Discorso poster

The problem we are facing is the point cloud segmentation in which we have to assign a label to each point of the cloud. In recent years, with the advent of autonomous and assisted driving, the problem of recognizing and classifying objects captured by cameras and sensors has become very important.
Unfortunately, the difficulties in this field are many.
Fully supervised methods have achieved really good results but in the real world there are many more categories of objects to be classified and manual 3D annotations for these objects is extremely time-consuming and expensive.
For these reason our method exploits Zero-shot learning which is able to recognize unseen objects by utilizing additional informations like word embedding to transfer the knowledge of seen categories to unseen ones.
This type of learning achieves good performance in 2D, but has some limitations in 3D. The main proposed methods in 3D are divided in two categories: generative and projection-based methods.
Generative methods trains a generator to generate fake features using seen classes as training set and a classifier to discriminate real and synthetic features. This kind of methods do not perform well because 3D feature are more difficult to generate and every time we see a new object we have to retrain (then infeasible for real-world applications).
Projection-based methods aim to align visual features to corresponding semantic features, by the seen-class supervision. In this way, the unseen classes can be recognized using similarities between visual and semantic features. This method doesn't retrain for each new unseen object but has limited performance. In fact, visual features can match only a subset of the semantic features.

Current autonomous vehicles and robots are equipped with multiple sensors, like LiDARs and cameras. Considering that images contain rich appearance features and point clouds contain accurate location and geometry features, we aim to make these two uni-modal visual data complement each other and generate more comprehensive visual features to better align with semantic ones.
Our model focuses on transductive and generalized zero-shot learning, so the training set is composed by both seen and unseen classes by tuples containing frame of point clouds with corresponding image and word embedding and for seen classes also the corresponding ground truth label.
Our model is composed by four modules.
Firstly, there is visual-semantic feature extraction module that extracts features from the various type of input data. To extract 3D features from lidar point clouds it is used Cylinder3D, for 2D features of images it is used ResUnet and to obtain word embeddings it is used W2V and Glove followed by a MLP.
Secondly, there is the semantic-visual feature enhancement module that reduces the gap between semantic and visual domain by the cross-attention mechanism which learns the semantic-visual projection automatically and enhance each feature with valuable knowledge of the other.
Then, there is the semantic guided visual feature fusion module that combines the enhanced visual features under the guide of the semantic ones. In particular, the two weight matrices are calculated with a multi-head attention mechanism and they represent the importance of the two visual features for the corresponding semantic feature. Then is applied a element-wise multiplication between the weight matrix and corresponding enhanced visual feature and passed to a MLP with softmax function.
Finally, there is the semantic-visual alignment module that aligns visual and semantic feature spaces under seen classes supervision, in order to transfer knowledge from seen to unseen classes. Two loss functions are used, the first one for seen classes aims to have compact

distribution within classes and distinguishable one between classes, the second one for unseen classes aims to push unseen classes' features apart to seen ones.
At the end, during the inference the model uses the fused visual feature and the semantic feature in order to segment the scene. The class of each point is determined by the similarity between its fused visual feature and semantic features of all classes (seen and unseen).

We used two different datasets: SemanticKITTY which contains 22 sequences with 20 classes and nuScenes which contains 40k annotated samples with 6 images of 17 classes.

The results written in the table speak for themselves. Our method outperforms previous state of the art methods with improvement rates of 52% and 49% in average for unseen mIoU. Also from the qualitative results is impressive. For example it is able to detect the traffic-sign whereas TGP does not. Furthermore it does not increase too much the inference time yielding real-time performance.

Finally we studied the importance of each module for our model.
For example, replacing semantic guided visual feature fusion module with a simple concatenation and a MLP we discover a drop of 3.5% of the unseen mIoU.