Google AI Blog

The latest from Google Research

4D-Net: Learning Multi-Modal Alignment for 3D and Image Inputs in Time

Wednesday, February 23, 2022

Posted by AJ Piergiovanni and Anelia Angelova, Research Scientists, Google Research

While not immediately obvious, all of us experience the world in four dimensions (4D). For example, when walking or driving down the street we observe a stream of visual inputs, snapshots of the 3D world, which, when taken together in time, creates a 4D visual input. Today's autonomous vehicles and robots are able to capture much of this information through various onboard sensing mechanisms, such as LiDAR and cameras.

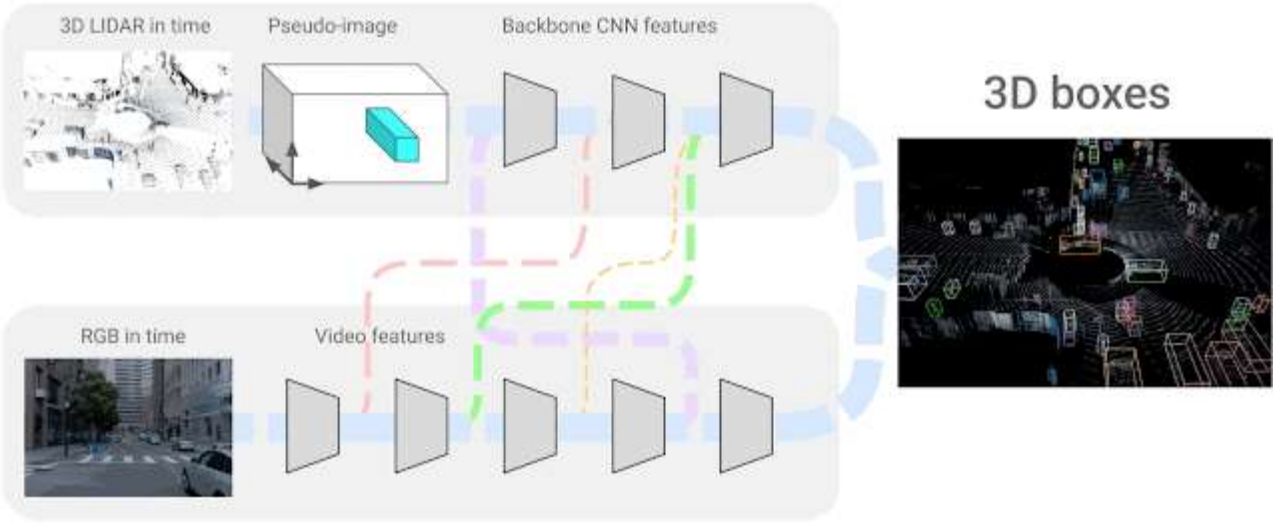
LiDAR is a ubiquitous sensor that uses light pulses to reliably measure the 3D coordinates of objects in a scene, however, it is also sparse and has a limited range — the farther one is from a sensor, the fewer points will be returned. This means that far-away objects might only get a handful of points, or none at all, and might not be seen by LiDAR alone. At the same time, images from the onboard camera, which is a dense input, are incredibly useful for semantic understanding, such as detecting and segmenting objects. With high resolution, cameras can be very effective at detecting objects far away, but are less accurate in measuring the distance.

Autonomous vehicles collect data from both LiDAR and onboard camera sensors. Each sensor measurement is recorded at regular time intervals, providing an accurate representation of the 4D world. However, very few research algorithms use both of these in combination, especially when taken "in time", i.e., as a temporally ordered sequence of data, mostly due to two major challenges. When using both sensing modalities simultaneously, 1) it is difficult to maintain computational efficiency, and 2) pairing the information from one sensor to another adds further complexity since there is not always a direct correspondence between LiDAR points and onboard camera RGB image inputs.

In "4D-Net for Learned Multi-Modal Alignment", published at ICCV 2021, we present a neural network that can process 4D data, which we call *4D-Net*. This is the first attempt to effectively combine both types of sensors, 3D LiDAR point clouds and onboard camera RGB images, when both are in time. We also introduce a *dynamic connection learning* method, which incorporates 4D information from a scene by performing connection learning across both feature representations. Finally, we demonstrate that 4D-Net is better able to use motion cues and dense image information to detect distant objects while maintaining computational efficiency.

