Negative Results in Computer Vision: A Perspective

Ali Borji

PII: S0262-8856(17)30167-1
DOI: doi: 10.1016/j.imavis.2017.10.001
Reference: IMAVIS 3653

To appear in: *Image and Vision Computing*

Received date: 4 October 2017
Accepted date: 16 October 2017

Please cite this article as: Ali Borji, Negative Results in Computer Vision: A Perspective, *Image and Vision Computing* (2017), doi: 10.1016/j.imavis.2017.10.001

Noname manuscript No.
(will be inserted by the editor)

Negative Results in Computer Vision: A Perspective

Ali Borji

the date of receipt and acceptance should be inserted later

Abstract A negative result is when the outcome of an experiment or a model is not what is expected or when a hypothesis does not hold. Despite being often overlooked in the scientific community, negative results are results and they carry value. While this topic has been extensively discussed in other fields such as social sciences and biosciences, less attention has been paid to it in the computer vision community. The unique characteristics of computer vision, particularly its experimental aspect, call for a special treatment of this matter. In this manuscript, I will address what makes negative results important, how they should be disseminated and incentivized, and what lessons can be learned from cognitive vision research in this regard. Further, I will discuss matters such as experimental design, statistical hypothesis testing, explanatory versus predictive modeling, performance evaluation, model comparison, reproducibility of findings, the confluence of computer vision and human vision, as well as computer vision research culture.

1 Introduction

What is a negative result? One may characterize a negative result as "when a hypothesis does not hold" or "when the outcome of an experiment or a model is not what is expected". Such a definition, however, could be one out of many possible definitions. One may argue that an unexpected result is actually a good useful positive result to share. Another possible definition is that a negative result is when the performance is not better given metrics such as accuracy. Regardless of how negative results are defined, such challenging and sometimes inconclusive findings are often discouraged and buried in the drawers and computers. Therefore, the publication record reflects only a tiny slice of the conducted research. In some sense they fabricate the "dark matter" of science. Such findings, however, still hold value. At the very least they can save resources by preventing researchers from repeating the same experiments. Perhaps the main reason for an overwhelmingly high number of negative results not put forward for dissemination is the lack of incentives. Interestingly, some researchers have even argued that most published findings are false [1]. Some also claim that hiding negative results is unethical. Nevertheless, negative results have been and continue to be constructive in the advancement of the science (e.g., Michelson-Morley experiment [2]).

To answer whether negative results are important in computer vision, should be published, or even if it makes sense to talk about them, first we need to investigate how computer vision research is conducted relative to scientific practices and methodologies conducted in other fields such as social or biological sciences. Computer vision research consists of a mixture of theoretical and experimental research. A small fraction of publications introduce principled theories for vision tasks (e.g., optical flow [3]). 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 relatively less hypothesis-driven and more practical. Some negative results offer invaluable insights regarding strength and shortcomings of existing models and theories, while others provide smart baselines. The emphasis has traditionally been placed on improving

A. Borji (corresponding author)
Center for Research in Computer Vision, University of Central Florida, Orlando, FL., USA
E-mail: aborji@crcv.ucf.edu

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.

Computer vision has its own unique characteristics making it distinct from other fields, thereby demanding a specific treatment of negative results. Firstly, vision is an extremely hard problem which has baffled many smart people throughout the history. The complexity of the problem makes it difficult to run controlled experiments and come up with a universal theory of vision. Secondly, often a lot of variables are involved in building vision algorithms and in analyzing large amounts of data. Further, fair comparison of several competing models using multiple evaluation scores exacerbates the problem. To address these, it would be very helpful to borrow from other fields (e.g., natural sciences) where experimental design and statistical testing are integral parts of the scientific research.

The common practice in experimental hypothesis-driven fields (e.g., cognitive science) include carefully formulating a hypothesis, identifying and controlling confounding factors, designing the right stimulus set, collecting high quality data, and performing appropriate statistical tests. These are complicated to perform in computer vision research as often many factors are involved. In particular, statistical analysis becomes very challenging in presence of many parameters and models. This makes it complicated to decide which statistical test is needed or when statistical analysis is critical to conduct. Principled and systematic gauging of the progress (rather than relying on trials and error and luck) helps judge what truly works and what does not and, hence steer the research in the right direction. For instance, we might have not given up on neural networks easily if we did more careful rigorous analyses in the past.

Notice that dealing with negative results is a very controversial topic and still unsettled in many fields. So, do not expect this writing to touch on all of the aspects. Rather, here, I try to shed light on some less explored matters and put computer vision in a broader perspective with respect to science in general, and its related fields such as Neuroscience and Cognitive Science, in particular. Indeed, further discussion is needed in the vision community to converge to a consensus regarding treatment of negative results.

