Supporting TVM in Autoware

Original proposal to use TVM in Autoware

https://discourse.ros.org/t/unified-ml-inference-in-autoware-a-proposal/14058/14

State of ML in Autoware

Node	Model	File Format	Inference Engine
lidar_apollo_cnn_seg_detect 4	Unkown 13	caffe	caffe
lidar_point_pillars 24	PointsPillars 5	onnx	TensorRT
trafficlight_recognizer/region_tlr_mxnet 4	Unkown	MxNet	MxNet
trafficlight_recognizer/region_tlr_ssd 7	SSD-unkown	caffee	ssdcaffe
vision_darknet_detect 10	Yolov2/3	darknet	darknet
vision_segment_enet_detect 3	ENet	caffe	caffe for ENet
vision_ssd_detect 3	SSD	caffe	ssdcaffe

Problem with current approach

- Use a varied range of formats and frameworks.
  - For deployment this presents challenges
    - Varied degree of hardware acceleration support
    - Lock into one acceleration hardware, difficult to port 7 frameworks to a different ML accelerator hardware.
- Use of frameworks that are forks or no longer actively maintained.
  - This presents challenges in the long term support and future update of these nodes
- Use of frameworks that are proprietary and require special licenses and sign-ups to use.
- Documentation and trained weights lacking
  - Makes it difficult to even compile the nodes. For example 13.
  - Difficult to re-train the model on custom data-sets. For example 5.

Solution Proposal

Unify all ML workload deployment in Autoware with a single workflow and a single ML inference framework. Organize and document all pre-trained models used in Autoware in a Model Zoo.

Current TVM related GitHub Issues/PRs

Issues:

- https://github.com/autowarefoundation/autoware.universe/issues/908
- https://github.com/autowarefoundation/autoware.universe/issues/628

PRs:

- https://github.com/autowarefoundation/autoware.universe/pull/1181

Universe packages that depends on TensorRT

- Lidar centerpoint
- pointpainting_fusion
- Lidar Apollo Instance Sgementation
- TensorRT YOLO
- Traffic Light Classifier
- Traffic Light SSD Fine Detector

Current Concern

- Are we really going to be able to port all the modules to TVM?
  - ARM would like to have Autoware to use TVM to be a default ML inference library
- Do we have enough engineers to work/review the tasks?
  - Are there any tutorials?
    - @liuzf1988 could create a tutorial for porting once lidar_centerpoint is done
    - Maybe we can share the tutorial to CoE to increase the number of engineers who can use TVM
  - candidates?
- Who will be maintaining ModelZoo?
  - Ambroise is comfortable as a maintainer, but would like someone from AWF to do the review
  - Xinyu is interested in learning TVM so he can start looking into it.

Other Comments:

- Fatih: For the current PRs, we can just merge them as long as it doesn't break building process. We can fix as further issue arises.
- Should we force everyone to use TVM?
  - Fatih: We can allow different inference engine, and compare the performance. We can give a choice for the users
  - We can have both TVM implementation and TensorRT implementation existing in the repository unless one is confirmed that it outperforms the other in all aspect
  - -> We will be making TVM as "recommended" inference engine for Autoware