EXPLAINING 3D OBJECT DETECTION THROUGH SHAPLEY VALUE-BASED ATTRIBUTION MAP

Michihiro Kuroki and Toshihiko Yamasaki

Dept. Information & Communication Engineering, The University of Tokyo, Tokyo

Appendix

Transformation of Equation

In this section, we describe the details of the transformation from Eq. 5 to Eq. 6 in the main paper. The expected value in Eq. 5 can be represented as follows:

$$\phi_i(f, \mathcal{X}) \approx \frac{1}{d} \sum_{k=1}^{d} \mathbb{E}\big[f(\mathcal{X}_{\mathbf{s}^k}) - f(\mathcal{X}_{\mathbf{s}^{k\prime}}) \bigm| \mathbf{s}_i^k = 1, \mathbf{s}_i^{k\prime} = 0\big], \tag{5}$$

$$= \frac{1}{d} \sum_{k=1}^{d} G_{\mathcal{X},k,i}. \tag{5a}$$

$$G_{\mathcal{X},k,i} = \mathbb{E}\big[f(\mathcal{X}_{\mathbf{s}^k}) - f(\mathcal{X}_{\mathbf{s}^{k\prime}}) \bigm| \mathbf{s}_i^k = 1, \mathbf{s}_i^{k\prime} = 0\big]. \tag{5b}$$

The expected value of Eq. 5b can be expressed as the summation of all combinations of two mask patterns. We denote two binary masks as $\mathbf{s}^{k,a}$ and $\mathbf{s}^{k,b}$, which exhibit patterns similar to that of $\mathbf{s}^k$. If duplication is permitted, we need to consider two conditions among the masks, namely $\mathbf{s}_i^{k,a} = 1, \mathbf{s}_i^{k,b} = 0$ and $\mathbf{s}_i^{k,a} = 0, \mathbf{s}_i^{k,b} = 1$.

$$\begin{aligned} G_{\mathcal{X},k,i} = \sum_{\mathbf{m}^{k,a}} \sum_{\mathbf{m}^{k,b}} \Big\{ &\Big(f(\mathcal{X}_{\mathbf{m}^{k,a}}) - f(\mathcal{X}_{\mathbf{m}^{k,b}})\Big) P\Big[\mathbf{s}^{k,a} = \mathbf{m}^{k,a}, \mathbf{s}^{k,b} = \mathbf{m}^{k,b} \Bigm| \mathbf{s}_i^{k,a} = 1, \mathbf{s}_i^{k,b} = 0\Big] \\ &+ \Big(f(\mathcal{X}_{\mathbf{m}^{k,a}}) - f(\mathcal{X}_{\mathbf{m}^{k,b}})\Big) P\Big[\mathbf{s}^{k,a} = \mathbf{m}^{k,a}, \mathbf{s}^{k,b} = \mathbf{m}^{k,b} \Bigm| \mathbf{s}_i^{k,a} = 0, , \mathbf{s}_i^{k,b} = 1\Big] \Big\}. \end{aligned} \tag{5c}$$

Here, P denotes probability. This equation can be further transformed as follows:

$$\begin{aligned} G_{\mathcal{X},k,i} = \sum_{\mathbf{m}^{k,a}} \sum_{\mathbf{m}^{k,b}} \Bigg\{ &\frac{\big(f(\mathcal{X}_{\mathbf{m}^{k,a}}) - f(\mathcal{X}_{\mathbf{m}^{k,b}})\big) P\big[\mathbf{s}^{k,a} = \mathbf{m}^{k,a}, \mathbf{s}^{k,b} = \mathbf{m}^{k,b}, \mathbf{s}_i^{k,a} = 1, \mathbf{s}_i^{k,b} = 0\big]}{P[\mathbf{s}_i^{k,a} = 1, \mathbf{s}_i^{k,b} = 0]} \\ &+ \frac{\big(f(\mathcal{X}_{\mathbf{m}^{k,a}}) - f(\mathcal{X}_{\mathbf{m}^{k,b}})\big) P\big[\mathbf{s}^{k,a} = \mathbf{m}^{k,a}, \mathbf{s}^{k,b} = \mathbf{m}^{k,b}, \mathbf{s}_i^{k,a} = 0, \mathbf{s}_i^{k,b} = 1\big]}{P[\mathbf{s}_i^{k,a} = 0, \mathbf{s}_i^{k,b} = 1]} \Bigg\}. \end{aligned} \tag{5d}$$

