DATA ANALYSIS METHOD AND SYSTEM

TECHNICAL FIELD

1. This invention relates generally to the data processing field, and more specifically to a new and useful method for processing more input streams in the data processing field.

BACKGROUND

1. It is oftentimes advantageous to dynamically increase the amount of throughput analyzed by analysis models, such as neural networks. This is oftentimes done by increasing the batch size at inference. However, this increased throughput can come at the expense of latency and/or accuracy, which, in real-time use cases, can be extremely detrimental to overall system performance.
2. Thus, there is a need in the data processing field to create a new and useful system and method to increase throughput s with lossless or comparable model performance. This invention provides such new and useful system and method.

BRIEF DESCRIPTION OF THE FIGURES

1. FIGURE 1 is a flowchart representation of the method.
2. FIGURE 2 is an example of the system and method.
3. FIGURE 3 is an illustrative example of the method.
4. FIGURE 4 is a flowchart representation of an illustrative example of the method.

DESCRIPTION OF THE PREFERRED EMBODIMENTS

1. The following description of the preferred embodiments of the invention is not intended to limit the invention to these preferred embodiments, but rather to enable any person skilled in the art to make and use this invention.
2. As shown in FIGURE 1, the method to scale model inputs can include: receiving a measurement set S100; optionally identifying measurements of interest from the measurement set S200; selecting measurements to composite S300; generating composite measurements from the selected measurements S400; and analyzing a batch of measurements including composited and uncomposited measurements S500. The method functions to allow more measurements (e.g., examples) to be processed using the same processing architecture, without retraining, hyperparameter adjustment, and/or hardware upgrades, while maintaining the same or similar performance (e.g., recall, precision, accuracy) and speed. Additionally or alternatively, this method can enable a machine with fixed computational resources (e.g., capable of processing only U computational units per unit time) to concurrently process more than U units of data (e.g., N images), while preserving or maximizing the quality of the output at each timestep.
3. In an example, shown in FIGURE 4, the method includes: receiving a set of N images; filtering out images satisfying a filtering condition (e.g., amount of activity detected in the scene is less than a predetermined threshold); determining the number of remaining images within the set (C); optionally determining a batch size (B) for the analysis model; determining a number of images to select for composition (or proportion of the remaining images to composite) based on the number of remaining images (C) and the batch size of the trained analysis model (B); selecting up to or at least the number of images from the remaining images; downscaling the selected images; generating a set of composite images (multiplexed images), each formed from a grid of downscaled images (e.g., grid of 4 downscaled images); optionally batching the composite images and the remaining (uncomposited) images, wherein the resultant batch size is equal to or less than the batch size of the trained analysis model; and providing the batched images to the analysis model for analysis. In this example, the batch size (B) can be smaller than N, smaller than C, and/or otherwise related to the image set. In specific examples, image filtering and/or image selection for composition can be performed using the metadata associated with the respective image, which can expedite the respective processes.
4. The method can be performed using a system, including: a sensor system and a central processing system, example shown in FIGURE 2. However, the system can include other components. The system functions to monitor and analyze a monitored space, but can be otherwise used. A different system instance (and method instance) is preferably used for each monitored space; alternatively, the same system instance, component instance, and/or method instance can be used for multiple monitored spaces (e.g., the same central processing system can be used to determine safety events for multiple spaces).
5. The sensor system functions to sample measurements of a monitored scene, and can optionally generate metadata for each measurement. The system preferably includes multiple sensor systems (e.g., N sensor systems), but can alternatively include a single sensor system, or any suitable number of sensor systems. The sensor systems can be distributed within a monitored space (e.g., physical space), and monitor the same or different monitored scene. Each sensor system preferably generates a single measurement timeseries (e.g., stream, etc.), but can alternatively generate multiple measurement timeseries.
6. Each sensor system can include one or more: cameras, videocameras, depth sensors, proximity sensors, temperature sensors, light sensors, motion sensors, and/or other sensors. Each sensor system can optionally include local analysis models (e.g., executing on onboard, local hardware), such as optic flow, video change detection, motion detection, object classifiers (e.g., binary classifiers that detect the presence of given object; multiclass classifier that detects presence of one or more objects from an object set; etc.), and/or models for other analyses. The local analyses results can be included in the metadata associated with the measurement, or be otherwise used.
