NOVA IMS Information Management School

MAA

Mestrado em Métodos Analíticos Avançados
Master Program in Advanced Analytics

Data Labeling tools for Computer Vision:
a Review

Pedro Miguel Lima de Sousa Reis

Dissertation presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

DATA LABELING TOOLS FOR COMPUTER VISION: A REVIEW

by

Pedro Miguel Lima de Sousa Reis

Dissertation presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics

Advisor: Roberto André Pereira Henriques, PhD

November 2021

To my beloved Carolina, who is the purest and kindest human being I know, with whom my life is happier every day. To Paula and Miguel for their endless love and support. To Rui and Rita that always worry and wonder about what and how I am currently doing. To Manuel and Conceição Maria for always being there unconditionally, pushing me further and believing in me since I was a child, whose words of encouragement and tenacity still echo in my ears. To Rui, who has always let me stand on his shoulders. To João, Alexandrino, Miguel, Pinhal, Brites and Sarabando – each one with its superpower. This dissertation is dedicated to all my family and friends. Finally, I'd also like to acknowledge NTT DATA, formerly known as everis, and the Data & Analytics team, for the support and inspiration throughout all my master's degree as a full-time worker-student.

ABSTRACT

Large volumes of labeled data are required to train Machine Learning models in order to solve today's computer vision challenges. The recent exacerbated hype and investment in Data Labeling tools and services has led to many ad-hoc labeling tools. In this review, a detailed comparison between a selection of data labeling tools is framed to ensure the best software choice to holistically optimize the data labeling process in a Computer Vision problem. This analysis is built on multiple domains of features and functionalities related to Computer Vision, Natural Language Processing, Automation, and Quality Assurance, enabling its application to the most prevalent data labeling use cases across the scientific community and global market.

KEYWORDS

Review; Computer Vision; Image Annotation; Data Labeling software; Supervised Machine Learning; Methodologies and Tools

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AGI	Artificial General Intelligence
AI	Artificial Intelligence
API	Application Programming Interface
CAGR	Compound Annual Growth Rate
CDO	Chief Data Officer
CI/CD	Continuous Integration/Continuous Development
CNN	Convolutional Neural Network
COVID-19	Coronavirus Disease 2019
CV	Computer Vision
DARPA	Defense Advanced Research Projects Agency
FTE	Full-time Equivalent employee
GAN	Generative Adversarial Network
GPU	Graphics Processing Unit
HDR	High Dynamic Range
HITL	Human-In-The-Loop
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
IT	Information Technology
MIT	Massachusetts Institute of Technology
ML	Machine Learning
NER	Named Entity Recognition
NLG	Natural Language Generation
NLP	Natural Language Processing
QA	Quality Assurance
RPA	Robotic Process Automation
SaaS	Software-as-a-Service
SDK	Software Development Kit
SVR	Single-View Reconstruction
UI	User Interface
UX	User Experience

1. INTRODUCTION

The concept of Computer Vision (CV) has undergone many changes since the beginning of the 21st century. In 2003, Computer Vision was just envisioned as an exciting but disorganized field (Forsyth & Ponce, 2003), but ten years later this definition rapidly turned into a field that focused on the automated extraction of information from images to infer something about the world (Prince, 2012; Solem, 2012). Currently, scientific and business communities are already used to hear on the concept of CV in a newspaper headline highlighting an investment of millions of dollars on the resolution of current global challenges, such as fighting Coronavirus Disease 2019 (COVID-19) (Ulhaq et al., 2020) or preventing climate changes (Ramachandra, 2019).

As a field that seeks to extract information from image data automatically, current in-use solutions or systems work with Computer Vision through two different approaches. Handcrafted approaches, for example, apply CV techniques by using sets of rules to solve a specific challenge. Some use-cases of it are 1) using a surveillance camera to detect movement in a room; 2) detecting green or deforested areas in satellite images, and; 3) cropping some part of a document from a picture. On the other hand, Computer Vision solutions may also be associated with machine learning models or other Artificial Intelligence (AI) applications, such as object recognition and detection, image classification or anomaly detection solutions. These Computer Vision systems are often inspired by the properties and characteristics of human vision. Conversely, these algorithms can also offer insights into how the information extracted from images is interpreted in the human brain.

The ability of artificially intelligent systems to see like humans has been a subject of increasing interest and does not appear to be slowing down any time soon. However, the process of deciphering images, due to the more significant amount of data that needs analysis, is more complex than understanding other forms of binary information. Nonetheless, the usage of artificial neural networks is making computer vision more capable of identifying patterns from images than the human visual cognitive system (Scheirer et al., 2014).

Also, computer vision technologies will not only be less demanding to train, but also be capable of perceiving more from images than they are doing within the present. Together with other technologies or other subsets of AI, these can be used in order to build even more powerful and robust applications. For instance, image captioning techniques can be combined with Natural Language Generation (NLG) applications are used to decipher the surrounding objects for visually impaired individuals (Kim, 2020). In a near future, computer vision will play a vital role within the development of Artificial General Intelligence (AGI) and Artificial Superintelligence by granting the capacity to handle data as similarly or even better than the human visual system (Pueyo, 2018). Taking this into consideration, it can be hard for the scientific community to accept that today's computer vision capabilities together with its

applications and benefits remain unexplored. Because the evolution of Artificial Intelligence surprisingly supplements it and because it is being adopted in more and more challenges and industries, the future of computer vision seems to be full of promises and incredible results. The long run will clear the way for AI systems that are as human as us.

Be that as it may, there is a group of challenges that must undoubtedly be overcome first, being the demystification of the black box of AI one of the biggest (Burkart & Huber, 2021). This can be because the push for explainable, transparent applications also goes along with the drive for securing AI safety, as it is among the highest priorities for researchers. Furthermore, explainability might be one of the engines of innovation that surgically drive the researchers towards the opportunities for improvement in systems. Many computer vision applications, whereas being successful, are still undecipherable when it involves their inner workings. Until model explainability is attained, it is only prudent to use Machine Learning models or other AI applications in fields where the experimentation process is not harmful, and the risk of failing is not very high.

However, training efficiency is also important, especially in the Deep Learning context which is commonly applied in Computer Vision applications. Both model size and training data volume are increasing over time to obtain a competitive advantage in the market that is thirsty for better results and performance *per se* (Abbas et al., 2021; Tan & Le, 2021). However, the more data is needed for an AI system, the more it is necessary to invest in the process of data gathering, labeling and preparation and in the infrastructures that support the training and industrialization of these models respecting the DataOps architectures and principles (Ereth, 2018). Regarding the data labeling process, it will most likely experience a vast operational cost reduction in the future, through automation or collaboration between humans and artificial intelligence, while reducing its environmental footprint will certainly also be a concern in the short term, as shown in (Ligozat et al., 2021).

To better comprehend what the future holds for the computer vision field, it is necessary to understand its history. As such, Chapter 2 of this dissertation presents a historical review that contains the highlights since its conception. Subsequently, in Chapter 3, a contextual picture of the Data Labeling field in the scientific and industrial community/market is provided, along with references to related studies. Then, in Chapter 4, a comparative analysis of a vast selection of data annotation tools is conducted, highlighting the observations that fuel the discussion on the topic in Chapter 5. Lastly, conclusions are drawn in Chapter 6 and a list of limitations and recommendations for future work is given, in order to continue the pursuit of the optimal Data Labeling framework.

1.1. DATA LABELING

Data labeling or annotation is a common practice for many research fields that often involves multiple experts working in collaborative workflows (Dutta & Zisserman, 2019; Fiedler et al., 2019;

Pulford, 2005; Q. Zhang et al., 2015). Multiple iterations of data exploration, label discussion and labeling guidelines refinement might also occur to ensure the quality of the final guidelines. In the end, annotators can autonomously label additional data using those guidelines to produce consistent final labels endowed with a high level of assertiveness (Chang et al., 2017). Labeling data is, however, seen as a seemingly simple task required for training many machine learning systems but is in fact fraught with problems. This task is so far unavoidably tedious, especially when providing a sufficient amount of labeled data to some more complex approaches, as deep learning algorithms. For most machine learning projects, data can be labeled by the domain expert, who has the specific knowledge to make correct annotations. However, these domain experts usually lack proficiency in the labeling software. This compels the labeling process to be done by an AI service provider or an outsourced third-party needed, which can be very expensive and/or inaccurate. To lower the costs of this task, a complete and intuitive labeling tool is needed. There are further requirements to consider when choosing the right tool for data labeling (Dutta & Zisserman, 2019; Said et al., 2017):

- License compliance: if using an external tool, it might occur that the tool only permits non-profit-organizations to use it.
- Data security: when handling sensitive data, it is wise to avoid data storage in the software supplier servers.
- User experience (UX): the tool should be intuitively operable by collaborators with less technical experience and easy to set up.
- Use case coverage: the application should have enough functionality to cover different use cases in the future.
- Costs: it should not exceed the financial and time frame – the paid tools are usually faster, while the free ones might harm the project calendar.