$$\begin{aligned} G_{\mathcal{X},k,i} = \frac{1}{P[\mathbf{s}_i^k = 1] \cdot P[\mathbf{s}_i^k = 0]} \sum_{\mathbf{m}^{k,a}} \sum_{\mathbf{m}^{k,b}} \Big\{ &\big(f(\mathcal{X}_{\mathbf{m}^{k,a}}) - f(\mathcal{X}_{\mathbf{m}^{k,b}})\big) P\big[\mathbf{s}^{k,a} = \mathbf{m}^{k,a}, \mathbf{s}^{k,b} = \mathbf{m}^{k,b}, \mathbf{s}_i^{k,a} = 1, \mathbf{s}_i^{k,b} = 0\big] \\ &+ \big(f(\mathcal{X}_{\mathbf{m}^{k,a}}) - f(\mathcal{X}_{\mathbf{m}^{k,b}})\big) P\big[\mathbf{s}^{k,a} = \mathbf{m}^{k,a}, \mathbf{s}^{k,b} = \mathbf{m}^{k,b}, \mathbf{s}_i^{k,a} = 1, \mathbf{s}_i^{k,b} = 0\big] \Big\}, \end{aligned} \tag{5e}$$

$$= \frac{1}{P[\mathbf{s}_i^k = 1] \cdot P[\mathbf{s}_i^k = 0]} \sum_{\mathbf{m}^{k,a}} \sum_{\mathbf{m}^{k,b}} \Big\{ \Big(f(\mathcal{X}_{\mathbf{m}^{k,a}}) - f(\mathcal{X}_{\mathbf{m}^{k,b}}) \Big) \Big(\mathbf{m}_i^{k,a} - \mathbf{m}_i^{k,b} \Big) P\Big[\mathbf{s}^{k,a} = \mathbf{m}^{k,a}, \mathbf{s}^{k,b} = \mathbf{m}^{k,b} \Big] \Big\}. \tag{5f}$$

We now aim to reformulate the summation over $\mathbf{m}^{k,b}$ in terms of its expected value.

$$\sum_{\mathbf{m}^{k,a}} \sum_{\mathbf{m}^{k,b}} \Big\{ \Big(f(\mathcal{X}_{\mathbf{m}^{k,a}}) - f(\mathcal{X}_{\mathbf{m}^{k,b}}) \Big) \Big(\mathbf{m}_i^{k,a} - \mathbf{m}_i^{k,b} \Big) P\Big[\mathbf{s}^{k,a} = \mathbf{m}^{k,a}, \mathbf{s}^{k,b} = \mathbf{m}^{k,b} \Big] \Big\},$$

$$= \sum_{\mathbf{m}^{k,a}} \Big\{ f(\mathcal{X}_{\mathbf{m}^{k,a}}) \cdot \mathbf{m}_i^{k,a} - f(\mathcal{X}_{\mathbf{m}^{k,a}}) \cdot \mathbb{E}\big[\mathbf{s}_i^{k,b}\big] - \mathbb{E}\big[f(\mathcal{X}_{\mathbf{s}^{k,b}})\big] \cdot \mathbf{m}_i^{k,a} + \mathbb{E}\big[f(\mathcal{X}_{\mathbf{s}^{k,b}}) \cdot \mathbf{s}_i^{k,b}\big] \Big\} P\Big[\mathbf{s}^{k,a} = \mathbf{m}^{k,a} \Big], \tag{5g}$$

$$\approx \sum_{\mathbf{m}^{k,a}} \Big\{ f(\mathcal{X}_{\mathbf{m}^{k,a}}) \cdot \mathbf{m}_i^{k,a} - f(\mathcal{X}_{\mathbf{m}^{k,a}}) \cdot \mathbb{E}\big[\mathbf{s}_i^{k,b}\big] - \mathbb{E}\big[f(\mathcal{X}_{\mathbf{s}^{k,b}})\big] \cdot \mathbf{m}_i^{k,a} + \mathbb{E}\big[f(\mathcal{X}_{\mathbf{s}^{k,b}})\big] \cdot \mathbb{E}\big[\mathbf{s}_i^{k,b}\big] \Big\} P\Big[\mathbf{s}^{k,a} = \mathbf{m}^{k,a} \Big], \tag{5h}$$

$$= \sum_{\mathbf{m}^{k,a}} \Big\{ \Big(f(\mathcal{X}_{\mathbf{m}^{k,a}}) - \mathbb{E}\big[f(\mathcal{X}_{\mathbf{s}^{k,b}})\big] \Big) \Big(\mathbf{m}_i^{k,a} - \mathbb{E}\big[\mathbf{s}_i^{k,b}\big] \Big) \Big\} P\Big[\mathbf{s}^{k,a} = \mathbf{m}^{k,a} \Big]. \tag{5i}$$

In the transformation, we assumed independence between $f(\mathcal{X}_{\mathbf{s}^{k,b}})$ and $\mathbf{s}_i^{k,b}$. Given that $\mathbf{m}^{k,a}$ and $\mathbf{m}^{k,b}$ follow the same distribution of $\mathbf{s}^k$, we can rewrite Eq. 5i as follows.