In what follows, first I elaborate on science versus engineering and where computer vision fits. I will continue with a comparison of computer and human vision research and how they relate to each other in terms of goals, research methodologies and practices. This is followed by discussions of negative results and statistical analysis in the context of computer vision. Section 6 considers the dissemination of negative results. Finally, a wrap up is presented in the epilogue.

2 Computer Vision: Engineering or Science?

Let's start with the question of whether computer vision is a scientific or an engineering discipline, or both. Science is concerned with understanding fundamental laws of nature, whereas engineering involves the application of science to create technology, products and services useful for society. Science asks questions about nature while engineers design solutions to problems.

As a scientific discipline, computer vision is concerned with gaining high-level understanding from digital images, video sequences, views from multiple cameras, or multi-dimensional data. It seeks to automate tasks that the human visual system can do and involves the development of a theoretical and algorithmic basis to achieve automatic visual understanding. Further, it deals with constructing a physical model of the scene (i.e., how the scene is created), how light interacts with the scene, as well as low-, intermediate-, and high-level descriptions of the scene content [4]. In other words, the ultimate goal of computer vision is image understanding, the ability not only to recover image structure but also to know what it represents. As a technological and engineering discipline, computer vision seeks to apply its theories and models for the construction of computer vision systems and applications.

Science and engineering are complementary and are beautifully and happily married in computer vision. We have a very solid in-depth scientific understanding of phenomena such as image formation, depth perception, stereoscopic vision, color perception and optical flow. Some engineering applications, among many, include biometrics (robust face and fingerprint recognition), optical character recognition, gesture recognition, motion capture, game playing, structure from motion, image stitching, machine inspection, retail, 3D model building, medical imaging, automotive safety, autonomous cars, assistive systems, and surveillance (in traffic and security). In this respect, computer vision is both theoretical (e.g., optical flow formulation) and experimental (e.g., model replication, parameters tuning, hacks, and tricks).

3 Computer Vision and Biological Vision

Vision is a broad interdisciplinary area. Both computer and human vision systems share the same objective which is converting light into useful signals from which accurate models of the physical world can be constructed. This information helps an agent (e.g., be it a robot or a human) live, act, and survive in its environment.

For a long time, human vision research has been concentrated on understanding the principles and mechanisms by which biological visual systems (with higher emphasis on primate vision) operate. This is in essence a reverse engineering (or inverse graphics) task. Likewise, computer vision research seeks a theory and engineering implementation. Despite sharing the same goal, they own unique characteristics. Early human visual sensory mechanisms, including the retina and the Lateral Geniculate Nucleus (LGN), are much more elaborate than current digital cameras (CCD sensors). Neural networks in higher visual areas (e.g., visual ventral stream) accommodate a sophisticated hierarchical processing through cascades of filtering (modeled as convolution), pooling, lateral inhibition, and normalization mechanisms. The result is a selective and invariant representation of the objects and scenes. This is somewhat akin to what Convolutional Neural Networks (CNNs) [5] do. Almost half of the human brain (considered to the the most complex known physical systems and thus a major scientific challenge) is devoted directly or indirectly to vision. The entire brain needs about 20 watts to operate (enough to run a dim light bulb). A processor as smart as the brain requires at least 10 to 20 megawatts of electricity to operate [6]. As to processing speed, the brain is still faster than the fastest supercomputers [7]. A remarkable capability of human vision is attention (a.k.a active vision) which allows selecting the most relevant and informative part of the massive incoming visual stimulus (at a rate of 10^8-10^9 bits/sec) [8]. Both human and computer vision systems have their own biases. Human vision is extremely sensitive to faces and optical illusions. Similarly, computer vision systems get easily fooled by adversarial examples [9]. One thing that we know, almost for sure, is that vision should be solved by frameworks that start with extracting simple features and build increasingly more complex ones. This is mainly because the world we live in is compositional.

There has indeed been a cross-pollination in the two fields (e.g., [10–21]). On the one hand, experimental paradigms and psychophysics tools in cognitive vision have been exploited to study the behavior of computer vision algorithms or to interpret how they work. For example, Parikh and Zitnick [22] employed the image jumbling paradigm, introduced in [23], to inspect whether some computer vision algorithms capture local or global scene information. Deng et al. [24] used the bubbling paradigm, proposed by Gosselin and Schyns [25], to model fine grained object recognition. The rapid (or ultra rapid) serial visual presentation [26, 27], has been utilized to investigate the quality of images generated by Generative Adversarial Networks [28]. Vondrick et al. [29] and Fong et al., [30] leveraged human recognition biases to improve machine classifiers. On the other hand, computational tools have been exploited heavily to understand how human vision works. For example, deep convolutional networks have recently been used to study the representational space in the visual ventral stream (e.g., [31]). Moreover, a plethora of computer vision, image processing, and machine learning tools have been utilized in biological vision research for the purposes such as stimulus design, discovering cues humans might rely on in solving a task, and modeling single neurons and neural populations.