As such, the motivation behind this dissertation is the study the current offer of image and video labeling tools through a detailed comparative analysis of the characteristics and functionalities of a selection of tools, in order to elect the most efficient tools for this process.

2. HISTORICAL OVERVIEW

To know where the most recent developments in Computer Vision are heading, it is necessary to understand not only its first embryonic steps but also its history as a whole. In this subchapter, we explore the birth of the Computer Vision field and its evolution since then, going through the most remarkable highlights of each decade.

It is commonly accepted that the father of Computer Vision is Lawrence Gilman Roberts (1937 – 2018) (Shapiro, 2020). He was an American engineer whose alma mater was Massachusetts Institute of Technology (MIT), where he received his bachelor's degree, master's degree, and PhD, all in electrical engineering. His PhD thesis dates back to 1963 and is considered as one of the foundational works of the field of Computer Vision, named "Machine Perception of Three-Dimensional Solids" (Roberts, 1963). In his seminal work, Roberts attempted to construct and display the full three-dimensional array of solid objects from a single two-dimensional image. This work enabled the usage of projective images formation models, where 3D lines map to 2D lines and polyhedral faces to polygons. This constituted a computational approach to a Single-View Reconstruction (SVR) that first extracts lines in an image and then matches the projected 3D lines of polyhedrons to the extracted lines (Roberts, 1963). Later in 2001, he earned one of the four "Father of the Internet" the Charles Stark Draper Prize awards from the U.S. National Academy of Engineering (DodgeSpecial, 2001).

Figure 1 – Evolution of the image generation capabilities by Generative Adversarial Networks (GANs) from 2014 to 2018 (Saxena & Cao, 2020)

In 1966, according to one well-known story, Marvin Minsky (1927 – 2016), an American cognitive and computer scientist who co-founded the MIT Computer Science and Artificial Intelligence Laboratory, challenged one of his undergraduate students Gerald Jay Sussman to "spend the summer linking a camera to a computer and getting the computer to describe what it saw" (Papert, 1966). The required time for this task was undoubtedly underestimated since there are still many research groups working on this topic in the present, more than 50 years later, like studying resource-efficient models (Kopuklu et al., 2019), Generative Adversarial Networks (GANs) (Saxena & Cao, 2020), Self-Supervised

Learning (Zhai et al., 2019) or Transformers and the usage of Self-Attention (Nguyen & Salazar, 2019). Thereupon, the Computer Vision field began to diverge from the already prevalent Digital Image Processing due to the researchers' desire to extract three-dimensional structures from images in order to achieve a full understanding of the scene (Azriel Rosenfeld & Kak, 2019; Azriel Rosenfeld & Pfaltz, 1966).

In the late 1960s, only the world's top universities could allocate funding for research in AI, such as MIT or the University of Cambridge. This decade marked the 10th anniversary of Alan Turing's Imitation Game (formerly known as the Turing test), in which Alan Turing discussed how to build intelligent machines and how to test their intelligence (Turing, 1950). Also in the 1960s, Isaac Asimov also declared the Three Laws of Robotics, a set of rules that triggered a panoply of discussions on machine ethics.

Soon after, in the 70s, many foundational algorithms started to be drawn and established until today, such as extracting edges from images, labeling of lines, stereo correspondence, optical flow, and motion estimation (Szeliski, 2010). Artificial Intelligence was already an ambitious market, innovative and full of scientific investment. Expectations were high, and, as such, failing the promised scientific advances was also striking. The United States Department of Defense research agency (DARPA) invested from 1971 to 1975 in a Speech Understanding Research program carried out by Carnegie Mellon University (Norvig, 2019). At the time, it was believed that the evolution of Speech Understanding techniques would be the basis for successfully achieving Speech Recognition. This hypothesis was later extinguished when confronted. The failure of this investment has led to severe discouragement and frustration for DARPA (Norvig, 2019).

On the other hand, in 1973, the paper “Artificial Intelligence: A General Survey”, known as *The Lighthill report*, was published (Agar, 2020). This report by James Lighthill was created for the British Science Research Council and emerged as a pessimistic prognosis for AI, conveying the idea that all the discoveries already made until then did not have the impact the researchers promised initially, mainly in the fields of Robotics and Natural Language Processing (NLP). British government investments in this area were advised against researchers' inability to transpose solutions to problems with a very restricted scope to more realistic problems. These events led to a societal demotivation towards AI and consequently Computer Vision, which triggered drastic falls in the investments in the field and severe criticism by the media. This phenomenon was later called *The Winter of AI*, having been a time of great pessimism (Norvig, 2019).

However, as scientific advances in Computer Vision gallop over time, it can be seen that the philosophy behind this field never ceases to follow it. As such, David Marr describes a visual information processing system in three different ways in his book released in 1982 (Marr, 1982) – the computational theory, the algorithms and representations, and the hardware implementation. To

explain these, Marr interrogates what the purpose of the computation or task is and what associated restrictions are already known or may be applied to that matter in the future, what algorithms are used to calculate the desired result and how are algorithms and representations mapped to specific specialized hardware and how can hardware restrictions be employed to pick an algorithm. Marr's last question has to do with the increase in the usage of Graphics Processing Units (GPUs) in Computer Vision, as this issue started to be relevant again in the end of the 80s decade (Szeliski, 2010). We can infer that Marr had at the time thought that scientific and statistical approaches should always go hand in hand with the engineering approaches, such as building efficient and robust algorithms, in order to design successful computer vision algorithms. Thus, almost 40 years ago, he was already embracing a philosophy that is still admirable for framing and solving more complex challenges that are still unresolved today.

Also, in the 1980s decade, researchers began to focus on more sophisticated techniques for image analysis, like the usage of image pyramids to perform tasks such as image blending and coarse-to-fine correspondence search (A. Rosenfeld, 1984). Later, image pyramids started to be displaced or augmented by wavelets in some applications. Regarding edge and contour detection methods, dynamically evolving trackers like Snakes (Kass et al., 1988) or the 3D physically-based models (Demetri Terzopoulos & Witkin, 1988) were developed. Plus, shape-from-X stereo techniques including shape from shading, shape from texture and shape from focus were perceived by the scientific community as capable of being described using the same mathematical framework if they were made more robust using regularization and placed as variational optimization problems (Bertero et al., 1988; Poggio et al., 1987; S. et al., 1987; D. Terzopoulos, 1983; Demetri Terzopoulos, 1986). By the same time, discrete Markov Random Field models were also pointed out as formulations of the same problems, enabling the use of better global search and optimization algorithms, such as Simulated Annealing (Aarts & Korst, 1987). On the other hand, 3D data processing continued to grow energetically during this decade (Besl & Jain, 1985; Faugeras & Hebert, 1986).

In the 90s, a lot of the previously mentioned topics continued to be explored. The most important development in computer vision was the expanded collaboration with computer graphics. Tracking algorithms including contour tracking such as Snakes (Kass et al., 1988), Particle Filters (Blake & Isard, 1998) and Level Sets (Malladi et al., 1995), as well as intensity-based techniques (Jianbo Shi & Tomasi, 1994; Lucas & Kanade, 1981; Rehg & Kanade, 1994), often applied to tracking faces (Lanitis et al., 1997; I. Matthews et al., 2007; J. Matthews & Baker, 2004) and bodies also improved a lot (Hilton et al., 2006; Moeslund et al., 2006; Sidenbladh et al., 2000). Techniques for image segmentation based on minimum energy (David & Jayant, 1989) and minimum description length (Leclerc, 1989), normalized cuts (Jianbo Shi & Malik, 2000) and mean shift (Comaniciu & Meer, 2002) were also developed. In addition, principal component *eigenface* analysis started to be applied in face recognition tasks (Debevec & Malik, 1997)

and linear dynamical systems for curve tracking (Blake & Isard, 1998), marking the appearance of statistical learning techniques.

Soon after, the 2000s decade was marked by the appearance of computational photography, which consisted of the remarkable appearance of image-based rendering techniques, such as capturing High Dynamic Range (HDR) images (Debevec & Malik, 1997) and panoramic image stitching. In addition, algorithms for automatically selecting overlapping image regions (Agarwala et al., 2004) and for merging images captured with flash with images captured without flash have also emerged (Petschnigg et al., 2004).

One of the fashions that arose in this decade was based on feature-based techniques combined with machine learning for object recognition (Fergus et al., 2007; Mundy, 2006). These techniques are also exemplified by scene recognition (J. Zhang et al., 2007) and location recognition (Brown & Lowe, 2007). Although feature-based techniques hold the largest share of research in this field at the time, there are also groups pursuing contour-based recognition (Belongie et al., 2002) and region segmentation (Mori et al., 2004).

