. De Fruyt, J. De Houwer, J. J. Denissen, K. Fiedler, S. Fiedler, D. C. Funder, R. Kliegl, B. A. Nosek, et al. Recommendations for increasing replicability in psychology. *European Journal of Personality*, 27(2):108–119, 2013.

[2] A. Borji and L. Itti. State-of-the-art in visual attention modeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):185–207, 2013.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[4] J. Devlin, S. Gupta, R. Girshick, M. Mitchell, and C. L. Zitnick. Exploring nearest neighbor approaches for image captioning. *arXiv preprint arXiv:1505.04467*, 2015.

[5] C. W. Dunnett. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, 50(272):1096–1121, 1955.

[6] B. K. Horn and B. G. Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981.

[7] H. Lian, Y. Ruan, R. Liang, X. Liu, and Z. Fan. Short-term effect of ambient temperature and the risk of stroke: a systematic review and meta-analysis. *International journal of environmental research and public health*, 12(8):9068–9088, 2015.

[8] A. Nguyen, J. Yosinski, and J. Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 427–436, 2015.

[9] R. Rosenthal. The file drawer problem and tolerance for null results. *Psychological bulletin*, 86(3):638, 1979.

[10] T. D. Sterling. Publication decisions and their possible effects on inferences drawn from tests of significanceor vice versa. *Journal of the American statistical association*, 54(285):30–34, 1959.

[11] B. W. Tatler. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of vision*, 7(14):4–4, 2007.

[12] C. Vondrick, A. Khosla, T. Malisiewicz, and A. Torralba. Hoggles: Visualizing object detection features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1–8, 2013.