In terms of performance, while computer vision has made large strides, it is still nowhere near human vision. In general, it seems that computer vision is better than human vision in some restricted tasks where variability is relatively low (e.g., optical character recognition, fingerprint recognition, frontal face recognition, etc). However, it lags far behind in cases where variability is high (e.g., view invariant object recognition [32]). Current state of the art computer vision techniques revolve around deep learning models [33], in particular CNNs [5]. These models have outperformed traditional techniques on a wide variety of vision problems. It is even claimed that CNNs outperform humans on classic hard problems such as scene recognition [34]. Nevertheless, while nature has evolved a very efficient robust biological solution to the problem of vision, computer vision is still looking for a general computational theory, let alone efficient physical implementations (e.g., on Silicon).

While both research communities have accumulated a great wealth of knowledge, they are struggling with some common long-standing challenging problems. Perhaps the biggest of all is the invariance problem which is believed to be the holy grail of vision. Humans are remarkably good at recognizing objects under drastic variations (e.g., illumination, rotation, blur, and occlusion) but the mechanisms behind this capacity in biological vision are still unknown. There has been a great deal of research in computer vision to come up with invariant representations. Although the current state of art algorithms (i.e., CNNs) provide partial invariance to some transformations (e.g., translation, rotation, and scale), a principled theory is yet to be developed (See for instance [35]). This is where cross talks in both fields can be extremely useful. Another challenge, related to the first one, is the role of feedback and top-down modulation in visual processing. The resurgence of deep neural networks has raised the hope that maybe a universal solution to deal with all vision problems is within our reach. This is even more conceivable when we notice that a) CNNs are rooted in the seminal findings of Hubel and Wiesel [36], and b) biological vision systems might be following similar mechanisms as in CNNs (e.g., [21,31,37,38]).

As for the research practice, different methodologies have been adopted in the two fields, driven by different (sometimes short term) goals and constraints (e.g., building a technology versus proving a hypothesis). Research in computer vision is traditionally benchmark-driven where algorithms that perform better than others are favored. The emphasis is on improving accuracy. Meanwhile, human vision research is primarily hypothesis-driven. Hypotheses, null and alternative, are carefully formulated, confounding factors are controlled, an appropriate stimulus set is created, behavioral or physiological data is carefully collected, and appropriate statistical tests are conducted. The outcome is valid until proven otherwise. It is the collection of evidences, for or against a hypothesis, that drives the science. Here, I would like to bring an example from my own work in eye movement research. A study by Greene and Wolfe [39] reported a failure in predicting an observer's task from his fixations (essentially a negative result) effectively negating the anecdotal finding of Yarbus [40]. In a reinvestigation [32], we repeated the experiment by considering a larger set of parameters over a bigger dataset. To their contrary, we concluded that Yarbus' hypothesis still holds. Several follow-up studies also supported our finding thus reinforcing the original hypothesis.

There is a discrepancy of opinions on the relationship between computer and biological vision. Some researchers argue that the human visual system provides the most compelling reference model. We can learn much from it as an existence proof and as a great source of inspiration. Therefore, there should be a symbiosis between the two communities as they address the same problem. Some others, to the contrary, argue that since the two systems function under very different constraints, the artificial vision solution does not necessarily need to mimic the biological solution (in reference to the flying). Nonetheless, it seems, to me, that the time is ripe for both communities to learn from each other to address challenging problems in vision. A great wealth of biological data (behavioral and neurophysiological) and computational models (e.g., different CNN architectures) are available that should be linked for understanding the visual system and building better algorithms.

4 Negative Results in Computer Vision

"Those who cannot remember the past are condemned to repeat it." - George Santayana

Let me first define some buzzwords before turning the discussion to computer vision.

Publication bias: coined by Theodore Sterling in 1959 [41], 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" [42]. It is sometimes referred to as the "file drawer effect," or "file drawer problem" [43]. 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.

Confirmation bias (a.k.a confirmatory bias or myside bias): is a type of cognitive bias in which one tends to favor, accept, interpret, or recall findings that align with his preexisting beliefs or hypotheses [44].

Negative results either go completely unpublished or are somehow turned into positive results through adjustments (e.g., selective reporting, post-hoc reinterpretation, 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"). I will elaborate on this further in the next section.

There are arguments for and against publishing negative results. Let's look at some supporting arguments first. Such results should be part of the scientific record for the sake of completeness. Without them, literature surveys and meta-analyses would be biased. They can save a lot of efforts by preventing researchers to conduct redundant investigations. Further, they motivate critical evaluation, analytical thinking and discussions, thus contributing to the intellectual sophistication of the community. Some counter arguments include the followings. Negative results should be less favored due to the scarcity of space. This seems to be no longer an issue in the age of digital publication. Some people argue that negative results can lead to a phenomenon known as the "cluttered office phenomenon" where in an office full of academic papers, it is hard to tell the good ones from the poor ones [45]. This resonates with computer vision research where the field is already replete with an overwhelmingly high volume of publications per year, making separating the wheat from the chaff daunting.