On the other hand, the application of increasingly sophisticated machine learning techniques has also gained a lot of attention from researchers in visual recognition for computer vision problems, as is the example of the object detection framework Viola-Jones (Viola & Jones, 2001), which coincided with the increase in labeled data on the Internet, turning the learning of categories of objects more and more accessible.

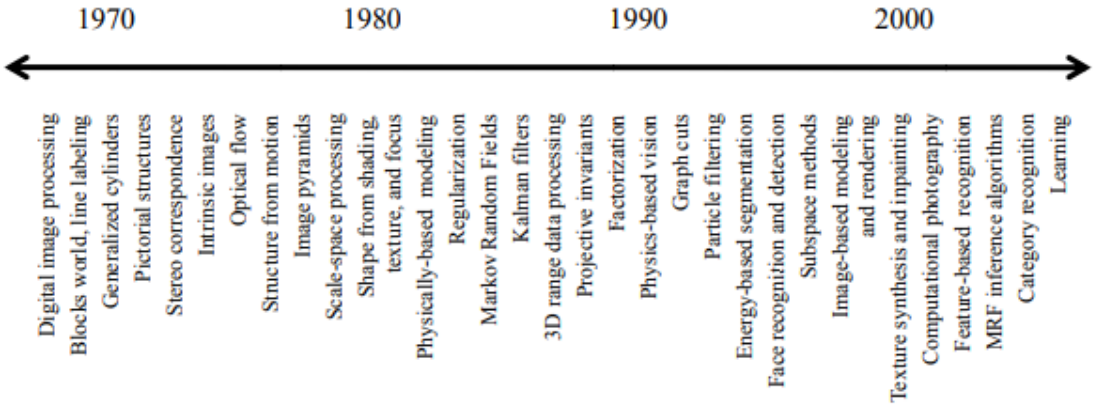

Figure 2 – Evolution of the most active research topics in the Computer Vision field over time (Szeliski, 2010)

Nonetheless, the biggest revolution in computer vision since the invention of computers themselves took place in the early 2010s, when researchers started using no hand-engineered features as the previous developments until then did. In 2012, a computer vision algorithm known as AlexNet achieved a 10% improvement over its competitors at the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) (Krizhevsky et al., 2017). This model relied on a Convolutional Neural Network

(CNN), and its breakthrough consisted in its ability to use a GPU to train the computer vision model significantly faster and for longer – AlexNet was trained over 6 days on two GPUs that were accessible to the consumer. Since then, Data Science and all its subdomains have evolved remarkably, both mathematically, in terms of the infrastructure it uses, and in terms of computing and large-scale processing. Finally, the global market has understood the advantages of applying the data-related technologies and methodological approaches invented to date, and has been able to materialize them either in products or services, or in the internal processes of organizations, such as in churn prediction and sales forecasting (Google, 2021). Even C-level executive positions such as Chief Data Officers (CDOs) have been created to manage the creation and governance of data processes and to ensure a data-driven culture in organizations.

The Computer Vision field has inherently benefited from this evolution, and is faced, perhaps for the first time, with the challenge of not only continuing to improve its performance, but also of optimizing the costs related to the necessary effort. This is where the data labeling component comes in, which may well be the driving force behind all future developments.

Figure 3 – Frequency that the term "Data Science" was searched for on Google, in the world and in multiple languages, from 2004 until the present (Google, 2021)

3. LITERATURE REVIEW

(Dutta & Zisserman, 2019) defined data labeling in the context of machine learning as the process of detecting and tagging data samples while attaching meaning and/or context to digital data. This process can be manual but is usually performed or assisted by software. Data labeling is an important part of data preprocessing for Machine Learning (ML), particularly for Supervised Learning. Both input and output data are labeled for classification to provide a learning basis for future data processing. For example, a system training to identify animals in images might be provided with multiple images of various types of animals from which it would learn the standard features of each, enabling it to correctly identify the animals in unlabeled images (Whytock et al., 2021).

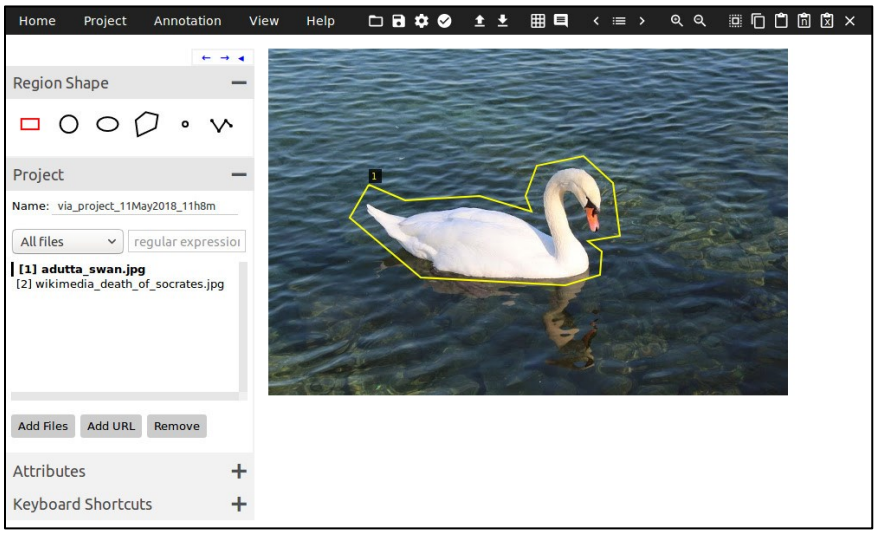

Figure 4 – Data Labeling annotation tool ImageTagger (Fiedler et al., 2019) framing an Object Detection task

Machine Learning and Deep Learning systems often require massive amounts of data to establish a foundation for reliable learning patterns. Data facilitated to the learning process must be labeled or annotated, which means that everything, or sometimes only the most important things, must be identified and localized in the image. It must also be labeled based on data features that help the model organize the data into patterns that produce a desired answer. A properly labeled dataset provides a ground truth that the ML model uses to check its predictions for accuracy and to continue refining its algorithm. Errors in this procedure impair the quality of the training dataset and the performance of any predictive models it is used for (Kshetri, 2021). To mitigate this, many organizations take a Human-In-The-Loop (HITL) approach, which is called a “data labeler” maintaining human involvement in training and testing data models throughout their iterative growth (Monarch, 2021). There are several procedures to structure and label data while maintaining human involvement (Fiedler et al., 2019). Either by using crowdsourcing, where a third-party platform gives an enterprise

access to many workers at once, and/or by using contractors, where an enterprise can hire temporary freelance workers to process and label data. A recent report from AI research and advisory firm Cognilytica found that over 80% of the time enterprises spend on AI projects goes toward preparing, cleaning, and labeling data (C. Research, 2019). Manual data labeling is the most time-consuming and expensive method, but it might be warranted for important applications (Fiedler et al., 2019). Some experts do believe that data labeling may present a new low-skilled job opportunity to replace the ones that are nullified by automation, because there is an ever-growing surplus of data and machines that need to process it to perform the tasks necessary for advanced ML and AI, which will create more and more low-skilled jobs and needs to hire more operational profiles (Kshetri, 2021).

Apart from that, the evolution of image classification models shows a clear upward trend in the emergence of new image classification models, stemming from an investment in research (Facebook AI Research, 2021). This is also true for semantic segmentation, language modelling, time series forecasting, speech recognition, among other methods (Wason, 2018).

Figure 5 – Evolution of Image Classification models on ImageNet dataset: Top 1 Accuracy (left) and Top 5 Accuracy (right) (Facebook AI Research, 2021)

Evidently, the evolution in data annotation techniques and software follows a similar trend in recent years, since hardware limitations are a hindrance in increasing the performance of the methods listed above, and research and development effort is directed towards this area. According to (G. V. Research, 2021), the global data annotation market was valued at US$ 695.5 million in 2019, is currently valued at US$ 1.66 billion and is expected to reach US$ 8.22 billion by 2028. The growing data annotation industry, which is expected to grow at a Compound Annual Growth Rate (CAGR) of 25.6% from 2021 to 2028 (G. V. Research, 2021), is expected to experience enormous expansion in the near future.

Plus, Gartner classified Data Labeling and Annotation Services as entering the *Trough of Disillusionment* (Figure 6) as this trend just walked by the *Peak of Inflated Expectations* (Gartner Inc., 2021). This means that this field has just surfed its wave of exacerbated hype and investment and it is finally slowing down while implementations fail to deliver. It typically happens before more instances

of how these services can benefit the enterprise start to take shape and become more broadly recognized, and is crucial to start gathering more maturity in the Data and AI market. Gartner estimates it to reach the *Plateau of Productivity* in 5 to 10 years (Gartner Inc., 2021). Thus, data labeling is a process that will constantly evolve and change to meet the business and technical objectives, that is, labeling tasks today are very prone to be different in three months. Through the time, a data labeling team evolves and finds better ways to label training data for improved quality and model performance, creating guidelines and sharing information on how to deal with the rarest use cases or scenarios.

