Faculty of Mathematics

and Computer Science

How to write a machine learning related scientific article

Com˘anac Dragos, -Mihail

dragos.comanac@stud.ubbcluj.ro

Abstract

c 2022 .

Keywords:

Object detection; Deep learning

1. Introduction

Writing a scientific article in the domain of machine learning can be a challenging but rewarding task. It involves

organizing, and communicating complex ideas and concepts in a clear and concise manner. It also requires a strong

understanding of the topic, as well as the ability to conduct research, organize and analyze data, and communicate

effectively. As such, the main benefit is that it provides a very good way to share knowledge acquired in a very specific

domain such as machine learning.

Given all of these benefits, it is very tempting for a beginner researcher in machine learning to get into putting his

or hers new ideas into this form. Even though for someone accustomed to this kind of approach of expressing ideas,

writing a scientific paper may seem like second nature, for the first timers, it might seem like a more challenging task

than it actually is, or more challenging than the actual research subject. This problem might be especially worse in the

case of machine learning research, because the people involved into this field are most likely stronger with numbers

and code, rather than expressing an idea in a format that is to be ideally understood by other fellow researchers.

It is important to note that there are several alternatives to writing a traditional scientific paper that can be used

to communicate research findings and share information with the scientific community. Some researchers create online

resources or databases to share their research findings and make them more widely accessible to others. These

resources can be helpful for researchers who are looking for specific data or information on a particular topic. Many

scientists use social media or other online platforms to share their research findings and engage with the scientific

community. This can be an effective way to get feedback on ongoing work and to communicate research findings to a

wider audience. Also, some researchers create videos or interactive resources to communicate their research findings

in a more engaging and accessible way. These resources can be helpful for researchers who want to explain complex

concepts or ideas in an easy-to-understand way.

c 2022 .

2 Com˘anac Dragos, -Mihail

But these methods usually are not rigorous enough compared to a scientific paper, and can easily fade away in time

and be forgotten, as opposed to scientific papers which can be continuously cited.

Having all of these arguments in mind, the aim of this paper is to give a concise and easy to understand guide for

writing a research paper on a machine learning related topic. The target audience is composed of especially researchers

that are looking to write their first paper. Therefore, we aim to cover aspects from the structure and contents of the

paper, to dissemination of the paper.We also believe that the approach we present can be extended very easily to other

domains also or to bachelor or masters thesis.

2. Placement in the general field

Machine learning is a field of artificial intelligence (AI) that involves the development of algorithms and models

that can learn from data without being explicitly programmed. Machine learning algorithms can be trained on large

datasets and can automatically improve their performance over time by learning from the data.

There are two main categories of machine learning: supervised learning and unsupervised learning. In supervised

learning, the algorithm is trained on labeled data, which includes both input data and the corresponding correct output.

The algorithm uses this data to learn how to map the input data to the correct output. In unsupervised learning, the

algorithm is not given any labeled data and must discover patterns and relationships in the data on its own.

Machine learning has a wide range of applications, including image and speech recognition, natural language processing,

fraud detection, and predictive modeling. It is an active area of research and has the potential to revolutionize

many industries and fields.

3. Structure of a research article

To write a scientific paper, researchers typically follow a standard format and structure that includes an introduction,

a literature review, a methods section, a results section, and a discussion section. The goal of a scientific paper is to

communicate the results of a research study and contribute new knowledge to the field. The content and focus of a

scientific paper will depend on the specific topic and research question being studied, as well as the discipline and

field in which the research is being conducted.

In Fig. 1 we present a high overview of the structure of a scientific paper. The hourglass model is a framework

for organizing the content of a scientific paper that is often used to help writers structure their papers in a logical and

effective way. The model is called the hourglass because it starts with a broad introduction that sets the stage for the

research, and then narrows down to focus on the specific research question and methods, before expanding again to

discuss the results and implications of the research. Another point is that if the body is removed, the contents should

still make sense and be continuous.

Fig. 1. General structure of a scientific paper, also known as the king model [1]

Com˘anac Dragos, -Mihail 3

Overall, the hourglass model is intended to help writers structure their scientific papers in a way that is logical and

easy for readers to follow. It provides a framework for organizing the content of a scientific paper in a way that clearly

communicates the research question, methods, results, and implications of the study.

