December 2022

Automatic Expansion of Data Sets for Machine Learning Models Using Crowdsourcing

Tomasz Mikolajewski

AUTOMATIC EXPANSION OF DATA SETS FOR MACHINE LEARNING MODELS USING CROWDSOURCING

Introduction

Machine learning models generally require a diverse and exhaustive data set in order to create a diverse and complete model. However, creating a diverse and complete model is difficult, expensive, and challenging. Traditionally, collecting such a diverse and exhaustive data set requires building a specification on what data to collect and translating it into multiple languages. Additionally, designing an accurate specification that results in collecting high quality data is especially challenging. The collection of data is often accomplished through working with crowdsourcing entities; however, these entities often do not have the expertise required to design accurate specifications.

Providing accurate specifications in multiple languages is even more challenging, resulting in poor quality data. In particular, instructions are overly technical and are often misunderstood when they are not provided to tech savvy users. As a result, translations that accurately cover all subtlety are not obtained and the diversity of the data collected is limited. Providing users with the ability to automatically expand a data set for a machine learning model by using crowdsourcing will ensure diversity and completeness of the data set, an accurate specification in multiple languages, and a complete machine learning model that can serve users around the world.

Summary

Computer-implemented systems and methods for providing automatic expansion of data sets for machine learning models with the disclosed technology can be accomplished by

automatically creating crowdsourcing tasks to send to users around the world to describe the data sets in their language and returning the expanded data set for a machine learning model.

In some instances, a user may upload a data set to the system. In response, the system may automatically create crowdsourcing tasks. The crowdsourcing tasks may ask users around the world (i.e., crowd workers) to describe the provided data set in their local language. The system may automatically determine the project specification for the task and institute crowdsourcing to ask users around the world to collect data according to the descriptions that were collected, instead of the user creating a project specification. The user may then receive the resulting expanded data set with the data points that have been collected from crowd workers around the world.

For example, a user may upload a photo of a cat to the system. The system can automatically create crowdsourcing tasks with instructions for crowd workers around the world to describe the photo and send the photo to the crowd workers. The crowd workers may then describe the photo as a "cat" in their local language. The system can use the crowd workers' descriptions of the photo to automatically determine a project specification for the task and create more crowdsourcing tasks to collect additional photos of cats and annotate them with text indicating "cat." The system may send to the same group of crowd workers, or a different group of crowd workers, the request to provide more annotated photos of cats. The user who uploaded the photo of the cat can then receive the expanded data set to use for a machine learning model.

In some instances, a project specification may be automatically determined by the system in order to complete a machine learning project, instead of a user manually creating a project specification, thus reducing the time of the data collecting process. The project specification may specify the data to be used for the machine learning model, how to collect the data, and how to

design the project in order to result in high quality and diverse data. The project specification may be automatically determined based on the descriptions of the data that were provided by the crowd workers in the crowdsourcing task that asked them to describe the data in their language. For instance, the project specification may indicate that the data to be collected is data that matches the descriptions provided and that the data is to be annotated in the languages used in the crowdsourcing task. Additional crowdsourcing tasks may then be automatically created to instruct crowd workers around the world to collect data according to the details in the project specification. The project specification may be automatically provided in multiple languages in order for crowd workers all over the world to understand the data collecting task, instead of the user translating the task into multiple languages.

For example, a user may upload a basic data set to the system, such as a photo of a cat, and the crowdsourcing tasks that were automatically created to ask crowd workers around the world to describe the photo may return descriptions indicating the photo is of a cat. Instead of a user creating a project specification that specifies that more photos of cats are to be gathered and annotated in multiple languages, the system may automatically create such a project specification. Another crowdsourcing task can then be automatically created to instruct crowd workers around the world to collect photos of cats and annotate them as “cat” in their language. All photos collected can then be added to the data set, resulting in a diverse and complete data set of annotated photos of cats in multiple languages to be used for a machine learning project.