7. The central processing system functions to perform all or a portion of the method. The central processing system can additionally detect security events (e.g., as discussed in US Application No. 16/137,782 filed 21-Sep-2018; US Application No. 16/816,907 filed 12-Mar-2020; US Application No. 16/696,682 filed 26-Nov-2019; US Application No. 16/695,538 filed 26-Nov-2019; each of which are incorporated in its entirety by this reference), generate user interfaces, and/or perform other functions. The central processing system can be an on-premises system (e.g., collocated with the sensor systems, located within the monitored space, etc.); remote from the monitored space; or otherwise arranged relative to the monitored space. The central processing system can include one or more CPUs, GPUs, microprocessors, servers, cloud computing, distributed computing, and/or any other suitable components. The central processing system can be connected to each of the sensor systems by a wireless or wired connection (e.g., a network switch). In a specific example, the central processing system and set of sensor systems cooperatively form a closed-circuit television system (e.g., with or without a television or other user interface).
8. The method can also be used with an analysis model, which functions to extract features or otherwise analyze each measurement. The analysis model is preferably a trained model, but can alternatively be untrained. The analysis model is preferably trained or tuned using a training batch size (e.g., using batch gradient descent, minibatch gradient descent, etc.) to obtain a predetermined set of performance metrics (e.g., latency, accuracy), but can alternatively be trained using any other suitable set of hyperparameters. The analysis model preferably only concurrently accepts an input having an inference batch size (e.g., number of measurements) (B) or less (e.g., to maintain the model accuracy or speed), but can additionally or alternatively have a fixed number of inlet heads, accept more or less inputs, and/or be otherwise configured. The batch size used during inference can be the training batch size, a target batch size, optimal batch size, the batch size that is known to confer the desired performance metrics, and/or other batch size. The batch size used during inference is preferably fixed (e.g., does not vary between iterations or instances of the method), but can alternatively be variable (e.g., based on the number of inputs, performance requirements, etc.). The analysis model can be a neural network, a set of equations (e.g., a bundle adjustment), and/or another model. The analysis model can be: single network, a cascade of networks, a neural network ensemble, and/or otherwise constructed. The analysis model can be: a classifier (e.g., binary, multiclass), regression, and/or other network.
9. A different instance of the method is preferably performed for each timestep or each analysis epoch. Alternatively, each instance can span multiple timesteps or analysis epochs, or multiple instances of the method can be run in series or in parallel for each timestep or analysis epoch. Each method instance preferably processes measurements sampled during the same timestep or analysis epoch, but can alternatively process measurements from other timesteps or analysis epochs. All or portions of the method are preferably performed in real- or near-real time (e.g. within a predetermined time period from measurement sampling), but can be performed asynchronously.
10. Receiving a measurement set S100 functions to obtain information about the monitored scenes for subsequent analysis. The measurement set preferably includes a measurement from each sensor system of the system (e.g., N measurements from N systems), example shown in FIGURE 2, but can alternatively include more or less measurements. The measurement can be: an image, a video (e.g., timeseries of images, image stream), point cloud, proximity measurement, temperature measurement, and/or any other suitable measurement. The measurements within the measurement set preferably all have the same aspect ratio, dimensions, resolution, and/or other measurement characteristics, but can alternatively have different characteristics. In the latter variant, the method can optionally crop, resize, infill, or otherwise process the measurements to consolidate the measurement characteristics.