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. Examples include a) not having enough participants to do a meaningful analysis, b) not having a control condition for an intervention, c) faulty measurements, or d) inconclusive or wrong statistical analysis. Nevertheless, negative results carry value, although modest. In theory, information from an experiment is not absolutely zero, if done correctly. Even when an experiment gives the exact same result as before (i.e., replication), a higher confidence towards that finding is gained. In some sense, this resembles probability density estimation but over different explanations for a phenomenon. Not every negative result is interesting though. As an example, 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 [9] 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 [46]). In saliency modeling [8], naive baselines such as the average fixation map or a central Gaussian blob outperform several fixation prediction models [47]. In object detection, some works have identified the cases where histogram of oriented gradients [48] fails on certain detection problems [49]. Smart baselines (e.g., a 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. Some of this type of papers tend to have titles that start with "In defense of $\cdots$". An example is the famous paper by Richard Hartley entitled "In defense of the eight-point algorithm" [50] where he shows that applying a very simple normalization before computing the fundamental matrix from a set of eight or more points results in a robust method that is comparable with the best iterative algorithms. SLIC superpixels [51] is another example where authors show how standard k-means with spatially uniform seeds can work very well. Thus, baseline methods are complementary to negative results and help gauge the progress and move the field forward. In addition to these, visualization techniques have also proven to be very effective in understanding, interpreting, and evaluating computer vision models (e.g., [52, 53]).

A related problem here is the issue of replicability or reproducibility. Replicability of findings is believed to be at the heart of empirical sciences [54]. As in other fields, 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 to reproduce the findings. Thus, in both cases there is a risk factor involved. To mitigate the risks, the community needs to advocate for well-documented solid results to ease the reproducibility process.

Confirming reproducibility requires a substantial amount of work even the source code is available due to lack of algorithm details, implementation matters, unavailable data, gratuitously long running time, etc. Proper incentives are therefore needed to encourage reproducibility. One incentive is that ones research needs to be reproducible to have high impact. Some questions that need to be thought of in this regard are as follows. Should we have explicit seals of reproducibility? Should it be part of the publication process? Should reviewers check the code? Should acceptance be conditional on the release of data and code?

Overall, due to the nature of the computer vision research, in particular its engineering aspect, the problem of the negative results and replicability is less pressing compared to biological research. Reporting wrong results can be detrimental in natural sciences (e.g., clinical and medical research) because it has important implications (lives are on the line). According to a vision psychologist, Elissa Aminoff, "if it (a computer vision algorithm) works, and is working better than everyone else's, it will eventually be made available for everyone to use. If there was a replication issue, it wouldn't get very far. So, I think in computer vision the discussion is more useful for theoretical advancement, rather than for replication issues."

5 Statistical hypothesis testing

"Extraordinary claims demand extraordinary evidence." [55] - Carl Sagan

How can we tell for sure a result is negative? It makes sense to announce a result negative only when careful rigorous statistical tests have been conducted. In what follows, I highlight some concerns regarding statistical analysis and testing in the context of computer vision.

Scientific research is usually done in two frontiers: experimental and theoretical [56]. In the theoretical frontier, the pursuit is for a comprehensive theory or principles (often expressed in mathematical forms) to explain visual phenomena (e.g., image formation). In the experimental frontier, researchers utilize two types of approaches: exploratory analysis (pattern mining) and hypothesis testing. In the first approach, they focus on running experiments to gather data in searching for hypotheses. In the second approach, first a well-defined hypothesis is explicitly formulated and then controlled experiments are performed to test its validity. Computer vision researchers use both frontiers although majority of the effort is on building tools and engineering solutions.

Statistical testing is an integral part of scientific research in many fields (especially experimental ones). Unfortunately, awakening to standard analysis of experimental results has been slow in computer vision and most of analyses are often carried out ad-hoc and heuristically. The input data to computer vision algorithms and estimates produced by them are often noisy. Thus, there is an inherent uncertainty associated with results produced by these algorithms. These uncertainties are expressed in terms of statistical distributions, and distributions' means and covariances. Reporting uncertainty is often neglected in computer vision or is done in a wrong way [57]. Further, many researchers misunderstand confidence intervals and standard error bars [58, 59].

Hypothesis tests are used in determining whether the outcome of a study would lead to a rejection of the null hypothesis for a pre-specified level of significance. The null hypothesis represents what is believed by default, before seeing any evidence. Statistical significance, p-value, is a probability value indicating whether the outcome of an experiment can happen accidentally or not. The smaller the p-value, the larger the significance. If the p-value is less than the required significance level, then we say the null hypothesis is rejected at the given level (e.g., 5% or 1%) of significance (i.e., leading to a conclusion). If the p-value is not less than the required significance level, then the test has no result (not conclusive or negative results). In this case the evidence is insufficient to support a conclusion.