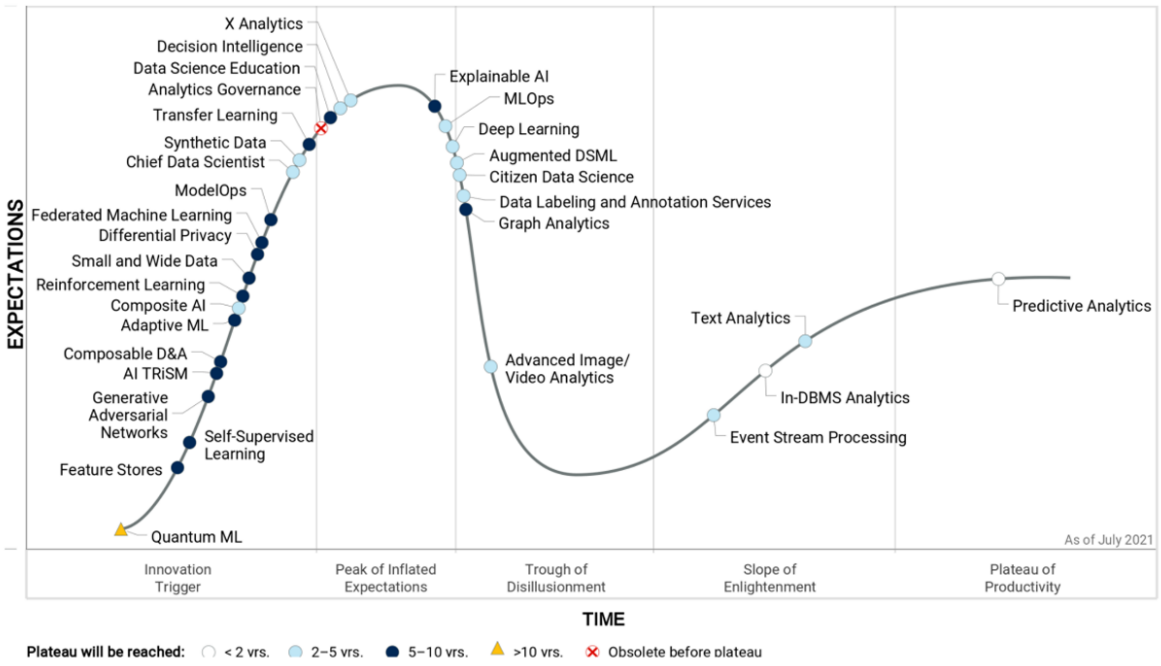

Figure 6 – Hype Cycle for Artificial Intelligence 2021 (Gartner Inc., 2021), as of July 2021

Relatively to the software, the best data labeling tools must be user-friendly in terms of User Interface and User Experience (UI/UX) and break the work down into atomic and smaller tasks to maximize labeling quality (Dutta & Zisserman, 2019). When a complex task is transformed into a set of atomic components, it is easy to measure and quantify each of those tasks. It also allows the identification of which tasks are best suited to humans and which ones can be automated. To optimize both data quality and the workforce investment, there are plenty aspects to consider when choosing the ideal data labeling tool. In the following chapter, an in-depth analysis and comparison of a hand-picked selection of tools is presented, focusing on the multiple functional and technical issues related to the topics of Computer Vision, Natural Language Processing (NLP), Automation, Quality Assurance (QA) and Management (Dutta & Zisserman, 2019; Said et al., 2017). These topics are dealt with as clusters of features or functionalities that are apparently prevalent within the data annotation tools and services market and might deviate slightly from our focus on the Computer Vision field.

Lastly, (Gaur et al., 2018) conducted a similar work to review the state-of-the-art in video annotation. However, given the latest market updates in Data Labeling services, this review lacks a view of the present and it doesn't analyze the prevalent concepts or features that are common to data labeling tools in structured terms.

4. DATA LABELING TOOLS

The tool selection that follows was performed based on data that was manually collected until July 2021. These tools were selected based on their general reputation, market adoption, and the features they provide to speed up and solve the data labeling task in Machine Learning problems. Even though this dissertation consists of a detailed and structured analysis, the definition of a criterion for the selection of the tools to be studied is not trivial, for several reasons. First, a panoply of new data labeling tools has been created and made available lately, given their demand, which makes choosing them difficult, as they are often released with very immature documentation. On the other hand, there are tools produced by large technological companies worldwide, and there are others that are developed by particulars as side-projects or even as hobbies, which makes their comparison impracticable due to the lack of resources associated with the latter. The increasing market pressure for developing new data labeling tools can be explained by its own needs and materialized in Figure 7, using Google Trends (Google, 2021) for the term "Supervised Machine Learning", that is immediately associated to the data labeling process because of its dependency on labeled data. Finally, personal experience and preference might have an undesirable impact on the meticulous definition of the study potential that tools may hold. Thus, the selection of the data labeling tools was based on the knowledge acquired throughout the work experience, from the sharing of people and forums of reference in the domain, from publications in scientific journals, from code sharing platforms, from news on market adoption, and from mentions in relevant conferences in the subject.

Figure 7 – Frequency that the term "Supervised Machine Learning" was searched for on Google, in the world and in multiple languages, from 2016 until the present

As such, Table 1 enumerates the considered tools that were analyzed in this dissertation, along with the respective reference and with each corresponding developing entity or brand, if applicable.

Tool Name	Developer/Brand	Reference
Colabeler	Colabeler	N/A
CVAT	Intel	A cost-effective, fast, and robust annotation tool (Said et al., 2017)
diffgram	Diffgram	N/A
ImageTagger	Hamburg Bit-Bots, University of Hamburg	ImageTagger: An Open Source Online Platform for Collaborative Image Labeling (Fiedler et al., 2019)
Label Studio	Heartex	N/A
Labelbox	LabelBox	N/A
LabelD	N/A	N/A
LabelImg	N/A	N/A
LabelMe	Computer Science and Artificial Intelligence Laboratory, MIT	LabelMe: A Database and Web-Based Tool for Image Annotation (Russell et al., 2008)
makesense.ai	makesense.ai	N/A
Playment	Playment, TELUS International	N/A
Ratsnake	N/A	Ratsnake: A Versatile Image Annotation Tool with Application to Computer-Aided Diagnosis (Iakovidis et al., 2014)
RectLabel	N/A	N/A
Remo.ai	Rediscovery.io	N/A
V7 Darwin	V7	N/A
VGG Image Annotation	Visual Geometry Group, University of Oxford	The VIA annotation software for images, audio and video (Dutta & Zisserman, 2019)
VoTT	Microsoft	N/A
COCO Annotator	N/A	N/A
EVA	N/A	N/A
SuperAnnotate	SuperAnnotate	N/A

Table 1 – Manually selected data annotation tools and respective developer teams and references, if applicable

During this revision, it was realized that a large part of the studied software tools does not have an associated article or published data. Information about their authors is also difficult to reach, which demonstrates that these tools are evidently divided into 3 groups: 1) tools that were developed specifically for commercial purposes; 2) tools that were developed according to the authors' own needs, and; 3) tools that are focused on scientific research and improvement in order to make the data annotation process as agile as possible. For these reasons, Table 1 displays developer team/brand as "N/A" when information about the software's authors is not publicly available or when they were a

dynamic and changeable team of contributors throughout time and where every developer had specific Information Technology (IT) knowledge and contributed to an Open-Source project.

Taking this into consideration, this first analysis demonstrates in a glance that the main developer teams or brands involved in Data Annotation tools or frameworks are big tech companies, specialized tech start-ups mainly based on Silicon Valley in the United States of America, or globally renowned academic research groups. It proves that nowadays, the data annotation tools are a priority for the most recognized IT companies and scientific entities around the world and their investments. These tools are often sold as SaaS (Software-as-a-Service), monetizing not only the product itself but also the optional service of outsourcing data labelers, constituting an extremely valuable asset to preserve, given the high demand that only tends to increase even further (Moulik, 2020).

Aiming a deep analysis of the technical features of the selected state-of-the-art data annotation tools, multiple clusters of functionalities coupled with their respective descriptions, grouped by functional fields of activity are followed. Each subchapter corresponds to a specific cluster, where the analyzed features help to peel off the multiple tools and a table is presented for that purpose, where the "✓" mark indicates the presence of a given functionality, "X" denotes its absence, and the "?" shows that there was no information available on that given subject at the time of this dissertation.

4.1. COMPUTER VISION FEATURES

Regarding CV-related functionalities, functionalities as *Bounding Boxes*, *Polygons*, *Lines*, *Key-points*, *Cuboids*, *Image classification* and *Video labeling* were selected to enable a full comparison between the selected tools.