If the hourglass model represents the backbone of the paper, the king model gives the whole picture. As such, the

hourglass model is wrapped with the title, abstract and references.

3.1. Title

The title of a research paper is an important element of the paper and plays a number of important roles. First of

all, it should provide a clear and concise summary of the research: The title should clearly and concisely summarize

the main focus and key findings of the research.

A good title should be interesting and attention-grabbing, and should encourage the reader to want to read more

about the research. Also, the title should clearly identify the research area and the specific research question or problem

being addressed.

Research paper titles should generally be concise and should not be too long. A title that is too long may be

truncated when it is published or indexed, which can make it difficult for readers to understand the main focus of the

paper. Aim for a title that is no more than 12-15 words long.

The title should include key terms that are relevant to the research area and that will help readers to understand

the focus of the paper. These key terms should be carefully chosen to accurately reflect the content of the paper and

should be used consistently throughout the paper.

While it is important to use key terms in the title, it is also important to avoid using jargon or technical terms that

may be unfamiliar to many readers. The title should be written in plain language that is accessible to a broad audience.

A title that is too general or vague may not accurately reflect the content of the paper and may not be engaging for

readers. Aim for a title that is specific and focused, and that clearly conveys the research question or problem being

addressed.

Also, the title should be written with the intended audience in mind. It should be written in a way that is appropriate

for the target audience and should be written in language that they will understand.

Finally, the title should accurately reflect the content of the paper and should not be misleading or overly general.

In general, a good research paper title should be informative, concise, and engaging, and should help to convey the

key points of the paper to the reader. It is important to spend some time considering the title of a research paper, as it

can have a significant impact on the reader’s impression of the paper and their willingness to read it.

3.2. Abstract

An abstract is a brief summary of a research paper that is typically presented at the beginning of the document. We

can also see in in Fig. 1 that the abstract is simply a compressed version of the hourglass model, meaning that it covers

a little bit from each part, going from general to specific and back to general. It is typically a single paragraph that

provides an overview of the main points and findings of the paper. The abstract should be concise and should clearly

and concisely describe the purpose, methods, results, and conclusions of the research and it should be self-contained

so that readers can understand the main points of the paper without having to read the full paper. It should also include

relevant keywords and phrases that will help readers find and understand the paper. The abstract is important because

it gives readers an idea of the content and scope of the paper, and it can help them decide whether they want to read

the full paper or not.

The abstract should include the purpose of the research, the methods used to conduct the research, the main results

or findings of the research, and the conclusions or implications of the research. It should also include relevant keywords

and phrases that will help readers find and understand the paper.

The length of an abstract can vary depending on the specific guidelines of the journal or conference, but it is usually

around 200-300 words in length. It is important to write the abstract after the rest of the paper has been completed, so

that the main points and findings of the research can be accurately summarized. Although, it is interesting to note that

some researchers prefer starting with the abstract so that they have a good idea about what they want to explain in the

rest of the paper, and later come and refine the contents of the abstract. In the end is a matter of preference, but the

abstract should be reviewed at the end anyways and it might be helpful to ask someone else for an opinion about it..

4 Com˘anac Dragos, -Mihail

The abstract should be written in a clear and concise manner, using simple language that is easy to understand.

Technical terms or jargon should be avoided unless they are necessary. Also, the abstract should be written in present

tense, as it is a summary of the current state of the research and It is generally best to avoid using abbreviations in the

abstract, as they can be confusing for readers who are not familiar with them.

It is also important to note that the abstract should be formatted according to the guidelines of the journal or

conference where the paper will be published. This may include specific font, font size, and margin requirements.

In summary, the abstract is a crucial part of a research paper because it provides a brief overview of the main points

and findings of the paper, and it helps readers decide whether they want to read the full paper or not. The first sentence

of the abstract should be attention-grabbing and clearly state the purpose of the research and using transitional phrases,

such as ”the aim of this study is to”, ”the results show that”, or ”the conclusion was is that” can help to clearly connect

the different sections of the abstract.