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[1]. 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 (train, validation, and test sets). An important complication 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 [60]. 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. Multiple comparison challenge has been studied in considerable detail and much progress have been achieved on this topic since the 90's (See for example [1, 61–63]).

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 with an example. Assume you aim to investigate the hypothesis that exercise (independent variable) causes weight loss (dependent variable). Let's say you collect data from 2n subjects (n in the control group; who did not exercised) and conclude that indeed

[1] "To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of." – Ronald Fisher

exercise leads to weight loss. Is this a reliable finding? Maybe not, due to several concerns:

- Some subjects (under treatment) might have been using drugs so the weight loss could be attributed to that,
- Control group might have included mostly female or old subjects (i.e., unbalanced groups), so gender or age might be confounding factors,
- Subjects in the treatment group might have eaten less than subjects in the control group during the course of the experiment,
- Some subjects might have been athletes,
- Some subjects might have spent less time exercising than others,
- Some subjects might have eaten immediately after the exercise while others did not,
- Etcetera.

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 is not the true reason of better reading ability (the true reason might be age or education). Another example is the myth in ancient Germany where people believed that storks deliver babies (See here [64, 65] for a discussion of this). Notice that these were only few concerns regarding statistical testing. There are, of course, several other factors to be carefully thought about when running experiments and statistical testing.

Let me bring two examples related to computer vision. In the first example, let's say you have designed a system that tells whether a scene is captured in China or in the United States [2]. 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. In the second example, assume you have a model that predicts whether the resolution of an aerial image is low, intermediate or high. Accidentally, your low-, intermediate-, and high resolution terrain images come from areas covered with snow, rock, and vegetation, respectively. Now your model might have learned to classify different regions instead of resolution. All in all, since most of the computer vision models are data- driven, it is crucial to understand what aspects of the data they are capturing. This will help explain what models actually learn.

It is becoming increasingly popular to resort to inexpensive crowdsourcing platforms (e.g., Amazon Mechanical Turk [66]) to collect annotated data for training supervised data-hungry models [67]. This has pros and cons. The pros are a) such large scale datasets provide more statistical power and are rich for statistical hypothesis testing, and b) the learned models have high expressive power (since the stimulus set is a good representative of the real world). There are several cons associated with it as well. Firstly, there is less control over the stimuli. It is often very cumbersome (sometimes impossible) to inspect all images to see if they are appropriate. Secondly, there is less control over the data collection setup. It is hard to tell whether subjects are qualified, follow the procedure or are actively engaged in the task. Some strategies have been devised to ensure a certain degree of data quality including a) secretly injecting some images for which annotations are already known (a.k.a catch trials), b) designing grading tasks (e.g., asking workers to grade each other's work), and c) collecting multiple annotations for every image to reduce noise. Third, due to these uncertainties and noise in the data it sometimes becomes very challenging to conduct meaningful statistical tests. For instance, it is hard to verify the claim that CNNs (e.g., ResNet [34]) outperform humans on the ImageNet recognition task. To harness such problems, cognitive vision researchers have traditionally been conducting laboratory experiments where they had (almost) full control over the stimuli and subjects. But that is about to change due to the big data revolution. So, as time goes by, hopefully better ways will emerge to deal with these problems.

Note that not all agree that statistical testing is needed in computer vision. According to one of the reviewers of a prior version of this piece: "... clearly many of the papers do not require such a test. Indeed, most computer vision papers presenting an empirical study are reporting a pre-defined metric on a pre-defined data; hence, I see no value in a statistical test. Since statistical test is designed to handle unknowns about the metric and data you choose.". What I believe is that sometimes it might be acceptable, to some degree, to tweak parameters, crunch numbers, etcetera in order to build an engineering product, but when it comes to making precise scientific statements, conducting careful statistical tests becomes inevitable.

[2] Or telling whether an image is natural or generated by a generative model.

6 Dissemination of negative results

Dissemination of positive results is often straightforward. How about when results do not support a desired hypothesis or are inconclusive? Here, I discuss the issues surrounding communicating negative results. Some journals such as PLOS ONE, Journal of Negative Results in Biomedicine, and Journal of Articles in Support of the Null Hypothesis explicitly welcome negative, null, or inconclusive results. As of now, there is no systematic way of disseminating negative results in computer vision. Not only that, but also commentary and opinion papers like this one are often not welcome as journals prefer technical papers.

Computer vision has a unique model of publication. While there are several prestigious journals (e.g., IEEE PAMI, IJCV, CVIU, IVC) 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 ranging from 20% 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 to disclose 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.

ArXiv and blogs (e.g., [68]) are two 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 folk, sporadic, and noisy thoughts. Nevertheless, occasionally people exploit these venues for communication or settling a matter. For example, I came across an ArXiv paper [69] debating the results of a previously peer-reviewed published paper [70] which introduced a new biologically-inspired method for mitigating the adversarial perturbations in CNNs.