4D-Net

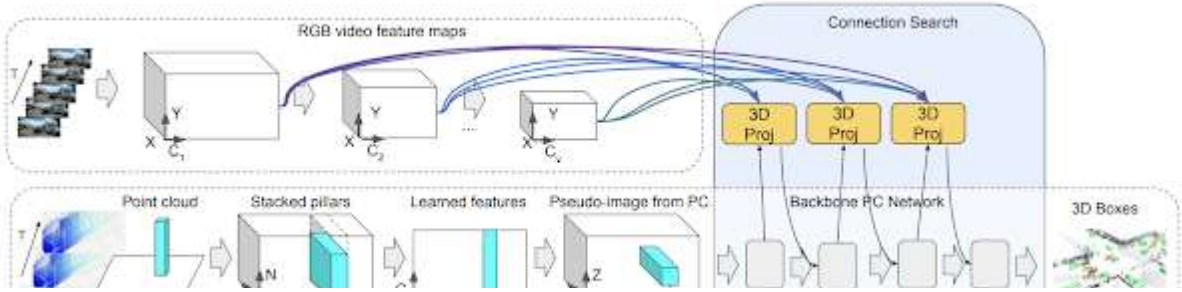
In our scenario, we use 4D inputs (3D point clouds and onboard camera image data in time) to solve a very popular visual understanding task, the 3D box detection of objects. We study the question of how one can combine the two sensing modalities, which come from different domains and have features that do not necessarily match — i.e., sparse LiDAR inputs span the 3D space and dense camera images only produce 2D projections of a scene. The exact correspondence between their respective features is unknown, so we seek to learn the connections between these two sensor inputs and their feature representations. We consider neural network representations where each of the feature layers can be combined with other potential layers from other sensor inputs, as shown below.



4D-Net effectively combines 3D LiDAR point clouds in time with RGB images, also streamed in time as video, learning the connections between different sensors and their feature representations.

Dynamic Connection Learning Across Sensing Modalities

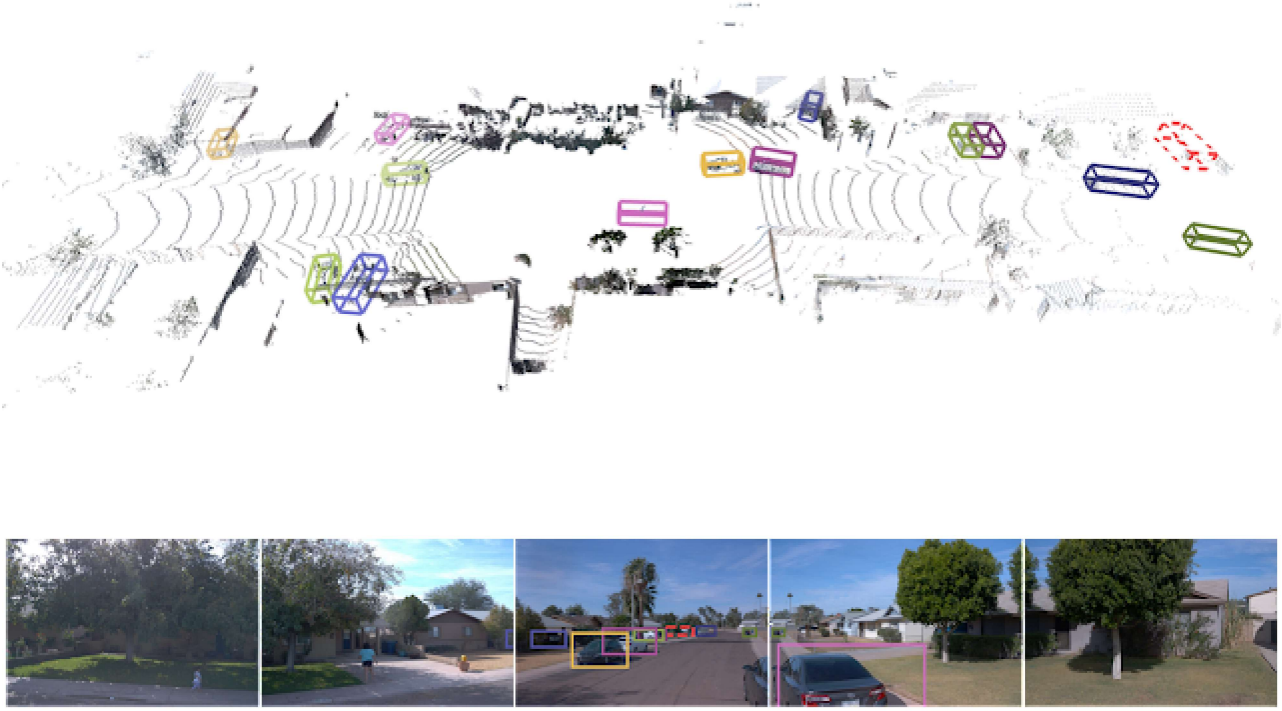
We use a light-weight neural architecture search to learn the connections between both types of sensor inputs and their feature representations, to obtain the most accurate 3D box detection. In the autonomous driving domain it is especially important to reliably detect objects at highly variable distances, with modern LiDAR sensors reaching several hundreds of meters in range. This implies that more distant objects will appear smaller in the images and the most valuable features for detecting them will be in earlier layers of the network, which better capture fine-scale features, as opposed to close-by objects represented by later layers. Based on this observation, we modify the connections to be dynamic and select among features from all layers using self-attention mechanisms. We apply a learnable linear layer, which is able to apply attention-weighting to all other layer weights and learn the best combination for the task at hand.