- *Bounding Boxes*: Functionality that allows the user to draw rectangular Bounding Boxes that define the location of the target objects in static images, typically suitable for object detection tasks.
- *Polygons*: Tool lets the user to draw polygons to delimit objects in static images, usually needed for instance segmentation tasks.
- *Lines*: Functionality to draw lines or vectors, usually employed in autonomous driving applications in lane when annotating the lanes on the highways, for instance.
- *Key-points*: Permits labeling through the connection of key-points to build a skeleton and to understand more easily what is labeled, widely known for motion tracking, facial landmark detection and hand gesture recognition.
- *Cuboids*: Allows 3D data annotation, saving the depth and height of each object of interest. Usually applied on Object Detection for self-driving vehicles.
- *Image Classification*: Functionality that associates an image as a whole to a specific category.
- *Video Labeling*: Tool permits an easy navigation between frames of a video and make annotations for each sequential image. Can also deal with videos more complexly, using a model to estimate the position of a previously annotated object in the following frames.

To permit a better understanding about these tools' functionalities on Computer Vision, Table 2 shows how they are used through the previously selected data labeling tools.

Tool Name	Bounding Boxes	Polygons	Lines	Key-points	Cuboids	Image Classification	Video Labeling
Colabeler	✓	✓	✓	X	X	X	✓
CVAT	✓	✓	✓	✓	✓	X	✓
diffgram	✓	✓	✓	✓	✓	✓	✓
ImageTagger	✓	✓	✓	✓	X	X	X
Label Studio	✓	✓	✓	✓	X	✓	X
Labelbox	✓	✓	✓	✓	X	✓	✓
LabelD	✓	?	?	?	?	✓	?
LabelImg	✓	X	X	X	X	X	✓
LabelMe	✓	✓	✓	✓	X	✓	✓
makesense.ai	✓	✓	✓	✓	X	✓	✓

Playment	✓	✓	✓	✓	✓	X	✓
Ratsnake	✓	✓	✓	✓	X	✓	✓
RectLabel	✓	✓	✓	✓	✓	X	X
Remo.ai	✓	✓	✓	X	X	✓	X
V7 Darwin	✓	✓	✓	✓	✓	✓	✓
VGG Image Annotation	✓	✓	✓	✓	✓	✓	✓
VoTT	✓	✓	X	X	X	✓	✓
COCO Annotator	✓	✓	?	✓	?	?	?
EVA	✓	X	X	X	X	X	✓
SuperAnnotate	✓	✓	✓	✓	✓	✓	✓

Table 2 – Computer Vision related features in the selected data annotation tools

Table 2 shows that all the selected data annotation tools contemplate the possibility of drawing Bounding Boxes as a labeling task. Besides, polygons drawing for instance segmentation tasks is also very common. On the other hand, Cuboids, Video Labeling and Image Classification seem to be the least existing features in the data annotation tools. The lack of functionality on Cuboids drawing and Video Labeling can be explained by the fact that its use is oriented towards uncommon and very specific use cases which need information on depth of the objects or real-time video processing, respectively. Plus, companies that develop data annotation tools to meet this requirement tend to use them in-house only in order to get a competitive advantage. The lack of the functionality that permits Image classification can mean that the developer brands consider this task as attainable by a manual process, having nearly no cost benefit to develop it.

4.2. NATURAL LANGUAGE PROCESSING FEATURES

Although the NLP-related feature cluster is not directly related to the Computer Vision field, it still makes sense to consider it since it is composed by a subset of features which are apparently prevalent in the most relevant tools on the market, and therefore it should be a subsection of this dissertation. Regarding these NLP functionalities, two features were analyzed:

- *Text classification*: The tool provides a feature that permits the categorization of a text sample (i.e., corpus), which can be seen in use cases like e-mail spam detection or text sentiment analysis.
- *Text entity labeling*: The software has the capacity of labeling text tokens (i.e., words or expressions) as entities like organizations or localities, for Named Entity Recognition (NER) applications.

It is expected that a very small percentage of the analyzed tools will provide NLP-related functionalities, since the labeling of textual data can be done very differently from one task to another and is complex enough for a tool to be focused only on labeling for NLP. Thus, the two selected features are basic features that might be found across the selected list of tools, being present in Table 3.

Tool Name	Text Classification	Text entity labeling
Colabeler	✓	✓
CVAT	X	X
diffgram	✓	✓
ImageTagger	X	X
Label Studio	✓	✓
Labelbox	✓	✓
LabelD	?	?
LabelImg	X	X
LabelMe	X	X
makesense.ai	X	X
Playment	X	X
Ratsnake	X	X
RectLabel	X	X
Remo.ai	X	X
V7 Darwin	X	X
VGG Image Annotation	X	X
VoTT	X	X
COCO Annotator	X	X
EVA	X	X
SuperAnnotate	X	X

Table 3 – NLP-related features in the selected data annotation tools

As can be seen in the table above, the only tools from the selection that are capable of addressing NLP problems are Colabeler, diffgram, Label Studio and Labelbox. The small coverage of these features by the selected tools can be explained by the fact that in the vast majority of cases, businesses do not have as much difficulty in manually labeling the text excerpts they have in their possession to provide to an ML model as they do with images and/or videos. This, because performing data labeling on images or videos involves more technological expertise than on text excerpts. For data labeling on image media, an annotator must manually select or identify the specific pixels or patterns that constitute a given category, a task that the platforms or services usually don't enable. The same is not true for the annotation of textual data, which is often associated and/or classified by existing fill-in forms in tools used by businesses seeking the same annotated textual data to be consumed by downstream Machine Learning models, and also because information systems that store text are generally designed to store structured text data, like Robotic Process Automations (RPAs) for e-mail classification.

4.3. AUTOMATION AND DEVELOPER-FRIENDLY FEATURES

Regarding specifications that are intended not only to automate and/or assist the labeling process, but also to leverage the power of the tool and its implementation, two features were identified:

- *HITL labeling*: The data annotation process within a tool calls a Machine Learning model to assist labeling, often automatically suggesting labels and asking for human validation, reducing the labeling effort.
- *Custom add-ins Software Development Kit (SDK)*: The tool allows the development of custom add-ins through a Software Development Kit in order to meet custom requirements.

It is relevant to note that not every tool should support custom add-ins, because the developer company might want to restrain the functionality scope of their own labeling tool for the sake of their business model or possible cybersecurity issues, for instance. On the other hand, HITL labeling is a very optimistic feature for any data labeling tool as it reduces time and effort, making the service expensive. Relatively to these two functionalities, Table 4 shows which tools ensure them.

Tool Name	HITL labeling	Custom add-ins SDK
Colabeler	X	✓
CVAT	✓	X
diffgram	✓	✓
ImageTagger	X	X
Label Studio	✓	✓
Labelbox	✓	✓
LabelD	?	?
LabelImg	X	X
LabelMe	X	X
makesense.ai	✓	X
Playment	X	X
Ratsnake	✓	X
RectLabel	✓	X
Remo.ai	✓	✓
V7 Darwin	✓	✓
VGG Image Annotation	X	✓
VoTT	✓	X
COCO Annotator	✓	X
EVA	X	X
SuperAnnotate	✓	X

Table 4 – Automation and developer-friendly features in the selected data annotation tools

HITL labeling is supported by more than half of the analyzed tools, while spreading rapidly in the latest months through the data labeling services and tools – especially in cloud service providers and outsourced third-parties, as it saves time and effort for the companies. On the other side, only a minority gives space for the development of tailor-made add-ins, presenting itself as a feature in decline, because data labeling enterprise solutions are intended for a niche market and will most likely end up covering all its needs, and because the market pressure on the open-source tools will also tend to cover the needed functionalities. Only 5 out of the 20 tools (diffgram, Label Studio, Labelbox, Remo.ai and V7 Darwin) provide both functionalities.

4.4. MANAGEMENT AND QUALITY ASSURANCE FEATURES

As with most operational tasks, their optimization involves monitoring and consequent improvement. Thus, the comparative analysis in this subchapter combines management and quality assurance functionalities:

- *Quality Management*: The tool offers the possibility of analyzing label quality with reports or dashboards.
- *Project Management*: Support for project management, planning and task assignment.
- *Data Management*: Functionality to analyze created labels, class volumetry or other metrics on labeled data.
- *Consensus*: The tool offers the possibility of cross-checking labels and label precision between multiple users.
- *Benchmarks*: Functionality to evaluate each user label precision and quality.
- *Performance metrics*: Creation of metrics for performance monitoring and evaluation of each data annotator.

Just as importantly, all the selected Management and QA tool functionalities are shown in Table 5.

