Title of the Invention

"A SYSTEM TO ENHANCE ACCURACY OF A COMPUTER VISION MODEL"

Technical Field of the Invention

[1] The present invention generally relates to an image processing system. More particularly, the present invention is related to a system and method for enhancing the accuracy of computer vision machine learning models.

Background of the Invention

[2] Machine Learning techniques have completely revolutionized the computer vision field in recent years. Implemented correctly, these approaches have proved to be both flexible across multiple problem spaces as well as very robust to changes and fluctuations in diverse input data. However, machine learning based computer vision systems require substantial computational resources, as well as significant time and effort in acquiring sufficient and representative training data and training the models.

[3] It is therefore tractable to use simpler models that can be trained on less data and require less computational resources during inference. The downside of such simpler models is that they rarely reach a comparable level of accuracy as more complex models. We can exemplify this by comparing two recent (state-of-the-art) models, one is CenterNet-ResNet50 (simpler) and the other one is EfficientDet-D7 (complex), both are developed in 2019 and trained on the same set of data. Both the

models perform the same function i.e detecting, locating and classifying 80 different types of objects in an input image.

[4] On the same hardware, the simpler model is able to analyze images at more than 10 times the speed of the complex model, but suffers from lower accuracy. The Mean Average Precision (MAP) measured on the evaluation data is 29.5 for the simpler model, compared to 51.2 for the complex model (higher is better). Conventionally, models such as these, make use of very little image preprocessing, for both training and inference.

[5] The common types of preprocessing fall into either one of two categories, Normalization or Augmentation. The proposed pre-processing technique can, in a sense, be regarded as the opposite of augmentation, although they can both successfully be used in conjunction. Augmentation in pre-processing aims to increase complexity of training data, deliberately making the training more challenging and diverse, with the intent to make the trained model more robust to variations.

[6] In light of the aforementioned present available systems and methods, there clearly exists a need for an alternative system. The proposed preprocessing technique, aims to reduce complexity from input images in a consistent manner, with the intent to make the inference job easier, allowing for simpler models to be used without loss of accuracy.

Brief Summary of the Invention

[7] The following presents a simplified summary of the disclosure in order to provide basic understanding to the reader. This summary is not an extensive overview of the disclosure, and it does not identify key/critical elements of the invention or delineate the scope of the invention. Its sole purpose is to present some concepts disclosed herein in a simplified form as a prelude to the more detailed description that is presented later.

[8] A principal objective of the present invention is the ability of a computer vision system to correctly predict the actual situation in an image.

[9] Another objective of the present invention is a form of image pre-processing step, which is applied to images both before training as well as inference of any computer vision model.

[10] Further objective of the present invention is to implement a preprocessing step is to reduce the complexity of the input images to a representation that is easier for the inference step to analyze the specific content of interest.

[11] Still another object of the present invention is the process lies not in the implementation details of the reducing pre-processing itself, but in applying such reduction technique on both annotated data before training the model, as well as during inference on the resulting model.

[12] According to a first aspect of the present invention, a system to enhance accuracy of a computer vision model is disclosed. The system includes a

pre-processing step to reduce image complexity of an input image from a given image source.

[13] In accordance with the first aspect of the present invention, wherein the reducing step includes subtracting or masking out a static background, blocking out or censoring specific areas of little value or with high risk of confusing interference, cropping or seam carving the image to remove sparse or irrelevant areas, enriching the image with apparent or optical movement vector fields, or replacing existing channels, enriching deliberately skewed or weighted normalization and regional or global image corrections such as adjusting contrast, saturation, color, level and other image enhancements.

[14] In accordance with the first aspect of the present invention, wherein the resulting output after applying the reduction step is an image or tensor that is substantially different from the input image.

[15] In accordance with the first aspect of the present invention, wherein the pre-processing step includes in applying image pre-processing step to images both on annotated data before training the model, as well as during inference on the resulting model.

[16] In accordance with the first aspect of the present invention, wherein the training phase includes initiating the input image of any form of image representation that may be used in computer vision applications, annotating the training dataset on the non-reduced images using any manual, automated or semi-automated process, image annotating that does not require annotations (labels) to be of any specific

format and could therefore be of any desired format, data consolidation to replace the annotations' references to input images to their reduced counterparts, training data produced with the proposed technique does not have to be structurally different from any traditional techniques and may be stored and handled using traditional processes, utilizing the pre-trained or untrained base models and conducting the model training using traditional training approaches.