11. S100 can optionally include determining metadata for each measurement. The metadata is preferably determined by and received from the source sensor system generating the measurement (example shown in FIGURE 3), but can alternatively be determined by a set of metadata extraction modules executed by the central processing system, be retrieved (e.g., based on the source sensor system identifier), and/or be determined by any other suitable system. Examples of metadata that can be associated with each measurement can include: the measurement characteristics (e.g., aspect ratio, resolution, point density, etc.), whether motion was detected, motion amount, timestamp, sensor identifier, whether an object was detected, detected object class, whether an activity was detected, detected activity class, interactions (e.g., between detected objects or entities), change across frames (e.g., sitting down to standing up), associated calendar events, measurement stream history (e.g., short term history, such as within the last 30s, 1min, 10mins, etc.; long term history, such as the last day, month, 3 months, year, etc.), historical patterns (e.g., minimum, maximum, average, or other statistical summary of the number of incidents detected from the measurement stream for a given time of day, day of week, week of year, etc.), scene history (e.g., short term history, long term history, etc.), security event history (e.g., short term history, long term history, etc.), reappearance of agents (e.g., users, vehicles, entities, etc.) on physically and/or temporally adjacent streams, agent parameters (e.g., extracted from the measurement stream, etc.), sensor context (e.g., physical location, monitored object or region class, etc.), scene type or environmental context (e.g., kitchen, front door, back door), ambient environment changes (e.g., lighting change exceeding a threshold, acoustic change exceeding a threshold, etc.), and/or other metadata.
12. Optionally identifying measurements of interest from the measurement set S200 functions to reduce the set of measurements to process in S500. A measurement of interest can be a measurement that has an above-threshold probability of depicting information indicative of an event of interest (e.g., security event); a measurement associated with a predetermined set of metadata values; and/or be otherwise defined. The measurements are preferably identified based on their respective metadata values (example shown in FIGURE 3), but can alternatively be identified based on: the measurement value, the context associated with the source sensor (e.g., measurements of an entryway are more frequently included in the resultant set and measurements of a drain pipe are less frequently included in the resultant set, etc.), time since measurements from the source sensor were run at full resolution (e.g., wherein the probability of analyzing a measurement at full resolution increases with time since a prior measurement from the same stream was analyzed at full resolution), and/or other parameters. S200 is preferably performed by a filtering module (examples shown in FIGURE 2 and FIGURE 3), but can be performed by another processing system. The filtering module can be a binary threshold filter that includes or excludes measurements from a filtered set based on whether the respective metadata or measurement value satisfies a predetermined threshold or condition, or otherwise filter the measurement set. The inclusion or exclusion parameters (filtering parameters, filtering conditions) can be specified: automatically; based on the use case (e.g., motion detection for a security system); based on the sensor system's environmental context (e.g., motion detection for streams from a camera monitoring an interior environment; object detection for streams from a camera monitoring an external environment); manually; or otherwise determined. In a first example, an activity filter is used to filter out measurements with less than a threshold amount of motion in the monitored scene. In a second example, the filtering module can retain measurements that have changed between timesteps and/or filter out measurements that have not changed between timesteps (e.g., by comparing hashes of images output by the same sensor). However, S200 can be otherwise performed.
13. Selecting measurements to composite S300 functions to pick a subset of the (resultant) measurement set for composition (e.g., multiplexing). The measurements can be selected from the resultant measurement set from S200 (filtered set, example shown in FIGURE 2 and FIGURE 3), the measurement set received from the sensor systems, and/or from any other suitable set of measurements. The measurements can be selected based on the respective metadata values (example shown in FIGURE 3), based on no information (e.g., randomly selected), based on the respective measurement value, and/or selected based on any other suitable data.
14. S300 can include: determining a number of measurements to select, and selecting at least the number of measurements. However, the measurements can be otherwise selected.
15. Determining a number of measurements to select functions to determine the minimum number of measurements required to satisfy the analysis model’s batch size.
16. In a first variation, the number of measurements to select is calculated based on the number of measurements of interest (C) and the batch size (B):
17. C = g\*M\*B + (1-M)\*B, where:
18. M is the proportion of batch inputs that should be composited measurements; g is the number of constituent measurements within a composite measurement; and g\*M\*B is the number of measurements to select for composition. M can be calculated (e.g., based on fixed B, fixed g, and C determined from S200), fixed, or otherwise determined. In an illustrative example, for a set of 80 measurements of interest, a target batch size of 32, and a grid value of 4 (4 measurements cooperatively form each composited measurement), M can be 50%, such that 64 of the measurements are composited to form 16 composite measurements. The 16 composite measurements, together with the 16 uncomposited measurements, cooperatively form a batch of 32 measurements for analysis model ingestion.