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 option for such discussions 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 together 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” (See [71] as an example). In all of these journals, all materials have to go through the peer review process. These practices enrich the scholarly work.

One point to stress here is that negative results should go through a thorough review process. Considering the intense race for publication, it only make sense to accept negative results that address problems which are important, challenging, and of high interest to the community. What we certainly do not need is a replete of negative results (or early progress reports) ending up in arXiv occupying the limited bandwidth of researchers.

How about publishing inconclusive results? I believe that such results are indeed publishable and treating them separately is hard. This is because obtaining a purely negative result in a data-driven discipline is puzzling since it needs to probe the entire parameter space and inspect lots of design choices.

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.

In addition to ArXiv and blogs, some other possibilities for publishing negative results are through specialized journals, conferences and workshops to the topic (e.g., http://negative.vision/) where papers can be presented and discussed in detail.

7 Epilogue

Here I summarize the main take-away lessons from this work.

- Firstly, solid, mature, trustworthy, important, interesting, well-documented, and peer-reviewed negative results that come from rigorous investigations should certainly be welcome. Such results (conclusive or inconclusive) 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 evaluation scores. One chief concern here is that hastily-derived and half-baked negative results can be dangerous and misleading. Notice that it is usually easier to obtain a negative result than a positive result (e.g., making a model work). What is very important, even more than sign, is the validity of the outcome. Overall, negativity towards negative results is counterproductive and such results should be published provided that they follow appropriate and sound scientific methodologies.
- Second, negative results should be properly and effectively disseminated, incentivized, shared, encouraged and discussed. Absence of negative results from the literature can cause serious problems (e.g., biasing researchers in certain directions). A mindset needs to be nurtured and encouraged to recognize and embrace such efforts (e.g., through dedicated journals and conferences to the topic). Emphasizing the statement by Torralba and Efros [72] that "too much value (is given) to winning a particular challenge", 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). If possible, it would be more effective to make the comments and discussions available to community.
- Third, 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. This is even more important when models start to saturate on a dataset raising the question of whether a 1% improvement in accuracy is meaningful. Overall, statistical analysis becomes increasingly more important as computer vision methods hit the market and become prevalent in daily life. Statistical analysis is also critical to support hypotheses and advance vision sciences. In this regard, students and computer vision researchers should be encouraged to strengthen their knowledge of statistics.
- Fourth, multi-disciplinary aspect of computer vision should be emphasized. Negative results and statistical testing are where the field can clearly learn a lot from from other fields, biological sciences in particular. To implement this, events that promote such emphasis should be encouraged (e.g., interdisciplinary tutorials and workshops in conferences, publication of interdisciplinary papers, invited talks, etc). In this respect, a synergy between computer and human vision can pay off well in the future. According to Rama Chellappa [73] "it is counterproductive if one or more groups of researchers to claim that their view of the elephant is the best one. Putting all our effort together, hopefully we can come up with a general solution to the grand problem of vision." This is particularly important in the age of deep learning where biologically-inspired neural networks have demonstrated a great promise.
- Fifth, perhaps less related to the main focus of this writing, is a concern on research attitude. Computer vision is enjoying a great phase of expansion and success. It might be even one of the fastest growing areas in computer science, thanks to industry backup. This has brought along a certain culture which is described well by Geman and Geman in "Science in the age of Selfies" [74]. These authors argue that researchers spend much of their time announcing ideas rather than formulating them. This partly has to be blamed on distractions caused by high information flow, web, social media, etc. Contributing to these, is the fact that researchers are rewarded for publishing more frequently than higher quality papers which has encouraged researchers to look for "minimum publishable units." This has led to a point that an overwhelmingly high number of publications appear each year making it practically impossible to track the progress even for senior scientists, let alone the newcomers. This concern has been raised by Alan Yuille [75]: "The conference cycle while adding dynamism often leads to a focus on short-term research, an emphasis on 'sound-bytes', and often small progress, improvements in performance on benchmarked datasets – rather than long-term quality research. This disrupts the balance between short-term research – picking the low-hanging fruit – and long-term research which builds the tools to pick the rest. We suggest reestablishing journal publications with rigorous peer review as the 'gold

standard' for referencing, for awarding prizes, for faculty appointments, and promotions." We should certainly benefit from relatively less emphasis on quantity of publications and relatively greater emphasis on quality of research. We need to openly and frankly discuss this in the community to find appropriate remedies.