[17] In accordance with the first aspect of the present invention, wherein the trained model with the proposed technique may be handled normally.

[18] In accordance with the first aspect of the present invention, wherein the inference includes inputting the trained model and reduced image to process inference step and performing the inference step using a reduced image.

[19] In accordance with the first aspect of the present invention, wherein reduction during the said pre-processing utilizes to remove, modify, enhance, enrich, emphasize or de-emphasize all or parts of the image data, in order to simplify the training and/or inference task.

[20] In accordance with the first aspect of the present invention, wherein the said system reliably simplifies the input images by removing irrelevant portions, such as the background as well as incorporate new valuable information such as movement vectors or the like.

[21] In accordance with the first aspect of the present invention, wherein the said system uses contextual information like previous images, meta-data from the image

or video sources, or the like, at the said inference or image acquisition step for simplifying the input images.

[22] In accordance with the first aspect of the present invention, wherein the said system greatly simplifies the inference task, as there is less noise to distract it from detecting the objects of interest.

[23] In accordance with the first aspect of the present invention, wherein the said system uses simple model and a comparatively small training dataset, to achieve accuracies that are close to what could normally only be expected from more complex models trained on large datasets.

[24] According to a second aspect of the present invention, a method of enhancing an accuracy of a computer vision model is disclosed method includes steps of pre-processing to reduce image complexity of an input image from a given image source. The pre-processing includes steps of importing, video or image content and/or contextual data as a primary input data to employ the image reduction process, aggregating, said input content and/or contextual data by combining into a unified context representation to determine image information that is of lesser import and additional data to embed in order to aid the model inference, accumulating, intermediate aggregations and/or contextual representations in storage components and reducing, the said aggregated context that is applied to the input image by means of transformation function, in order to reduce it to a representation that makes the training and/or inference easier.

[25] In accordance with the second aspect of the present invention, wherein the resulting output after applying a reduction step is an image or tensor that is substantially different from the input image.

[26] In accordance with the second aspect of the present invention, wherein the method involves in applying image pre-processing step to images both on annotated data before training the model, as well as during inference on the resulting model.

Brief Description of the Drawings

[27] Other objects and advantages of the present invention will become apparent to those skilled in the art upon reading the following detailed description of the preferred embodiments, in conjunction with the accompanying drawings, wherein like reference numerals have been used to designate like elements, and wherein:

[28] Fig. 1a is a flowchart depicting a high-level overview of the process in accordance with an exemplary embodiment of the present invention.

[29] Fig. 1b is a flowchart depicting the detailed process involved in pre-processing step for both training and inference steps.

[30] Fig. 2 shows an example of image reduction using threshold masking of a background context estimated from previous images.

Detailed Description of the Invention

[31] It is to be understood that the present disclosure is not limited in its application to the details of construction and the arrangement of components set forth in the following description or illustrated in the drawings. The present disclosure is capable of other embodiments and of being practiced or of being carried out in various ways. Also, it is to be understood that the phraseology and terminology used herein is for the purpose of description and should not be regarded as limiting.

[32] The use of "including", "comprising" or "having" and variations thereof herein is meant to encompass the items listed thereafter and equivalents thereof as well as additional items. The terms "a" and "an" herein do not denote a limitation of quantity, but rather denote the presence of at least one of the referenced item. Further, the use of terms "first", "second", and "third", and the like, herein do not denote any order, quantity, or importance, but rather are used to distinguish one element from another.

[33] According to a non-limiting exemplary embodiment of the present invention, a system to enhance accuracy of a computer vision model is disclosed. The system includes a pre-processing step to reduce image complexity of an input image from a given image source.

[34] In accordance with non-limiting exemplary embodiment of the present invention, the system includes object detection models from a given image source (e.g., camera, video file, sensor etc) as a pre-processing step for both training and inference steps.

[35] In accordance with the non-limiting exemplary embodiment of the present invention, at its core the process is a form of image pre-processing step that is applied to images both before training as well as inference of any computer vision model.