$$G_{\mathcal{X},k,i} \approx \frac{1}{P[\mathbf{s}_i^k = 1] \cdot P[\mathbf{s}_i^k = 0]} \sum_{\mathbf{m}^k} \Big\{ \Big(f(\mathcal{X}_{\mathbf{m}^k}) - \mathbb{E}\big[f(\mathcal{X}_{\mathbf{s}^k})\big] \Big) \Big(\mathbf{m}_i^k - \mathbb{E}\big[\mathbf{s}_i^k\big] \Big) \Big\} P\Big[\mathbf{s}^k = \mathbf{m}^k \Big], \tag{5j}$$

$$= \frac{1}{\mathbb{E}\big[\mathbf{s}_i^k\big]\big(1 - \mathbb{E}\big[\mathbf{s}_i^k\big]\big)} \sum_{\mathbf{m}^k} \Big\{ \Big(f(\mathcal{X}_{\mathbf{m}^k}) - \mathbb{E}\big[f(\mathcal{X}_{\mathbf{s}^k})\big] \Big) \Big(\mathbf{m}_i^k - \mathbb{E}\big[\mathbf{s}_i^k\big] \Big) \Big\} P\Big[\mathbf{s}^k = \mathbf{m}^k \Big]. \tag{5k}$$

Using the definition of covariance, we ultimately rewrite the summation as the expected values over $\mathbf{s}^k$.

$$G_{\mathcal{X},k,i} \approx \frac{\mathbb{E}\big[f(\mathcal{X}_{\mathbf{s}^k}) \cdot \mathbf{s}_i^k\big] - \mathbb{E}\big[f(\mathcal{X}_{\mathbf{s}^k})\big] \cdot \mathbb{E}\big[\mathbf{s}_i^k\big]}{\mathbb{E}\big[\mathbf{s}_i^k\big]\big(1 - \mathbb{E}\big[\mathbf{s}_i^k\big]\big)}. \tag{5l}$$

$$\phi_i(f, \mathcal{X}) \approx \frac{1}{d} \sum_{k=1}^{d} \frac{\mathbb{E}\big[f(\mathcal{X}_{\mathbf{s}^k}) \cdot \mathbf{s}_i^k\big] - \mathbb{E}\big[f(\mathcal{X}_{\mathbf{s}^k})\big] \cdot \mathbb{E}\big[\mathbf{s}_i^k\big]}{\mathbb{E}\big[\mathbf{s}_i^k\big] \cdot (1 - \mathbb{E}\big[\mathbf{s}_i^k\big])} \tag{6}$$

Pseudocode

The pseudocode describing our method is shown in Algorithm 1.

Algorithm 1 Pseudocode for computing attribution map Φ

Inputs: The number of samplings N, number of approximation layers L, object detector function F, input point cloud $\mathcal{X}$, explanation target detection $\mathcal{D}_t$, detection score function $Sim(\cdot)$, and all-ones mask $\mathbf{1}$.

Outputs: Attribution map Φ

$\Phi \leftarrow O$
for $l = 1, \ldots, L$ **do**
  $\Phi^l \leftarrow O$
  $\text{sum\_score} \leftarrow 0, \text{sum\_mask} \leftarrow O, \text{sum\_score\_mask} \leftarrow O$
  for $r = 1, \ldots, N$ **do**
    $\mathbf{s}^{l_r} \leftarrow$ The input point cloud space is divided into voxel units. The voxels are selected randomly with probability $p = \frac{l}{L+1}$, and a point i within the unselected voxels is masked (i.e. $\mathbf{s}_i^{l_r} = 0$).
    $f\left(\mathcal{X}_{\mathbf{s}^{l_r}}\right) \leftarrow \max_{\mathcal{D}_j \in F(\mathcal{X}_{\mathbf{s}^{l_r}})} Sim(\mathcal{D}_t, \mathcal{D}_j)$
    $\text{sum\_score} \leftarrow \text{sum\_score} + f\left(\mathcal{X}_{\mathbf{s}^{l_r}}\right)$
    $\text{sum\_mask} \leftarrow \text{sum\_mask} + \mathbf{s}^{l_r}$
    $\text{sum\_score\_mask} \leftarrow \text{sum\_score\_mask} + f\left(\mathcal{X}_{\mathbf{s}^{l_r}}\right) \cdot \mathbf{s}^{l_r}$
  end for
  $\overline{f\left(\mathcal{X}_{\mathbf{s}^l}\right)} \leftarrow \text{sum\_score}/N$
  $\overline{\mathbf{s}^l} \leftarrow \text{sum\_mask}/N$
  $\overline{f\left(\mathcal{X}_{\mathbf{s}^l}\right) \cdot \mathbf{s}^l} \leftarrow \text{sum\_score\_mask}/N$
  $\Phi_l \leftarrow \frac{1}{L} \cdot \left\{\overline{f\left(\mathcal{X}_{\mathbf{s}^l}\right) \cdot \mathbf{s}^l} - \overline{f\left(\mathcal{X}_{\mathbf{s}^l}\right)} \cdot \overline{\mathbf{s}^l}\right\} \oslash \left\{\overline{\mathbf{s}^l} \odot \left(\mathbf{1} - \overline{\mathbf{s}^l}\right)\right\}$
  $\Phi \leftarrow \Phi + \Phi_l$
end for
return Φ