Tool Name	Quality Management	Project Management	Data Management	Consensus	Benchmarks	Performance metrics
Colabeler	X	X	X	X	X	X
CVAT	X	✓	X	X	X	X
diffgram	✓	✓	X	✓	✓	✓
ImageTagger	X	X	?	X	X	X
Label Studio	X	X	X	✓	X	X
Labelbox	✓	X	X	✓	✓	✓
LabelD	?	?	X	?	?	?
LabelImg	X	X	X	X	X	X
LabelMe	X	X	X	✓	X	X
makesense.ai	X	X	X	X	X	X
Playment	X	X	✓	X	X	X
Ratsnake	X	X	✓	X	X	X
RectLabel	X	X	X	X	X	X
Remo.ai	X	✓	✓	X	X	X
V7 Darwin	✓	X	X	✓	✓	✓
VGG Image Annotation	X	X	X	X	X	X
VoTT	X	✓	X	X	✓	X
COCO Annotator	X	X	X	X	X	X
EVA	X	X	X	X	X	X
SuperAnnotate	X	X	X	X	X	X

Table 5 – Management and QA features in the selected data annotation tools

Table 5 shows that the overwhelming majority of the tools that have functionalities to ensure the management of the labeling work also provide ways to review and validate the quality of the results obtained, proving the need to make the labeling process iterative. About half of the analyzed tools have some concern in this regard, from which we conclude that it is a type of functionality that the market demands.

4.5. GENERAL COMPARATIVE ANALYSIS

Table 6 is shown below as a wrap up of all the performed analysis, where the categories of features prevalent in each data labeling tool are easily pointed out. The presented values were calculated based on the ratio of covered functionalities under each of the four clusters (Computer Vision, NLP, Automation and QA) scope for each analyzed tool.

Tool Name	Computer Vision	Natural Language Processing	Automation & Developer	Management and QA
Colabeler	57%	100%	50%	0%
CVAT	86%	0%	50%	17%
diffgram	100%	100%	100%	83%
ImageTagger	57%	0%	0%	0%
Label Studio	71%	100%	100%	17%
Labelbox	86%	100%	100%	67%
LabelD	29%	0%	0%	0%
LabelImg	29%	0%	0%	0%
LabelMe	86%	0%	0%	17%
makesense.ai	86%	0%	50%	0%
Playment	86%	0%	0%	17%
Ratsnake	86%	0%	50%	17%
RectLabel	71%	0%	50%	0%
Remo.ai	57%	0%	100%	33%
V7 Darwin	100%	0%	100%	67%
VGG Image Annotation	100%	0%	50%	0%
VoTT	57%	0%	50%	33%
COCO Annotator	43%	0%	50%	0%
EVA	29%	0%	0%	0%
SuperAnnotate	100%	0%	50%	0%

Table 6 – Comparative summary of the features grouped by category across the analyzed data annotation tools

The comparison between the relative coverage of the functionality categories by tool demonstrates only a minority of the tools focus on the NLP-related features, unlike the functionalities related to Computer Vision. Moreover, the selected Automation and Developer-friendly features are apparently representative along most of the analyzed tools, and Management and QA functionalities are more present in the tools that were developed by a company.

5. DISCUSSION

The objective of this dissertation is to compare the selected tools in a structured way, with a view to their potential use for data labeling purposes for Computer Vision challenges. In order to evaluate the tools in their totality, this comparative analysis was extended to features that are unnecessary for CV topics, but that might eventually support an organization's decision in selecting a tool. As such, multiple functionalities were highlighted that focus not only on image, video, and text data, but also on the automation and optimization of the labeling process and its quality assurance. However, Table 6 shows that the selection of tools might be biased towards the focus on Computer Vision, which means that this analysis has more significance in this field.

Initially, it was planned to create a point system to evaluate each data labeling tool fullness, where the presence of each feature would add up one point for a tool, and the decision on the best tool would be judged only by the maximum number of points obtained throughout the analysis. However, this comparison would not be fair or realistic since, for example, the weighting or importance of NLP-related features is likely to be lower if the organization under consideration is a software house that develops deep learning models to interpret images. In addition, the financial factor may also have an impact on the choice and therefore the choice of the data labeling tool presented in this dissertation will only be according to the author's perspective and might differ according to the circumstances of a reader who belongs to a specific organization or development team.

Given the diversity of purposes of the analyzed tools, one or more tools will be chosen for each analysis conducted. Starting with the main analysis of the functionalities with Computer Vision, the tools that present all the analyzed functionalities stand out evidently: diffgram, V7 Darwin, VGG Image Annotation and SuperAnnotate. Regarding the pillar of NLP functionalities, the tools that allow text classification and the identification of named entities are Colabeler, diffgram, Label Studio and Labelbox. It should be noted that the pool of tools that offer these two functionalities is quite different when compared to the Computer Vision oriented one, as text features are extremely underrepresented. This is explained by the scope restrictions of the different products, which are probably aimed at different niche markets or academic tracks, as previously explained. Concerning HITL and being open to the inclusion of customizable add-ins, diffgram, Label Studio, Labelbox, Remo.ai and V7 Darwin are the winners, meaning that these are the tools with the most versatility for multiple use cases, while having a reduced effort rate needed to achieve the dataset annotation goal. As for the pillar that relates to features for Management and QA, the tools that stand out the most are diffgram, Labelbox, Remo.ai and V7 Darwin. The functionalities oriented to project management, task planning and assignment and data management are certainly underdeveloped in the labeling tools market overall, but are mostly present in the corresponding development roadmaps, which foresee that they

will be rolled out during 2021 or 2022. From the entire palette of reviewed tools, there are two tools that stand out from the competition – V7 Darwin and diffgram.

V7 Darwin has all the necessary features for data annotation for a Computer Vision ML project, which are developed in a robust, quality and extremely user-friendly way. Its major focus is on ease of annotation and automation of labeling using machine learning models, while being simplistic and having an apparently short learning curve. Thus, it is one of the most promising tools for image and video labeling, despite being expensive and not being open-source which is a great disadvantage *per se* for not relying on its own community to evolve.

On the other hand, diffgram is the most complete open-source choice that holistically covers the entire data lifecycle, from data ingestion and mining to integration with cloud or on-premise machine learning pipelines. In addition, it is a platform which is only paid for teams consisting of more than 20 users, which ensures data storage and versioning, data labeling for multiple tasks, workflow management, and data security - all these features are accessible directly from its platform or its Application Programming Interface (API). Moreover, it is based on a very active GitHub repository and has more than 500 stars. Plus, it ensures interpolation inside videos to absolve the labelers' effort in having to label every single frame in every timestamp of the video, using an object tracker and smart frame comparison heuristics. Unfortunately, it does not yet cover the functionality needed for NLP problems, although these will be on next year's development roadmap, as well as audio-related functionality.

6. CONCLUSIONS

This dissertation frames the evolution of the domain of Computer Vision, from its inception to the present. Past all the ups and downs of the field, data emerges as the new oil of the 21st century, so there is a global shift in the bleeding-edge market trends for data-driven culture. As such, the need arises to start optimizing the data labeling process, envisioning the generation of datasets with more and better data even faster, in order to minimize time effort and financial costs, without penalizing the labeling process. However, a panoply of tools was created with these intents, creating the need to constantly review the state-of-the-art and to strategically pick the right software for one's needs. For this, a selection of the tools with the most mentions in the scientific and industrial community is proposed, on which a comparative analysis is made to elect the most revolutionary and complete tools to perform data labeling and respond to Computer Vision or Machine Learning problems in general, in any organization. However, the labeling process is typically expensive and tedious and might even harm the feasibility of a project. Thus, the scope of this comparison seeks to mitigate and address the bottleneck associated with the data labeling phase in a Machine Learning project.

From the technical point of view, the comparative analysis of existing data labeling frameworks pointed to the victory of diffgram software, whose functionalities cover prodigiously the entire pipeline of a data project, featuring data ingestion, annotation, integration, exploration and production in the form of Machine Learning models. Diffgram is open-source and maintains a public development roadmap, leveraging the community participation in its evolution.

In terms of more business-related insights about this tool, diffgram has a free version for teams with less than 20 users and can prevent multiple errors and data redundancy while avoiding the daily imports and exports of data between different tools. By centralizing the entire pipeline in one tool, a team can: 1) shorten its own learning curve, which is a huge advantage given the market pressure to quickly train Full-time Equivalent employees (FTEs) in new technologies; 2) minimize potential security problems; 3) reduce licensing costs, and; 4) avoid redundancy of stored data. Moreover, the fact of being open-source opens a panoply of possibilities of quickly adding and testing new features to the product.

7. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

A practical component was planned to accompany this dissertation – the work was intended to consist in the creation of a new data labeling tool, which was which was quickly seen as too ambitious for a master thesis dissertation. Then, the practical component shaped itself into appending a new feature/functionality to an already existing data labeling tool. However, as this work advanced throughout the existing documentation on the multiple analyzed tools, the more it allowed to realize that a functionality designed and developed in less than one year would not be able to compete with a tool developed by a niche company or a large IT leader, since it would inevitably fall short of quality and complexity of the features already present in other tools as they are rolled-out and productized by large, specialized, developer teams with much more critical mass.

Also, the lack of premium/enterprise licenses in non-open-source tools was one of the limitations found during this work. As a recommendation for future work, it is suggested to contact the owners of these tools to request and obtain premium/enterprise licenses for academic purposes. The usage of such licenses may allow this work to go into more detail at the technical level and more obvious inference of how the backends of the tools work.