19. In a second variation, the number of measurements to select is predetermined and/or fixed.
20. In a third variation, the number of measurements to select is based on the computing hardware performance (e.g., GPU performance). For example, the number of measurements to select can be looked up based on the current %GPU, current memory used, current energy used, and/or other parameters. In another example, the number of measurements to select can increase when the hardware performance is good, and decrease when the hardware performance is low (e.g., wherein the number of measurements can be calculated, iteratively determined, or otherwise determined).
21. In a fourth variation, the number of measurements to select is determined based on the number of sensor systems, sensor streams, input channels, and/or measurement data.
22. In a fifth variation, the number of measurements to select is determined based on performance of the policy (discussed below) and/or analysis model. In this variation, the number of measurements to select can be decreased when the performance drops below a threshold metric, increased when the performance increases above a threshold metric, and/or otherwise adjusted based on the performance.
23. However, all images can be selected for composition, no images can be selected for composition, or the number of measurements to select can be otherwise determined.
24. Selecting at least the number of measurements functions to pick the individual measurements to be composited. The system can select up to the minimum number of measurements required to satisfy the analysis model’s batch size, more than the minimum number (e.g., a multiple of the number of grids in a composite measurement), a predetermined number of measurements, and/or any other suitable number of measurements. Measurements can be evaluated and selected individually (e.g., until the minimum number of measurements is reached); evaluated in a batch, then selected based on the evaluation; and/or evaluated and selected in any order. In variants of the method including S200, the number of selected measurements can dynamically vary across analysis epochs and/or timesteps based on the number of measurements of interest identified in S200 (e.g., vary as a function of the number of measurements of interest).
25. The measurements are preferably selected using a policy or policy module enforcing the policy, but can be otherwise selected.
26. In a first variation, the measurements can be selected randomly, using a truly random model, a pseudorandom model, a quasirandom model, a low discrepancy sequence, and/or other random model.
27. In a second variation, the measurements can be selected based on the probability of a security event. In this variation, a security event probability can be determined based on the measurement and/or associated metadata, wherein measurements with low security event probabilities can be preferentially selected for composition or otherwise handled.
28. In a third variation, the measurements can be selected based on the scene complexity (e.g., predetermined or determined from the measurement). In this variation, measurements of less complex scenes can be preferentially selected for composition or otherwise handled.
29. In a fourth variation, the measurements can be selected based on the associated metadata value. For example, measurement with lower activity or motion values can be preferentially selected for composition. In another example, measurements from different sensor systems can be scheduled for compression during predetermined analysis epochs.
30. In a fifth variation, the measurements can be selected using a learned model, which determines whether each measurement should be composited. The learned model can classify each measurement: serially, as a batch, or in any order. In a first embodiment, the learned model classifies each measurement in the set as a measurement to composite or not composite. In a second embodiment, the learned model scores each measurement, wherein measurements satisfying a score condition can be selected for composition. In this embodiment, the learned model can score the measurement based on the scene content, the metadata values, and/or other values. For example, a measurement associated with a high-motion scene can be scored with a low composition score (e.g., the measurement should be analyzed full size or full resolution if possible), while a measurement associated with a low-motion scene can be scored with a high composition score. The measurements can then be ranked, included in a composition set based on whether the respective score satisfies a score threshold, or otherwise selected for composition based on the score. However, the learned model can otherwise facilitate measurement selection for composition.