3.3. Introduction

The introduction of a research paper is the section that comes at the beginning of the paper and provides an

overview of the main points and purpose of the study. The introduction serves several important purposes. It should

provide background information about the topic being studied, including any relevant theories or previous research

that have been conducted on the subject. This helps to establish the importance and relevance of the research.

The introduction should clearly state the research question or hypothesis that the study is attempting to address.

This helps to focus the study and guide the reader’s understanding of the rest of the paper.

It should also provide an overview of the structure of the paper, including the main sections or chapters and

their purpose. This helps to guide the reader through the paper and helps them understand the organization of the

information.

Finally, the introduction should be interesting and engaging, and it should make the reader want to read the rest

of the paper. As such, the introduction of a research paper is an important section that provides context, states the

research question or hypothesis, outlines the structure of the paper, and entices the reader to read on.

Regarding wording aspects, similar to the abstract, the introduction should be written in clear and concise language

that is easy to understand. Technical terms or jargon should be avoided unless they are necessary.

3.4. Body

The body of a research paper is the section that comes after the introduction and presents the main points and

findings of the study. It is the longest section of the paper and is typically divided into several chapters or sections,

depending on the specific requirements of the paper. The body of the paper should present the research in a logical

and organized manner, and it should include several elements such as a literature review, or placement in the general

field, methodology together with it’s results and how it compares to other approaches existent in the literature.

As it was the case up until this point, the body of the paper should be written in a clear and concise manner, and

it should be well-organized and easy to follow. Each section should be clearly labeled and should be structured in a

logical way, with subheadings and bullet points used as needed to break up the text and make it easier to read.

Another good practice is to use tables and figures which are useful tools for presenting and organizing data in a

research paper. They can help to clarify and simplify complex information, making it easier for readers to understand

and interpret the results of the study on which the paper is based.

While it is very good to use them, they should be used sparingly, by only including them if they are necessary to

understand some results or help in clarifying or highlighting some important points.

Usually, the body itself is organized in several sections, each with its own role that we will describe in what follows.

3.4.1. Literature review

This section usually comes right after the introduction and it is meant to give some context about the field of interest

for the research. Essentially, a literature review is a critical summary of what the scientific literature says about the

specific research field of interest. It is a systematic, explicit, and reproducible method for identifying, evaluating, and

synthesizing the existing body of recorded knowledge on a topic. It should be thorough, covering all of the important

and relevant studies on the topic.

Com˘anac Dragos, -Mihail 5

The purpose of the literature review is to provide context and background information for the research being

conducted. It should give the reader a sense of what has been done in the field, what the key findings are, and what

gaps in the literature exist that the research paper aims to fill.

As such, a literature review is not just a list of studies, but rather a critical evaluation of the research that has been

done on a topic. It should provide a balanced view of the literature and should not simply present one side of the

argument.

Besides the key findings of the research, the literature review should also cover any limitations or weaknesses in

the reviewed studies. Other important insights should be included such as patterns or trends in the research. This could

help in figuring out which gaps should be filled by the research paper.

For instance, if the focus of the research is to introduce a new neural network architecture, the literature review

should cover other similar neural network architectures. If the new neural network is a convolutional one, it might

be a good idea to compare it with the well known and established architectures such as MobileNets [3, 4] or residual

networks [2].

Finally, it is important to keep in mind that a literature review is an ongoing process. During the actual research,

new findings or other methods might arise that need to be addressed and included in the literature review.

3.4.2. Methodology

The methodology section is arguably the core and most important section of the paper. Here are described the

particular steps took into developing and testing the proposed machine learning model. If the other sections can be

somewhat extended to other topics as well, this one is the most specific to machine learning.

First of all, data is essential to any machine learning algorithm, therefore it should be clearly stated how was the

data used for training and testing collected. Its quality is crucial because it determines what the learning algorithm

can actually learn. This may include information about the sources of the data, any preprocessing or cleaning that was

done, and any sampling techniques.

Machine learning itself is a complex domain, comprising of several subdomains, each with its particularities and

type of data. For instance, the most popular type of learning is supervised learning.