Once a license is given, one of the possible essential aspects to explore is the comparison of object trackers regarding video labeling. An object tracker that presents a better performance can also encourage the choice for its tool, as the video labeling process can be hugely optimized, as the tracker itself can spare the data labeler of annotating all the frames in a video while using information collected in previous video frames to help the consequent ones. Furthermore, future works could consider using multiple tools to achieve a labeled dataset while maintaining the same team of data labelers, in order to measure and compare the efficiency of each labeling tool in terms of effort and time consumed in a real-life scenario.

8. BIBLIOGRAPHY

Aarts, E., & Korst, J. (1987). Simulated Annealing: Theory and Application. *Simulated Annealing: Theory and Application*, 7. https://link-springer-com.ezproxy2.library.colostate.edu/content/pdf/10.1007%2F978-94-015-7744-1_2.pdf

Abbas, A., Sutter, D., Zoufal, C., Lucchi, A., Figalli, A., & Woerner, S. (2021). The power of quantum neural networks. *Nature Computational Science*, *1*(6), 403–409. https://doi.org/10.1038/s43588-021-00084-1

Agar, J. O. N. (2020). What is science for? The Lighthill report on artificial intelligence reinterpreted. *British Journal for the History of Science*, *53*(3), 289–310. https://doi.org/10.1017/S0007087420000230

Agarwala, A., Dontcheva, M., Agrawala, M., Drucker, S., Colburn, A., Curless, B., Salesin, D., & Cohen, M. (2004). Interactive digital photomontage. *ACM SIGGRAPH 2004 Papers, SIGGRAPH 2004*, 294–302. https://doi.org/10.1145/1186562.1015718

Belongie, S., Malik, J., & Puzicha, J. (2002). Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(4), 509–522. https://doi.org/10.1109/34.993558

Bertero, M., Poggio, T. A., & Torre, V. (1988). Ill-Posed Problems in Early Vision. *Proceedings of the IEEE*, *76*(8), 869–889. https://doi.org/10.1109/5.5962

Besl, P. J., & Jain, R. C. (1985). Three-dimensional object recognition. *ACM Computing Surveys (CSUR)*, *17*(1), 75–145. https://doi.org/10.1145/4078.4081

Blake, A., & Isard, M. (1998). Active Contours. In *Active Contours*. Springer London. https://doi.org/10.1007/978-1-4471-1555-7

Brown, M., & Lowe, D. G. (2007). Automatic panoramic image stitching using invariant features. *International Journal of Computer Vision*, *74*(1), 59–73. https://doi.org/10.1007/s11263-006-0002-3

Burkart, N., & Huber, M. F. (2021). A Survey on the Explainability of Supervised Machine Learning. *Journal of Artificial Intelligence Research*, *70*, 245–317. https://doi.org/10.1613/jair.1.12228

Chang, J. C., Amershi, S., & Kamar, E. (2017). Revolt: Collaborative crowdsourcing for labeling machine learning datasets. *Conference on Human Factors in Computing Systems - Proceedings*, *2017-May*, 2334–2346. https://doi.org/10.1145/3025453.3026044

Comaniciu, D., & Meer, P. (2002). Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(5), 603–619. https://doi.org/10.1109/34.1000236

David, M., & Jayant, S. (1989). Optimal Approximations by Piecewise Smooth Functions and Associated

Variational Problems. *Communications on Pure and Applied Mathematics*, *42*, 577–685.

Debevec, P. E., & Malik, J. (1997). Recovering high dynamic range radiance maps from photographs. *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques - SIGGRAPH '97*, *3*(1), 369–378. https://doi.org/10.1145/258734.258884

DodgeSpecial, J. (2001). *Draper Prize Honors Four "Fathers of the Internet."* https://www.wsj.com/articles/SB982004616905008338

Dutta, A., & Zisserman, A. (2019). The VIA annotation software for images, audio and video. *MM 2019 - Proceedings of the 27th ACM International Conference on Multimedia*, 2276–2279. https://doi.org/10.1145/3343031.3350535

Ereth, J. (2018). DataOps – Towards a definition. *CEUR Workshop Proceedings*, *2191*, 104–112.

Facebook AI Research. (2021). *PapersWithCode.com*.

Faugeras, O. D., & Hebert, M. (1986). The Representation, Recognition, and Locating of 3-D Objects. *The International Journal of Robotics Research*, *5*(3), 27–52. https://doi.org/10.1177/027836498600500302

Fergus, R., Perona, P., & Zisserman, A. (2007). Weakly supervised scale-invariant learning of models for visual recognition. *International Journal of Computer Vision*, *71*(3), 273–303. https://doi.org/10.1007/s11263-006-8707-x

Fiedler, N., Bestmann, M., & Hendrich, N. (2019). ImageTagger: An Open Source Online Platform for Collaborative Image Labeling. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 11374 LNAI* (pp. 162–169). https://doi.org/10.1007/978-3-030-27544-0_13

Forsyth, D., & Ponce, J. (2003). *Computer Vision: A Modern Approach*. Prentice Hall Professional Technical Reference.

Gartner Inc. (2021). *Gartner*. Hype Cycle for Artificial Intelligence 2021. https://www.gartner.com/en/information-technology/insights/artificial-intelligence

Gaur, E., Saxena, V., & Singh, S. K. (2018). Video annotation tools: A Review. *Proceedings - IEEE 2018 International Conference on Advances in Computing, Communication Control and Networking, ICACCCN 2018*, 911–914. https://doi.org/10.1109/ICACCCN.2018.8748669

Google. (2021). *Google Trends*. https://www.google.com/trends

Hilton, A., Fua, P., & Ronfard, R. (2006). Modeling people: Vision-based understanding of a person's shape, appearance, movement, and behaviour. *Computer Vision and Image Understanding*, *104*(2–3), 87–89. https://doi.org/10.1016/j.cviu.2006.09.002

Iakovidis, D. K., Goudas, T., Smailis, C., & Maglogiannis, I. (2014). Ratsnake: A Versatile Image Annotation Tool with Application to Computer-Aided Diagnosis. *The Scientific World Journal*, *2014*, 1–12. https://doi.org/10.1155/2014/286856

Jianbo Shi, & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(8), 888–905. https://doi.org/10.1109/34.868688

Jianbo Shi, & Tomasi. (1994). Good features to track. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition CVPR-94*, 593–600. https://doi.org/10.1109/CVPR.1994.323794

Kass, M., Witkin, A., & Terzopoulos, D. (1988). Snakes: Active contour models. *International Journal of Computer Vision*, *1*(4), 321–331. https://doi.org/10.1007/BF00133570

Kim, J. (2020). Application on character recognition system on road sign for visually impaired: Case study approach and future. *International Journal of Electrical and Computer Engineering*, *10*(1), 778–785. https://doi.org/10.11591/ijece.v10i1.pp778-785

Kopuklu, O., Kose, N., Gunduz, A., & Rigoll, G. (2019). Resource efficient 3D convolutional neural networks. *Proceedings - 2019 International Conference on Computer Vision Workshop, ICCVW 2019*, 1910–1919. https://doi.org/10.1109/ICCVW.2019.00240

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386

Kshetri, N. (2021). Data Labeling for the Artificial Intelligence Industry: Economic Impacts in Developing Countries. *IT Professional*, *23*(2), 96–99. https://doi.org/10.1109/MITP.2020.2967905

Lanitis, A., Taylor, C. J., & Cootes, T. F. (1997). Automatic interpretation and coding of face images using flexible models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*(7), 743–756. https://doi.org/10.1109/34.598231

Leclerc, Y. G. (1989). Constructing simple stable descriptions for image partitioning. *International Journal of Computer Vision*, *3*(1), 73–102. https://doi.org/10.1007/BF00054839

Ligozat, A.-L., Lefèvre, J., Bugeau, A., & Combaz, J. (2021). Unraveling the hidden environmental impacts of AI solutions for environment. In *Proceedings of ACM Conference (Conference'17)* (Vol. 1, Issue 1). Association for Computing Machinery. http://arxiv.org/abs/2110.11822

Lucas, B. D., & Kanade, T. (1981). *Iterative Image Registration Technique With an Application To Stereo Vision. 2*(April 1981), 674–679.

Malladi, R., Sethian, J. A., & Vemuri, B. C. (1995). Shape Modeling with Front Propagation: A Level Set Approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *17*(2), 158–175. https://doi.org/10.1109/34.368173

Marr, D. (1982). Vision: a computational investigation into the human representation and processing of visual information. In *Vision: a computational investigation into the human representation and processing of visual information.* W. H. Freeman.