31. The learned model can be trained on a predetermined training set, wherein each training measurement is labelled with a “compose” or “not compose” label; trained on the runtime measurements, the learned model composition label, and the analysis model results (e.g., retrained when the analysis model performance drop below a threshold; penalize the runtime measurement-composition label combination when the analysis model performance drops below a threshold; etc.); trained by comparing the analysis results of the composited measurement and the analysis results of the uncomposited measurement; and/or otherwise trained.
32. In a sixth variation, the measurements can be selected using a set of heuristics. The heuristics can be applied to the measurement value, the metadata values, and/or any other suitable measurement data. The heuristics can be learned, specified by a user, and/or otherwise determined. For example, the heuristics can specify that measurements having metadata values below a threshold (e.g., activity values below a threshold) should be candidates for composition, while measurements with metadata values above a threshold (e.g., activity values above a threshold) should not be candidates for composition. In another example, the heuristics can specify that measurements from a sensor system should not be composited. In a first specific example, the heuristics can specify that measurements from a sensor system should not be composited for a predetermined number of epochs after a measurement from said sensor system was composited in a prior epoch. In another specific example, the heuristics can specify which measurements cannot be composited based on when the last measurement from the same sensor system was analyzed at full resolution (e.g., N minutes, measurement epochs, frames, etc.). In another specific example, the heuristics can specify deterministically rotating the composition assignment through the set of measurement streams (e.g., such that all streams receive the same number of composited/uncomposited opportunities over a predetermined period of time). However, any other suitable set of heuristics can be used.
33. In a seventh variation, the measurements can be selected using a score (e.g., affinity score) and/or using an affinity function (e.g., lossless affinity function). The score can be calculated based on: the time since the measurement stream’s last uncomposited analysis; the number of active entities detected in the measurement or measurement stream (e.g., people, vehicles, other mobile entities, etc.); the number of high-importance entities (e.g., firearms, weaponry, etc.); the presence or probability of a small object (e.g., with a critical dimension of less than 3ft, 2ft, 1ft, 6inches, etc.); whether the measurement stream is active (e.g., whether there is change detected in the measurement stream; whether the measurement stream’s sensor is turned on, etc.); importance of the measurement stream and/or entity detected therein to a downstream process (e.g., downstream detection model); and/or other parameters of or extracted from the measurement stream.
34. However, the measurements can be selected using a combination of the above, or be otherwise performed.
35. Generating composite measurements from the selected measurements S400 functions to composite a subset of the measurements into composite measurements, while leaving the remaining measurements uncomposited. While this reduces the resolution (and possibly detection accuracy) for the composited images, this enables the same, pretrained analysis model to analyze more measurements than the original batch size. In examples, the inventors have discovered that this can be accomplished with minimal (e.g., less than 10%, 5%, 1%, 0.5%, etc.) drop in performance (e.g., accuracy, precision, recall, etc.).
36. A composite measurement can be a synthetic measurement that is created from a set of constituent measurements. The composite measurement can have the same aspect ratio, dimensions, resolution, and/or other measurement parameters as the uncomposited measurements and/or training measurements, but can alternatively have different parameter values. For example, when the uncomposited measurements are 1024 x 768 px, the composited measurement is also 1024 x 768 px. However, the composite measurement can be otherwise related to the uncomposited measurements, or otherwise constructed.
37. The composite measurement is preferably a grid of downscaled constituent measurements (examples shown in FIGURE 2 and FIGURE 3), but can be otherwise constructed. The measurement identifier is tracked for each grid cell or grid position, such that the analysis result for the grid cell can be traced back to the source sensor system and/or monitored scene, but can be otherwise identified or tracked. The constituent measurements are preferably uniformly downscaled to fit the grid cell size (e.g., maintaining the constituent measurements’ relative aspect ratio), but can be non-uniformly downscaled or otherwise downscaled. The constituent measurements can be downscaled by resizing the measurement, using a nearest-neighbor interpolation, using a bilinear algorithm, using box sampling, using resampling (e.g., sinc, Lanczos), using vectorization, using Fourier-transform methods, and/or otherwise downscaling the constituent measurement. Alternatively, the constituent measurements can be overlaid to form the composite measurement, have key scene features extracted and composited into a single image, or otherwise composited into the composite measurement.