Detailed Description

Figure 1 depicts an example computing system 100 in which systems and methods in accordance with the present disclosure can be executed. The computing system comprises a user computing device 102 containing one or more processors 112, memory 114 which may contain

data 116 and instructions 118 configured to carry out the methods disclosed herein, and a user input component 122. The user input component can be, for example, a touch display or physical buttons within the user computing device 102. The computing system 100 further comprises a network 180 and a server computing system 130. The server computing system 130 comprises one or more processors 132, and memory 134 which may contain data 136 and instructions 138 configured to carry out the methods disclosed herein. It should be appreciated that any combination or order of systems and methods disclosed herein can be performed on the user computing device, server computing system, or similar. For example, all processes can be performed on the user computing device 102 or the server computing system 130.

Figure 2 depicts an example embodiment of automatically expanding a data set for a machine learning model by automatically creating crowdsourcing tasks to describe the data in multiple languages 200 according to the present disclosure. A user 202 may upload a data set 204 to the system. The data set 204 may include any type of data to be used to train a machine learning model, such as photos, videos, handwriting samples, facial expressions, or landmarks. For example, the user 202 may upload a data set 204 consisting of photos of crops. The system may then automatically create crowdsourcing tasks 206 that ask users around the world to describe the data set 204 in their local language. For example, the data set 204 may contain photos of crops and the crowdsourcing tasks 206 may instruct the crowd workers to label each photo of a crop in the data set 204 with the name of that crop in their local language. The system may then create more crowdsourcing tasks 208 to ask users around the world to collect data according to the descriptions that were collected in the previous crowdsourcing tasks 206. The data collected for the crowdsourcing tasks 208 may then be annotated with the same label that the crowd workers provided when describing the data in the prior crowdsourcing tasks 206.

For example, the crowdsourcing tasks 208 may ask users to provide more photos of crops that match the descriptions they provided in the previous crowdsourcing task 206, which can also be labeled with the same description that was provided previously. The system may then return the expanded data set 210 containing data points collected from the crowd workers all over the world. For example, the expanded data set 210 may contain the original photos of crops that were provided in the basic data set 204, the labeled photos that were described by the crowd workers in the first crowdsourcing tasks 206, and the labeled additional photos collected in the second crowdsourcing tasks 208.

Referring now to Figure 3, an example embodiment of automatically determining a project specification and collecting data according to the project specification 300 according to the present disclosure. A data set 302 may be uploaded by a user. The data set 302 may include any type of data to be used to train a machine learning model, such as photos, videos, handwriting samples, facial expressions, or landmarks. For example, the data set 302 may include a photo of a cat 304. Crowdsourcing tasks may be automatically created by the system to provide the photo to crowd workers around the world and ask them to describe the photo 304 in their local language. The crowd workers may describe the photo in their local language 306 and the descriptions may be returned to the system. The system may then use those descriptions to generate a project specification with instructions to crowd workers to collect data consisting of photos of cats and to annotate them with the text “cat” in their language 308. The project specification may be automatically provided in multiple languages, such as the languages that the crowd workers used to describe the photo of the cat in their local language 306, without the need to manually translate the task into such languages.

Additional crowdsourcing tasks can be automatically created by the system to instruct crowd workers around the world to collect photos of cats and annotate them with the text “cat” in their language, according to the project specification. The crowdsourcing task asking for annotated collections of cat photos may be sent to the same crowd workers who described the photo in their local language or to different crowd workers around the world. The crowd workers may return collections of photos of cats that are annotated with the word “cat” in their local language 310 to the system. The collections of photos of cats that are annotated with the word “cat” 310 may then be added to the expanded data set 312, consisting of multiple photos of cats annotated as “cat” in multiple languages that were collected automatically by creating crowdsourcing tasks and project specifications. The result is high quality and diverse data, as the data set consisting of one photo of a cat 302 has been expanded automatically into a diverse data set consisting of multiple photos of cats each annotated in multiple languages 312.