References

1. J. P. Ioannidis, Why most published research findings are false, PLos med 2 (8) (2005) e124.
2. R. S. Shankland, Michelson-morley experiment, American Journal of Physics 32 (1) (1964) 16–35.
3. B. K. Horn, B. G. Schunck, Determining optical flow, Artificial intelligence 17 (1-3) (1981) 185–203.
4. R. Szeliski, Computer vision: algorithms and applications, Springer Science & Business Media, 2010.
5. Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (11) (1998) 2278–2324.
6. http://discovermagazine.com/2009/oct/06-brain-like-chip-may-solve-computers-big-problem-energy/.
7. https://www.caseyresearch.com/articles/brain-vs-computer.
8. A. Borji, L. Itti, State-of-the-art in visual attention modeling, IEEE transactions on pattern analysis and machine intelligence 35 (1) (2013) 185–207.
9. A. Nguyen, J. Yosinski, 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, 2015, pp. 427–436.
10. N. Kruger, P. Janssen, S. Kalkan, M. Lappe, A. Leonardis, J. Piater, A. J. Rodriguez-Sanchez, L. Wiskott, Deep hierarchies in the primate visual cortex: What can we learn for computer vision?, IEEE transactions on pattern analysis and machine intelligence 35 (8) (2013) 1847–1871.
11. W. J. Scheirer, S. E. Anthony, K. Nakayama, D. D. Cox, Perceptual annotation: Measuring human vision to improve computer vision, IEEE transactions on pattern analysis and machine intelligence 36 (8) (2014) 1679–1686.
12. N. Pinto, D. D. Cox, J. J. DiCarlo, Why is real-world visual object recognition hard?, PLoS Comput Biol 4 (1) (2008) e27.
13. J. J. DiCarlo, D. D. Cox, Untangling invariant object recognition, Trends in cognitive sciences 11 (8) (2007) 333–341.
14. N. K. Medathati, H. Neumann, G. S. Masson, P. Kornprobst, Bio-inspired computer vision: Towards a synergistic approach of artificial and biological vision, Computer Vision and Image Understanding 150 (2016) 1–30.
15. C. Tan, S. Lallee, G. Orchard, Benchmarking neuromorphic vision: lessons learnt from computer vision, Frontiers in neuroscience 9 (2015) 374.
16. K. Fukushima, S. Miyake, Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition, in: Competition and cooperation in neural nets, Springer, 1982, pp. 267–285.
17. A. Borji, L. Itti, Human vs. computer in scene and object recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 113–120.
18. R. VanRullen, Perception science in the age of deep neural networks, Frontiers in psychology 8.
19. N. Kriegeskorte, Deep neural networks: a new framework for modeling biological vision and brain information processing, Annual Review of Vision Science 1 (2015) 417–446.
20. D. D. Cox, T. Dean, Neural networks and neuroscience-inspired computer vision, Current Biology 24 (18) (2014) R921–R929.
21. T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, T. Poggio, Robust object recognition with cortex-like mechanisms, IEEE transactions on pattern analysis and machine intelligence 29 (3).
22. D. Parikh, C. L. Zitnick, Finding the weakest link in person detectors, in: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE, 2011, pp. 1425–1432.
23. J. Vogel, A. Schwaninger, C. Wallraven, H. H. Bülthoff, Categorization of natural scenes: local vs. global information, in: Proceedings of the 3rd symposium on Applied perception in graphics and visualization, ACM, 2006, pp. 33–40.
24. J. Deng, J. Krause, L. Fei-Fei, Fine-grained crowdsourcing for fine-grained recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2013, pp. 580–587.
25. F. Gosselin, P. G. Schyns, Bubbles: a technique to reveal the use of information in recognition tasks, Vision research 41 (17) (2001) 2261–2271.
26. M. C. Potter, Meaning in visual search, Science 187 (4180) (1975) 965–966.
27. S. Thorpe, D. Fize, C. Marlot, Speed of processing in the human visual system, nature 381 (6582) (1996) 520.
28. E. L. Denton, S. Chintala, R. Fergus, et al., Deep generative image models using a laplacian pyramid of adversarial networks, in: Advances in neural information processing systems, 2015, pp. 1486–1494.
29. C. Vondrick, H. Pirsiavash, A. Oliva, A. Torralba, Learning visual biases from human imagination, in: Advances in neural information processing systems, 2015, pp. 289–297.
30. R. Fong, W. Scheirer, D. Cox, Using human brain activity to guide machine learning, arXiv preprint arXiv:1703.05463.
31. D. L. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, J. J. DiCarlo, Performance-optimized hierarchical models predict neural responses in higher visual cortex, Proceedings of the National Academy of Sciences 111 (23) (2014) 8619–8624.
32. A. Borji, L. Itti, Defending yarbus: Eye movements reveal observers' task, Journal of vision 14 (3) (2014) 29–29.
33. Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.
34. K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1026–1034.
35. M. Jaderberg, K. Simonyan, A. Zisserman, et al., Spatial transformer networks, in: Advances in Neural Information Processing Systems, 2015, pp. 2017–2025.
36. D. H. Hubel, T. N. Wiesel, Receptive fields, binocular interaction and functional architecture in the cat's visual cortex, The Journal of physiology 160 (1) (1962) 106–154.
37. M. Riesenhuber, T. Poggio, Hierarchical models of object recognition in cortex, Nature neuroscience 2 (11) (1999) 1019–1025.
38. F. Anselmi, T. Poggio, Representation learning in sensory cortex: a theory, Tech. rep., Center for Brains, Minds and Machines (CBMM) (2014).
39. M. R. Greene, T. Liu, J. M. Wolfe, Reconsidering yarbus: A failure to predict observers task from eye movement patterns, Vision research 62 (2012) 1–8.