38. The constituent measurements cooperatively forming a composite measurement are preferably measurements selected in S300 (examples shown in FIGURE 2 and FIGURE 3), but can additionally or alternatively be measurements received in S100, measurements identified in S200, filler measurements (e.g., with NAN or a predetermined value for each pixel, etc.), and/or other measurements. Filler measurements can be used to fill a gap in the composite measurement when the number of selected measurements is not a multiple of the composite measurement grid number, or otherwise used. The constituent measurements can optionally include padding between spatially adjacent measurements (e.g., adjacent measurements in the collation).
39. Each measurement selected in S300 preferably only appears once in the set of generated composite measurements, but can alternatively appear multiple times in the generated composite measurement set. Each composite measurement preferably includes multiple constituent measurements, but can alternatively include a single constituent measurement.
40. The number of constituent measurements per composite measurement (e.g., g) is preferably predetermined, but can alternatively vary across analysis epochs, based on the number of measurements of interest (C) (e.g., when the number of measurements of interest exceeds the number of potential measurements that could be analyzed, even if all measurements in the batch were composited, or when C > g\*B, etc.), or otherwise vary. The number of constituent measurements per composite measurement can be calculated as a fraction of: the uncomposited measurement height, width, and/or otherwise calculated. For example, 4, 16, or 32 constituent measurements can be included in a composited measurement.
41. Analyzing a batch of measurements S500 functions to extract features or determine high-level analyses from the measurements. The batch of measurements preferably includes composited measurements (e.g., from S400) and uncomposited measurements (e.g., the non-selected measurements from S300; raw measurements from S100; processed measurements from S100; etc.), examples shown in FIGURE 2 and FIGURE 3, but can alternatively include only uncomposited measurements, only composited measurements, reference measurements, and/or other measurements. S500 can concurrently or contemporaneously analyze the measurements in the batch, analyze the measurements within the batch in series, and/or analyze the measurements in the batch in any order. S500 is preferably performed using the analysis model discussed above, but can alternatively be performed using any model. The analyses can be: security event analyses, change detection analyses, anomaly detection analyses, and/or any other suitable analysis. In examples, S500 can be performed using one or more of the methods disclosed in: US Application No. 16/137,782 filed 21-Sep-2018; US Application No. 16/816,907 filed 12-Mar-2020; US Application No. 16/696,682 filed 26-Nov-2019; and/or US Application No. 16/695,538 filed 26-Nov-2019; each of which are incorporated in its entirety by this reference. However, S500 can be otherwise performed.
42. Different processes and/or elements discussed above can be performed and controlled by the same or different entities. In the latter variants, different subsystems can communicate via: APIs (e.g., using API requests and responses, API keys, etc.), requests (e.g., controlled by authentication and/or authorization credentials), and/or other communication channels.
43. Alternative embodiments implement the above methods and/or processing modules in non-transitory computer-readable media, storing computer-readable instructions that, when executed by a processing system, cause the processing system to perform the method(s) discussed herein. The instructions can be executed by computer-executable components integrated with the computer-readable medium and/or processing system. The computer-readable medium may include any suitable computer readable media such as RAMs, ROMs, flash memory, EEPROMs, optical devices (CD or DVD), hard drives, floppy drives, non-transitory computer readable media, or any suitable device. The computer-executable component can include a computing system and/or processing system (e.g., including one or more collocated or distributed, remote or local processors) connected to the non-transitory computer-readable medium, such as CPUs, GPUs, TPUS, microprocessors, or ASICs, but the instructions can alternatively or additionally be executed by any suitable dedicated hardware device.
44. Embodiments of the system and/or method can include every combination and permutation of the various system components and the various method processes, wherein one or more instances of the method and/or processes described herein can be performed asynchronously (e.g., sequentially), concurrently (e.g., in parallel), or in any other suitable order by and/or using one or more instances of the systems, elements, and/or entities described herein.
45. As a person skilled in the art will recognize from the previous detailed description and from the figures and claims, modifications and changes can be made to the preferred embodiments of the invention without departing from the scope of this invention defined in the following claims.

CLAIMS

We Claim:

1. The inventions as shown and/or described.