[36] In accordance with the non-limiting exemplary embodiment of the present invention, the purpose of the preprocessing step is to reduce the complexity of the input images to a representation that is easier for the inference step to analyze the specific content of interest. This can for example involve removing irrelevant clutter (such as cleaning it from a static background and suppressing image noise, etc.) including potentially adding additional information (such as motion vectors from video decoding) that aids the inference.

[37] In accordance with the non-limiting exemplary embodiment of the present invention, the pre-processing step includes importing, video or image content and/or contextual data as a primary input data to employ the image reduction process, aggregating the said input content and/or contextual data by combining into a unified context representation to determine image information that is of lesser import and additional data to embed in order to aid the model inference, accumulating the intermediate aggregations and/or contextual representations in storage components and reducing the said aggregated context that is applied to the input image by means of transformation function, in order to reduce it to a representation that makes the training and/or inference easier.

[38] In accordance with the non-limiting exemplary embodiment of the present invention, the reducing step includes subtracting or masking out a static background, blocking out or censoring specific areas of little value or with high risk of confusing interference, cropping or seam carving the image to remove sparse or irrelevant areas, enriching the image with apparent or optical movement vector fields, or replacing existing channels, enriching deliberately skewed or weighted normalization and regional or global image corrections such as adjusting contrast, saturation, color, level and other image enhancements.

[39] In accordance with the non-limiting exemplary embodiment of the present invention, wherein the pre-processing step includes in applying image pre-processing step to images both on annotated data before training the model, as well as during inference on the resulting model.

[40] In accordance with the non-limiting exemplary embodiment of the present invention, wherein the benefits of this process are ability to use less complex models (i.e. faster inference), reducing the training data size (i.e. less time spent on data acquisition), reducing the training duration (i.e. faster training) and increasing overall accuracy

[41] In accordance with the non-limiting exemplary embodiment of the present invention, the novel part of this process lies not in the implementation details of the reducing pre-processing itself, but in the idea of applying such reduction onto both annotated data before training the model, as well as during inference on the resulting model.

[42] In accordance with the non-limiting exemplary embodiment of the present invention, wherein the proposed pre-processing technique can, in a sense, be regarded as the opposite of augmentation, although they can both successfully be used in conjunction. Augmentation, pre-processing aims to increase complexity of training data, deliberately making the training more challenging and diverse, with the intent to make the trained model more robust to variations.

[43] In accordance with the non-limiting exemplary embodiment, wherein the proposed preprocessing technique, on the other hand, aims to reduce complexity from input images in a consistent manner, with the intent to make the inference job easier, allowing for simpler models to be used without loss of accuracy.

[44] In accordance with the non-limiting exemplary embodiment, the proposed reduction technique can, with a fast and simple model and a comparatively small training dataset, achieve accuracies that are close to what could normally only be expected from more complex models trained on large datasets.

[45] In accordance with the non-limiting exemplary embodiment, the image reduction preprocessing varies depending on the specifics of the approach, but simple yet efficient reduction techniques only require a few tens of milliseconds overhead, making it a tractable approach for boosting accuracy.

[46] In accordance with the non-limiting exemplary embodiment, the proposed reduction technique is not mutually exclusive with normal augmentation techniques, and they can both be used in conjunction. The proposed technique is not limited to object detection models, and can be applied to any computer vision machine learning

task and can be applied to any image processing model, without having to modify the model architecture.

[47] In accordance with the non-limiting exemplary embodiment, there are many different approaches to image reduction which could be chosen depending on use-case and which contextual data is available at image acquisition. Background subtraction is just a simple (but effective) example, suitable for stationary cameras.

[48] Referring to drawings, Fig. 1a is a flowchart (100) depicting a high-level overview of the process in accordance with an exemplary embodiment of the present invention. The process lies not in the implementation details of the reducing pre-processing itself, but in the idea of applying such reduction onto both annotated data before training the model, as well as during inference on the resulting model. The process is a form of image pre-processing step, which is applied to images both before training as well as inference of any computer vision model.

[49] In the training phase, at the first step initiating the input image of any form of image representation that may be used in computer vision applications, at the next step is annotating (2) the training dataset on the non-reduced images using any manual, automated or semi-automated process, at further step the data consolidation (4) helps to replace the annotations' references to input images to their reduced counterparts and in the final step the trained model (10) with the proposed technique may be handled normally.