Connection learning approach schematic, where connections between features from the 3D point cloud inputs are combined with the features from the RGB camera video inputs. Each connection learns the weighting for the corresponding inputs.

Results

We evaluate our results against state-of-the-art approaches on the Waymo Open Dataset benchmark, for which previous models have only leveraged 3D point clouds in time or a combination of a single point cloud and camera image data. 4D-Net uses both sensor inputs efficiently, processing 32 point clouds in time and 16 RGB frames within 164 milliseconds, and performs well compared to other methods. In comparison, the next best approach is less efficient and accurate because its neural net computation takes 300 milliseconds, and uses fewer sensor inputs than 4D-Net.



Results on a 3D scene. **Top**: 3D boxes, corresponding to detected vehicles, are shown in different colors; dotted line boxes are for objects that were missed. **Bottom**: The boxes are shown in the corresponding camera images for visualization purposes.

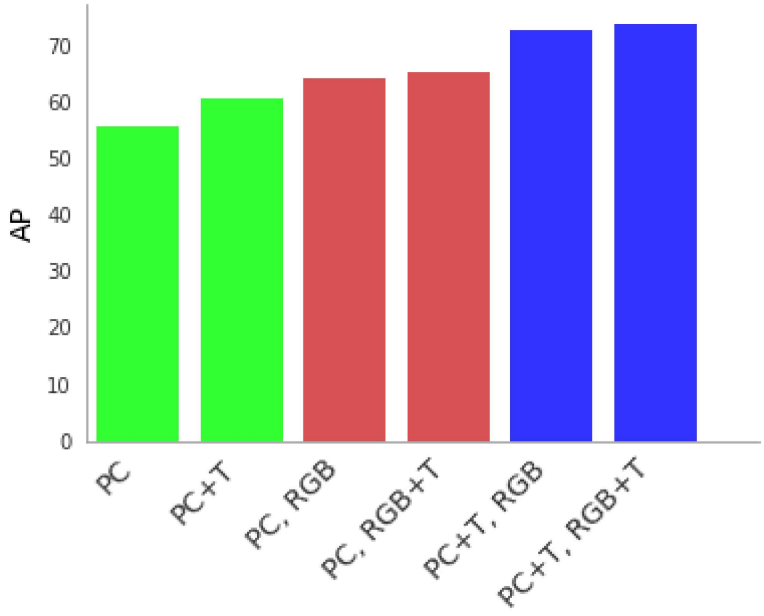
Detecting Far-Away Objects

Another benefit of 4D-Net is that it takes advantage of both the high resolution provided by RGB, which can accurately detect objects on the image plane, and the accurate depth that the point cloud data provides. As a result, objects at a greater distance that were previously missed by point cloud-only approaches can be detected by a 4D-Net. This is due to the fusion of camera data, which is able to detect distant objects, and efficiently propagate this information to the 3D part of the network to produce accurate detections.