The dataset must have a special structure in order to satisfy the supervised learning algorithms. Basically, such a

dataset must contain pairs of inputs and outputs from which a function that correlates them is learned.

These pairs must contain meaningful information. For example, it is a good idea to have the inputs to be samples

from the same distribution in order to have some patterns. These inputs can be anything ranging from simple feature

vectors to images or sequences of words and usually it is relatively easy to find lots of samples for input.

The tricky part is finding the correct output. Obviously, the outputs must be representative for the input and must

describe what it should be learned. This part of gathering the outputs is especially hard because it means that each

input must be processed and labeled. These labels can be quite costly, especially if expensive hardware or expert

people are used in the labeling process. For instance, medical data could be such an example. Also, cheap labels can

be quite costly in the end if there are a lot of inputs. And in this era of deep learning, a lot of data is required for an

algorithm to learn, which is a big problem from the labeling point of view. This need for an extensive amount of good

labels is one of the main pitfalls of supervised learning. As such, a lot of interest has been invested in unsupervised

methods, that do not require any labels, also in semi-supervised methods that learns from samples with and without

labels.

For instance, in the case of object detection, the inputs are images, and the output represent a set of bounding boxes

corresponding to the target objects in the image. They need to contain information about the class of the object and

about the location of the object in the image. A common way to describe the location is to give the coordinates of the

upper left and bottom right corners in the image scale.

In order to have a qualitative dataset, it is important to have labeling conventions. For instance, it is important to

clearly describe what is considered to be an object of interest. Otherwise, the learning algorithm might be confused.

Also, the input needs to be clear, at least for humans, because if a human can’t tell what is there, it would be difficult

for the learning algorithm.

Also, a common practice is to split the dataset into several parts. Firstly, a training dataset is necessary. Basically,

on this, the weights of the algorithm will be updated, meaning that the patterns that describe the data are extracted

from this information. Beside a training dataset, a validation dataset is necessary. No learning occurs on this data. As

6 Com˘anac Dragos, -Mihail

the name suggests, it is used to validate the quality of the algorithm. A common use is to save the models that perform

good on this dataset.

This need of several datasets comes from the problem of overfitting. The idea is that if the model is too complex

it will learn too well the data and it does not generalise well to unseen data. If we have a separation, we can spot

overfitting by comparing the loss or another metric on the training and validation sets. If the model overfits, a continous

improvement can be observed on the training set, while on validation the performance plateaus. As such, various

regularization methods can be applied in order to alleviate this issue. This is a complex problem on its own.

Also, sometimes a training and validation datasets are not enough. It can happen that the by setting various hyperparameters

of the learning algorithm, the validation set is also overfit. Therefore a new dataset, called test dataset is

used. The idea is that the final performance measurements are done on it because it is the only data that is not seen

during optimization, and such it would provide an idea about the performance of the algorithm in a real life scenario

with new data.

Before deep learning, a common choice was to have a larger proportion of the data dedicated to validation or testing

such as 20-30%. This was when the datasets were smaller. But nowadays, if the dataset is large enough, this percents

can be lower. For example, if the dataset contains millions of images, it may be enough to have only a fer percents

dedication for validation and testing each. The intuition is that it is more important to have data on which the model

can actually learn. This works best if the dataset splits are diverse enough, such that the test set is representative.

3.4.3. Results and comparison

3.5. Discussion

3.6. References

4. Dissemination of a research paper

5. Conclusions

References

[1] Derntl, M., 2014. Basics of research paper writing and publishing. International Journal of Technology Enhanced Learning 6, 105–123.

[2] He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image Recognition, in: IEEE Conference on Computer Vision and

Pattern Recognition (CVPR), pp. 770–778.

[3] Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., 2017. MobileNets: Efficient Convolutional

Neural Networks for Mobile Vision Applications. CoRR abs/1704.04861.

[4] Sandler, M., Howard, A.G., Zhu, M., Zhmoginov, A., Chen, L.C., 2018. MobileNetV2: Inverted Residuals and Linear Bottlenecks. IEEE/CVF

Conference on Computer Vision and Pattern Recognition , 4510–4520.