[50] In the inference phase (12), at the first step initiating the input image (6) of any form of image representation that may be used in computer vision applications,

at the next step is reducing, the said aggregated context that is applied to the input image by means of transformation function, in order to reduce it to a representation that makes the training and/or inference easier.

[51] The resulting output after applying reduction step (8) is an image or tensor that is substantially different from the input image and then consolidated into both annotated data before training the model, as well as used during inference (12) on the resulting model. The inference phase (12) further includes, inputting the trained model and reduced image to process inference step and performing the inference step using a reduced image.

[52] Fig. 1b is a flowchart (100b) depicting the detailed process involved in pre-processing step for both training and inference steps. The image pre-processing step is applied to images both before training as well as inference of any computer vision model. The purpose of the pre-processing step is to reduce the complexity of the input images to a representation that is easier for the inference step to analyse the specific content of interest.

[53] The pre-processing step to reduce image complexity of an input image from a given image source includes: importing, video (1) or image content (6) and/or contextual data (9) as a primary input data to employ the image reduction process. The video source (1) represents any apparatus or software that can generate video content. Examples include, but are not limited to, video files, digital cameras, thermal or multi-spectral imaging equipment, radar, sonar and medical imaging devices etc. The proposed techniques can be applied on images that do not originate from video. Video data is (3), however, a common source of image data and also,

directly or indirectly, provides a rich context as a secondary output, which in turn may be used in the reduction steps for improved results. This includes, but is not limited to, sequential image data as well as movement data. The video data (3) refers to the data that is produced by video sources. It may be compressed or uncompressed, historical recording or a live stream where as any video format applies.

[54] The image source (6) represents any apparatus or software that can generate image data. Examples include, but are not limited to, direct sources of image data such as cameras or imaging devices as well as software or hardware components that decodes video data, or other types of data. The main purpose of the image source is to provide input image data but it may also, directly or indirectly, provide a rich context as a secondary output, which in turn may be used in the reduction steps for improved results. This includes, but is not limited to, timestamps as well as movement data. The input image (7) refers to any form of image representation that may be used in computer vision applications.

[55] The contextual data (9) refers to any meta-data that is not technically part of the image data itself, but carries useful information. This could be meta-data about the images themselves, including, but is not limited to: timestamps, location (e.g. longitude and latitude), environment (e.g. indoors, outdoors, etc) and placement (e.g. pan, tilt, field of view, ect). Such information could come directly from the video source or image source that generates the image content, or it could come from an external source, like other sensors or configurations etc.

[56] Contextual data (9) may also include more complex information that may or may not be of indirect or statistical nature, including but not limited to: Current movement vector field such information is for example available as inter-frame motion compensation data in most modern video compression formats, or can be estimated with optical-flow techniques or similar, Statistical probability distributions such as heatmaps, describing “hot” and “cold” areas of the image where the various objects or events of interest are more or less likely to occur or where there are known areas of high noise, such as foliage. Such information may for example, have been aggregated from previous analysis of the images from the same image source and Background reference image or images such information may, for example, have been created from previous analysis of the images from the same image source.

[57] The contextual data may greatly aid the image reduction process, but is not required, and is thus optional. As exemplified in other sections (Aggregation and Accumulation) it is possible to generate a great deal of contextual information without this additional input.

[58] The next step is aggregation (13), in this step the input images and/or contextual data are combined into a unified context representation, used to determine which image information is of lesser import and what additional data could be of value to embed to aid the model inference. As a simple example, the context representation could just be an image representing the static background, as it would look if no foreground objects were present or visible. Such background context representation could for example be approximated by sampling the input images at regular intervals and selecting the pixel-wise median value across a number of such

samples. There are many different methods for approximating the background image from a sequence of images.

[59] Note that the usage of a background image as context is only for illustrative purposes, as it is a context representation that is easy to explain, generate and visualize. It is also computationally efficient and often provides reasonably good reduction results. The context may also include image variance statistics, motion vectors, camera settings, time information and any other information, including any sensor information or readings from any sources available at image acquisition. It is however important to note that the technique described herein is not limited to only such contexts, nor does the context have to be an image representation.