Is Data in Time Valuable?

To understand the value of the 4D-Net, we perform a series of ablation studies. We find that substantial improvements in detection accuracy are obtained if at least one of the sensor inputs is streamed in time. Considering both sensor inputs in time provides the largest improvements

in performance.

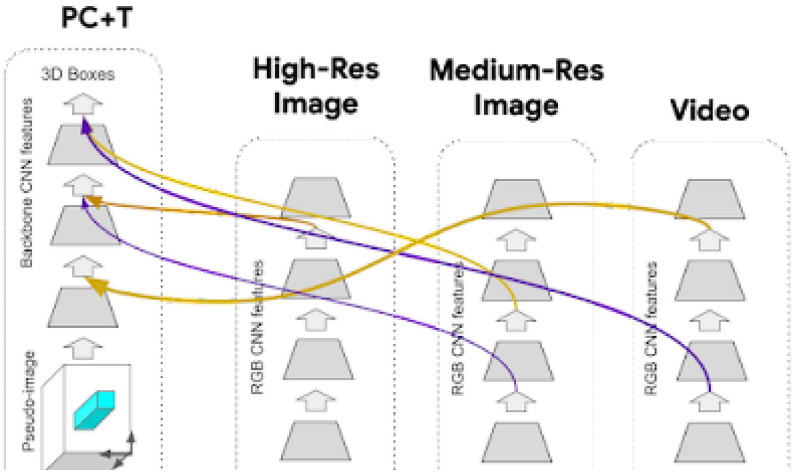


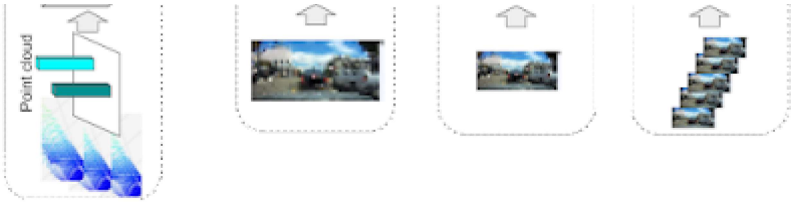
4D-Net performance for 3D object detection measured in average precision (AP) when using point clouds (PC), Point Clouds in Time (PC + T), RGB image inputs (RGB) and RGB images in Time (RGB + T). Combining both sensor inputs in time is best (rightmost columns in blue) compared to the left-most columns (green) which use a PC without RGB inputs. All joint methods use our 4D-Net multi-modal learning.

Multi-stream 4D-Net

Since the 4D-Net dynamic connection learning mechanism is general, we are not limited to only combining a point cloud stream with an RGB video stream. In fact, we find that it is very cost-effective to provide a large resolution single-image stream, and a low-resolution video stream in conjunction with 3D point cloud stream inputs. Below, we demonstrate examples of a four-stream architecture, which performs better than the two-stream one with point clouds in time and images in time.

Dynamic connection learning selects specific feature inputs to connect together. With multiple input streams, 4D-Net has to learn connections between multiple target feature representations, which is straightforward as the algorithm does not change and simply selects specific features from the union of inputs. This is an incredibly light-weight process that uses a differentiable architecture search, which can discover new wiring within the model architecture itself and thus effectively find new 4D-Net models.





Example multi-stream 4D-Net which consists of a stream of 3D point clouds in time (PC+T), and multiple image streams: a high-resolution single image stream, a medium-resolution single image stream and a video stream (of even lower resolution) images.

Summary

While deep learning has made tremendous advances in real-life applications, the research community is just beginning to explore learning from multiple sensing modalities. We present 4D-Net which learns how to combine 3D point clouds in time and RGB camera images in time, for the popular application of 3D object detection in autonomous driving. We demonstrate that 4D-Net is an effective approach for detecting objects, especially at distant ranges. We hope this work will provide researchers with a valuable resource for future 4D data research.

Acknowledgements

This work is done by AJ Piergiovanni, Vincent Casser, Michael Ryoo and Anelia Angelova. We thank our collaborators, Vincent Vanhoucke, Dragomir Anguelov and our colleagues at Waymo and Robotics at Google for their support and discussions. We also thank Tom Small for the graphics animation.