40. A. L. Yarbus, Eye movements during perception of complex objects, Springer, 1967.
41. 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) (1959) 30–34.
42. H. Lian, Y. Ruan, R. Liang, X. Liu, 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) (2015) 9068–9088.
43. R. Rosenthal, The file drawer problem and tolerance for null results., Psychological bulletin 86 (3) (1979) 638.
44. S. Plous, The psychology of judgment and decision making., Mcgraw-Hill Book Company, 1993.
45. L. D. Nelson, J. P. Simmons, U. Simonsohn, Let's publish fewer papers, Psychological Inquiry 23 (3) (2012) 291–293.
46. J. Devlin, S. Gupta, R. Girshick, M. Mitchell, C. L. Zitnick, Exploring nearest neighbor approaches for image captioning, arXiv preprint arXiv:1505.04467.
47. 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) (2007) 4–4.
48. N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, Vol. 1, IEEE, 2005, pp. 886–893.
49. C. Vondrick, A. Khosla, T. Malisiewicz, A. Torralba, Hoggles: Visualizing object detection features, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 1–8.
50. R. I. Hartley, In defense of the eight-point algorithm, IEEE Transactions on pattern analysis and machine intelligence 19 (6) (1997) 580–593.
51. R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Süsstrunk, Slic superpixels compared to state-of-the-art superpixel methods, IEEE transactions on pattern analysis and machine intelligence 34 (11) (2012) 2274–2282.
52. M. D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: European conference on computer vision, Springer, 2014, pp. 818–833.
53. J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, H. Lipson, Understanding neural networks through deep visualization, arXiv preprint arXiv:1506.06579.
54. 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) (2013) 108–119.
55. https://rationalwiki.org/wiki/extraordinary_claims_require_extraordinary_evidence.
56. R. M. Haralick, Computer vision theory: The lack thereof, Computer Vision, Graphics, and Image Processing 36 (2-3) (1986) 372–386.
57. http://www.nowozin.net/sebastian/blog/how-to-report-uncertainty.html.
58. S. Belia, F. Fidler, J. Williams, G. Cumming, Researchers misunderstand confidence intervals and standard error bars., Psychological methods 10 (4) (2005) 389.
59. http://scienceblogs.com/cognitivedaily/2008/07/31/most-researchers-dont-understa-1/.
60. C. W. Dunnett, A multiple comparison procedure for comparing several treatments with a control, Journal of the American Statistical Association 50 (272) (1955) 1096–1121.
61. D. Cox, B. Efron, Statistical thinking for 21st century scientists, Science Advances 3 (6) (2017) e1700768.
62. B. Efron, T. Hastie, Computer age statistical inference (2016).
63. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, Journal of the royal statistical society. Series B (Methodological) (1995) 289–300.
64. R. Matthews, Storks deliver babies (p= 0.008), Teaching Statistics 22 (2) (2000) 36–38.
65. https://priceonomics.com/do-storks-deliver-babies/.
66. http://www.mturk.com.
67. A. Kovashka, O. Russakovsky, L. Fei-Fei, K. Grauman, et al., Crowdsourcing in computer vision, Foundations and Trends® in Computer Graphics and Vision 10 (3) (2016) 177–243.
68. www.computervisionblog.com.
69. W. Brendel, M. Bethge, Comment on" biologically inspired protection of deep networks from adversarial attacks", arXiv preprint arXiv:1704.01547.
70. A. Nayebi, S. Ganguli, Biologically inspired protection of deep networks from adversarial attacks, arXiv preprint arXiv:1703.09202.
71. A. Borji, D. N. Sihite, L. Itti, Objects do not predict fixations better than early saliency: A re-analysis of einhäuser et al.'s data, Journal of vision 13 (10) (2013) 18–18.
72. A. Torralba, A. A. Efros, Unbiased look at dataset bias, in: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE, 2011, pp. 1521–1528.
73. R. Chellappa, Mathematical statistics and computer vision, Image and Vision Computing 30 (8) (2012) 467–468.
74. D. Geman, S. Geman, Opinion: Science in the age of selfies, Proceedings of the National Academy of Sciences 113 (34) (2016) 9384–9387.
75. A. L. Yuille, Computer vision needs a core and foundations, Image and Vision Computing 30 (8) (2012) 469–471.