[60] The context is basically any usable data structure, containing meta-data about the current image, which the later reduction step may utilize to remove, modify, enhance, enrich, emphasise or de-emphasise all or parts of the image data, in order to simplify the training and/or inference task. It is in the aggregation step that this context is created and/or consolidated.

[61] The next step is accumulation (14), the process may, or may not, store intermediate aggregations and/or contextual representations in storage components, such as for example a buffer, heap, tree, queue or other suitable data structure. The storage itself could be on any medium including but not limited to RAM and/or permanent physical storage.

[62] The next step is reduction (8), the aggregated context is applied to the input image by means of some transformation function, in order to reduce it to a

representation that makes the training and/or inference easier. A non-exhaustive list of example transformations that could be performed include, but are not limited to: subtracting or masking out the static background, blocking out or censoring specific areas of little value or with high risk of confusing interference, such as for example foliage (such areas may for example be identified in the aggregation step through simple color and frequency analysis of image sequences, or other techniques), cropping or seam carving the image to remove sparse or irrelevant areas, enriching the image with apparent or optical movement vector fields (may be extracted from many video compression formats, or obtained through various optical flow techniques).

[63] Such additional information could either be encoded as additional channels in the image representation (if the model supports it), or replacing existing channels (for example; an RGB image + XY movement vector field could be transformed to a compatible LXY image tensor, where L represents the lightness, i.e $(min(R,\ G,\ B)\ +\ max(R,\ G,\ B))\div 2)$, deliberately skewed or weighted normalization and regional or global image corrections such as adjusting contrast, saturation, color, level and other image enhancements.

[64] The resulting output after applying reduction step (8) is an image or tensor that is substantially different from the input image and this is because little or no information from the input image that is not relevant to the computer vision task at hand (be it classification, object detection, masking or key points extraction) has been removed or suppressed, and new information of value may have been embedded. The reduction process (8) is applied to images both in the training phase as well as in the inference phase.

[65] The image pre-processing step is applied to images both on annotated data (2) before training the model, as well as during inference on the resulting model. At the training phase the steps include, initiating the input image of any form of image representation that may be used in computer vision applications, annotating (2) the training dataset on the non-reduced images using any manual, automated or semi-automated process, image annotating (5) this does not require annotations (labels) to be of any specific format, and could therefore be of any desired format, data consolidation (15) this helps to replace the annotations' references to input images to their reduced counterparts, training data (16) produced with the proposed technique does not have to be structurally different from any traditional techniques and may be stored and handled using traditional processes, utilizing the pre-trained or untrained base models, conducting the model training (10) using traditional training approaches and the trained model (18) with the proposed technique may be handled normally.

[66] At the inference phase (12), the trained model and reduced image is inputted to process the inference step and performing the inference step using a reduced image (20).

[67] Fig. 2 shows an example (200) of image reduction using threshold masking of a background context estimated from previous images. For illustrative purposes, using the background image context, this could be as simple as subtracting and/or masking out the pixels in the input image that differ, by at least some threshold value, from the background image established in the aggregation step.

[68] At the initial step, the input image before reduction pre-processing refers to any form of image representation that may be used in computer vision applications. In the reduction step, the aggregated context is applied to the input image by means of some transformation function, in order to reduce it to a representation that makes the training and/or inference easier. The resulting output of the above reduction steps is an image or tensor that may be substantially different from the input image.

[69] While specific embodiments of the invention have been shown and described in detail to illustrate the inventive principles, it will be understood that the invention may be embodied otherwise without departing from such principles.

CLAIMS

I/We Claim

1. A system (100) to enhance accuracy of a computer vision model, wherein the system includes,

   a storage device to store a plurality of input images, wherein the plurality of said input images comprises one or more video (1) or image (6) content and/or contextual data (9); and

   a processor for:

      pre-processing, to reduce image complexity of the said input image, wherein the pre-processing includes:

         importing, the said input images as a primary input data to employ an image reduction process;

         aggregating (13), the said input images by combining into a unified context representation to determine image information that is of lesser import and additional data to embed in order to aid a model inference;

         accumulating (14), intermediate aggregations and/or contextual representations in as storage components;

         reducing (8), the said aggregated context that is applied to the input image by means of transformation function, in order to reduce it to a representation that makes the training and/or inference easier, wherein the reducing includes:

            subtracting or masking out a static background;

            blocking out or censoring specific areas of little value or with high risk of confusing interference;

cropping or seam carving the image to remove sparse or irrelevant areas;

enriching the image with apparent or optical movement vector fields, or replacing existing channels;

enriching deliberately skewed or weighted normalization; and

regional or global image corrections such as adjusting contrast, saturation, color, level and other image enhancements;

wherein, resulting output after applying reduction step (8) is an image or tensor that is substantially different from the input image; and

applying, image pre-processing step to images both on annotated data before training the model, as well as during inference on the resulting model;

training, wherein the training includes:

initiating, input image (7) of any form of image representation that may be used in computer vision applications;

annotating (2), training dataset on the non-reduced images using any manual, automated or semi-automated process;

image annotating (5), does not require annotations (labels) to be of any specific format, and could therefore be of any desired format;

data consolidation (15), to replace the annotations' references to input images to their reduced counterparts;

training data (16), produced with the proposed technique does not have to be structurally different from any traditional techniques and may be stored and handled using traditional processes;

utilizing, pre-trained or untrained base models (17);

conducting, model training (10) using traditional training approaches; and

wherein, the trained model (18) with the proposed technique may be handled normally;

inference (12), wherein the inference includes:

inputting, trained model and reduced image to process inference step; and

performing (20), inference step using a reduced image.

2. The system (100) as claimed in claim 1, wherein reduction during the said pre-processing utilizes to remove, modify, enhance, enrich, emphasize or de-emphasize all or parts of the image data, in order to simplify the training and/or inference task.

3. The system (100) as claimed in claim 1, wherein the said system reliably simplifies the input images by removing irrelevant portions, such as the background.

4. The system (100) as claimed in claim 1, wherein the said system uses contextual information like previous images, meta-data from the image or

video sources, or the like, at the said inference or image acquisition step for simplifying the input images.

5. The system (100) as claimed in claim 1, wherein the said system simplifies the inference task, as there is less noise to distract it from detecting the objects of interest.

6. The system (100) as claimed in claim 1, wherein the said system uses simple model and a comparatively small training dataset, to achieve accuracies that are close to what could normally only be expected from more complex models trained on large datasets.

7. A method (100) of enhancing an accuracy of a computer vision model, wherein the method includes steps of,

    pre-processing, to reduce image complexity of an input image (7) from a given image source, wherein the processing includes steps of:

        importing, video (1) or image (6) content and/or contextual data (9) as a primary input data to employ the image reduction process;

        aggregating (13), said input content and/or contextual data by combining into a unified context representation to determine image information that is of lesser import and additional data to embed in order to aid the model inference;

        accumulating (14), intermediate aggregations and/or contextual representations in storage components;

reducing (8), the said aggregated context that is applied to the input image by means of transformation function, in order to reduce it to a representation that makes the training and/or inference easier, wherein the reducing includes steps of:

subtracting or masking out a static background;

blocking out or censoring specific areas of little value or with high risk of confusing interference;

cropping or seam carving the image to remove sparse or irrelevant areas;

enriching the image with apparent or optical movement vector fields, or replacing existing channels;

enriching deliberately skewed or weighted normalization; and

regional or global image corrections such as adjusting contrast, saturation, color, level and other image enhancements;

wherein, resulting output after applying reduction step (8) is an image or tensor that is substantially different from the input image (7); and

applying, image pre-processing step to images both on annotated data before training the model, as well as during inference on the resulting model;

training, wherein the training includes steps of:

initiating, input image (7) of any form of image representation that may be used in computer vision applications;

annotating (2), training dataset on the non-reduced images using any manual, automated or semi-automated process;

image annotating (5), does not require annotations (labels) to be of any specific format, and could therefore be of any desired format;

data consolidation (15), to replace the annotations' references to input images to their reduced counterparts;

training data (16), produced with the proposed technique does not have to be structurally different from any traditional techniques and may be stored and handled using traditional processes;

utilizing, pre-trained or untrained base models (17);

conducting, model training (10) using traditional training approaches; and

wherein, the trained model (18) with the proposed technique may be handled normally;

inference (12), wherein the inference includes steps of:

inputting, trained model (18) and reduced image to process inference step; and

performing (20), inference step using a reduced image.