The data set 312 may be further expanded by repeating the process of describing a photo 304 in the local language, gathering more photos 306, annotating the photos in the local language 310, and adding the photos to the expanded data set 312. For example, a photo of a sport shoe 304 may be provided and crowd workers may describe the photo in their local language 306 as a “shoe.” Additional crowdsourcing tasks can be automatically created by the system to instruct crowd workers to collect data consisting of photos of shoes and to annotate them with the text “shoe” in their local language 308. The crowd workers may primarily return photos of sport shoes, as well as some photos of shoe types local to the crowd worker’s location, such as Japanese geta shoes, which can be annotated with the word “shoe” 310 and added to the expanded data set 312. In order to increase the diversity of the data set 312, the process may then be repeated to allow local shoe types to be collected. A photo of a Japanese geta shoe 304 may

be provided, crowd workers may describe the photo in their local language 306, and additional crowdsourcing tasks may be automatically created to further instruct crowd workers to collect data consisting of photos of Japanese geta shoes 308. The collected photos of Japanese geta shoes can then be annotated in the local language 310 and added to the expanded data set 312. In another example, the user 202 may select an image 304 that the user would like to be described and annotated in other languages, or the system may identify rare images in the data set 302 or 312 to be described and annotated in multiple languages. The data set 312 may then be further expanded by repeating the process of describing the image 304 provided by the user or identified by the system in the local language, gathering more photos 306, annotating the photos in the local language 310, and adding the photos to the expanded data set 312.

Figures

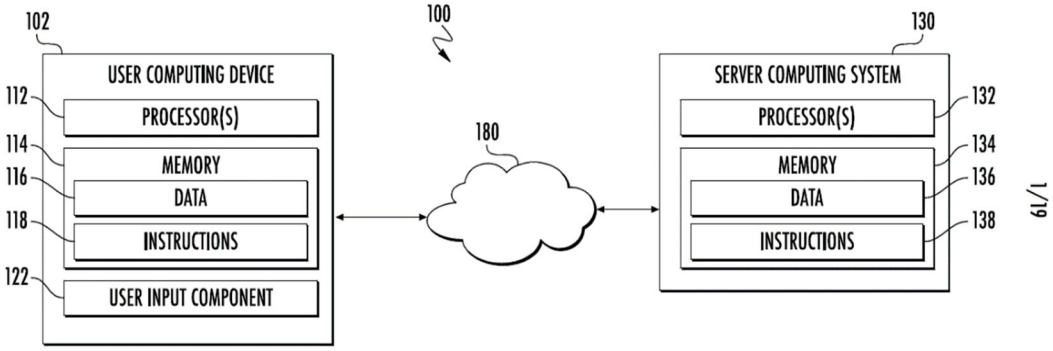

FIG. 1

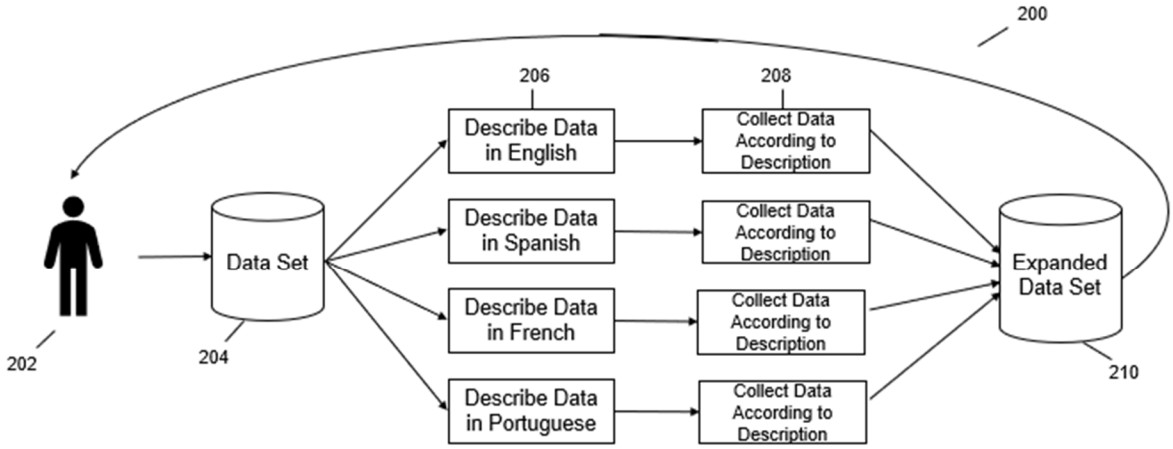
200
206
208
Describe Data in English
Collect Data According to Description
Describe Data in Spanish
Collect Data According to Description
Describe Data in French
Collect Data According to Description
Describe Data in Portuguese
Collect Data According to Description
Data Set
Expanded Data Set
202
204
210