Matthews, I., Xiao, J., & Baker, S. (2007). 2D vs. 3D Deformable Face Models: Representational Power, Construction, and Real-Time Fitting. *International Journal of Computer Vision*, *75*(1), 93–113. https://doi.org/10.1007/s11263-007-0043-2

Matthews, J., & Baker, S. (2004). Active appearance models revisited. *International Journal of Computer Vision*, *60*(2), 135–164. https://doi.org/10.1023/B:VISI.0000029666.37597.d3

Moeslund, T. B., Hilton, A., & Krüger, V. (2006). A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, *104*(2-3 SPEC. ISS.), 90–126. https://doi.org/10.1016/j.cviu.2006.08.002

Monarch, R. (Munro). (2021). *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Manning Publications Co.

Mori, G., Xiaofeng Ren, Efros, A. A., & Malik, J. (2004). Recovering human body configurations: combining segmentation and recognition. *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, *2*, 326–333. https://doi.org/10.1109/CVPR.2004.1315182

Moulik, S. (2020). Data as the New Currency—How Open Source Toolkits Have Made Labeled Data the Core Value in the AI Marketplace. *Academic Radiology*, *27*(1), 140–142. https://doi.org/10.1016/j.acra.2019.09.016

Mundy, J. L. (2006). *Toward Category-Level Object Recognition* (J. Ponce, M. Hebert, C. Schmid, & A. Zisserman (eds.); Vol. 4170). Springer Berlin Heidelberg. https://doi.org/10.1007/11957959

Nguyen, T. Q., & Salazar, J. (2019). *Transformers without Tears: Improving the Normalization of Self-Attention*. *1*. https://doi.org/10.5281/zenodo.3525484

Norvig, S. R. and P. (2019). Artificial Intelligence A Modern Approach 4th Ed. In *Journal of Chemical Information and Modeling* (Vol. 53, Issue 9).

Papert, S. (1966). *The summer vision project* (pp. 1–6). http://dspace.mit.edu/handle/1721.1/6125

Petschnigg, G., Szeliski, R., Agrawala, M., Cohen, M., Hoppe, H., & Toyama, K. (2004). Digital photography with flash and no-flash image pairs. *ACM Transactions on Graphics*, *23*(3), 664–672. https://doi.org/10.1145/1015706.1015777

Poggio, T., Torre, V., & Koch, C. (1987). Computational vision and regularization theory. *Readings in Computer Vision*, 638–643. https://doi.org/10.1016/b978-0-08-051581-6.50061-1

Prince, S. D. J. (University C. L. (2012). *Computer Vision: Models, Learning and Inference*. http://www.cambridge.org/9781107011793

Pueyo, S. (2018). Growth, degrowth, and the challenge of artificial superintelligence. *Journal of Cleaner Production*, *197*, 1731–1736. https://doi.org/10.1016/j.jclepro.2016.12.138

Pulford, G. W. (2005). Taxonomy of multiple target tracking methods. *IEE Proceedings: Radar, Sonar and Navigation*, *152*(5), 291–304. https://doi.org/10.1049/ip-rsn:20045064

Ramachandra, V. (2019). *Causal inference for climate change events from satellite image time series using computer vision and deep learning*. http://arxiv.org/abs/1910.11492

Rehg, J. M., & Kanade, T. (1994). Visual tracking of high DOF articulated structures: An application to

human hand tracking. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 801 LNCS* (pp. 35–46). https://doi.org/10.1007/BFb0028333

Research, C. (2019). *Data Engineering, Preparation, and Labeling for AI 2019: Technical Report*.

Research, G. V. (2021). *Data Collection And Labeling Market Size, Share & Trends Analysis Report*. https://doi.org/10.1109/CVPR.2004.1315182

Roberts, L. G. (1963). *Machine perception of three-dimensional solids*. *November*. http://dspace.mit.edu/handle/1721.1/11589

Rosenfeld, A. (1984). *Some Useful Properties of Pyramids*. 2–5. https://doi.org/10.1007/978-3-642-51590-3_1

Rosenfeld, Azriel, & Kak, A. C. (2019). Digital picture processing 2nd edition. In *Journal of Chemical Information and Modeling* (Vol. 1, Issue 1).

Rosenfeld, Azriel, & Pfaltz, J. L. (1966). Sequential Operations in Digital Picture Processing. *Journal of the ACM (JACM)*, *13*(4), 471–494. https://doi.org/10.1145/321356.321357

Russell, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2008). LabelMe: A Database and Web-Based Tool for Image Annotation. *International Journal of Computer Vision*, *77*(1–3), 157–173. https://doi.org/10.1007/s11263-007-0090-8

S., L. L., Blake, A., & Zisserman, A. (1987). Visual Reconstruction. In *Mathematics of Computation* (Vol. 53, Issue 188). https://doi.org/10.2307/2008745

Said, A. F., Kashyap, V., Choudhury, N., & Akhbari, F. (2017). A cost-effective, fast, and robust annotation tool. *Proceedings - Applied Imagery Pattern Recognition Workshop*, *2017-Octob*, 1–6. https://doi.org/10.1109/AIPR.2017.8457958

Saxena, D., & Cao, J. (2020). *Generative Adversarial Networks (GANs): Challenges, Solutions, and Future Directions*. *54*(3). http://arxiv.org/abs/2005.00065

Scheirer, W. J., Anthony, S. E., Nakayama, K., & Cox, D. D. (2014). Perceptual annotation: Measuring human vision to improve computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(8), 1679–1686. https://doi.org/10.1109/TPAMI.2013.2297711

Shapiro, L. G. (2020). Computer vision: the last 50 years. *International Journal of Parallel, Emergent and Distributed Systems*, *35*(2), 112–117. https://doi.org/10.1080/17445760.2018.1469018

Sidenbladh, H., Black, M. J., & Fleet, D. J. (2000). Stochastic Tracking of 3D Human Figures Using 2D Image Motion. In *Journal of Explosives Engineering* (Vol. 31, Issue 6, pp. 702–718). https://doi.org/10.1007/3-540-45053-X_45

Solem, J. E. (2012). *Programming Computer Vision with Python: Tools and algorithms for analyzing images* (Vol. 1, Issue 1). O'Reilly Media, Inc.

Szeliski, R. (2010). *Computer Vision: Algorithms and Applications* (Vol. 1). Springer-Verlag London.

https://doi.org/10.1007/978-1-84882-935-0

Tan, M., & Le, Q. V. (2021). *EfficientNetV2: Smaller Models and Faster Training*. http://arxiv.org/abs/2104.00298

Terzopoulos, D. (1983). Multilevel computational processes for visual surface reconstruction. *Computer Vision, Graphics, & Image Processing*, *24*(1), 52–96. https://doi.org/10.1016/0734-189X(83)90020-8

Terzopoulos, Demetri. (1986). Image Analysis Using Multigrid Relaxation Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-8*(2), 129–139. https://doi.org/10.1109/TPAMI.1986.4767767

Terzopoulos, Demetri, & Witkin, A. (1988). Physically Based Models with Rigid and Deformable Components. *IEEE Computer Graphics and Applications*, *8*(6), 41–51. https://doi.org/10.1109/38.20317

Turing, A. M. (1950). I. - Computing Machinery and Intelligence. *Mind*, *LIX*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Ulhaq, A., Born, J., Khan, A., Gomes, D. P. S., Chakraborty, S., & Paul, M. (2020). COVID-19 Control by Computer Vision Approaches: A Survey. *IEEE Access*, *8*, 179437–179456. https://doi.org/10.1109/access.2020.3027685

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, *1*, I-511-I–518. https://doi.org/10.1109/CVPR.2001.990517

Wason, R. (2018). Deep learning: Evolution and expansion. *Cognitive Systems Research*, *52*(August), 701–708. https://doi.org/10.1016/j.cogsys.2018.08.023

Whytock, R. C., Świeżewski, J., Zwerts, J. A., Bara-Słupski, T., Koumba Pambo, A. F., Rogala, M., Bahaa-el-din, L., Boekee, K., Brittain, S., Cardoso, A. W., Henschel, P., Lehmann, D., Momboua, B., Kiebou Opepa, C., Orbell, C., Pitman, R. T., Robinson, H. S., & Abernethy, K. A. (2021). Robust ecological analysis of camera trap data labelled by a machine learning model. *Methods in Ecology and Evolution*, *12*(6), 1080–1092. https://doi.org/10.1111/2041-210X.13576

Zhai, X., Oliver, A., Kolesnikov, A., & Lucas Beyer. (2019). S4L: Self-Supervised Semi-Supervised Learning. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1476–1485.

Zhang, J., Marszałek, M., Lazebnik, S., & Schmid, C. (2007). Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer Vision*, *73*(2), 213–238. https://doi.org/10.1007/s11263-006-9794-4

Zhang, Q., Song, X., Shao, X., Zhao, H., & Shibasaki, R. (2015). From RGB-D images to RGB images: Single labeling for mining visual models. *ACM Transactions on Intelligent Systems and Technology*, *6*(2),

1–29. https://doi.org/10.1145/2629701