FIG. 2

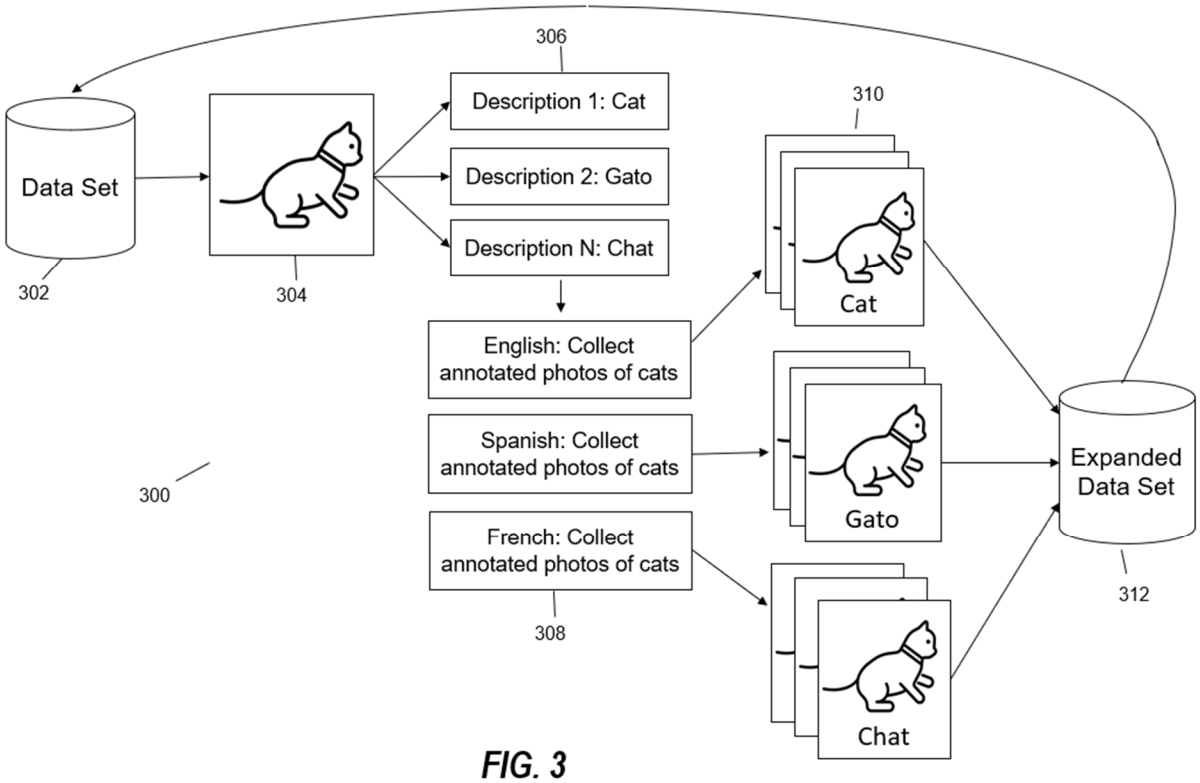
306
Description 1: Cat
Description 2: Gato
Description N: Chat
310
Data Set
Cat
English: Collect annotated photos of cats
Spanish: Collect annotated photos of cats
Gato
Expanded Data Set
French: Collect annotated photos of cats
Chat
300
302
304
308
312

FIG. 3

Abstract

The present disclosure describes computer-implemented systems and methods for automatic expansion of data sets for machine learning models by automatically creating crowdsourcing tasks to send to users around the world to describe a data set in their language and returning the expanded data set for the machine learning model. A user may provide an initial data set and receive a diverse and complete expanded data set with annotations in multiple languages in less time and without any additional effort from the user than in traditional data gathering for machine learning models.