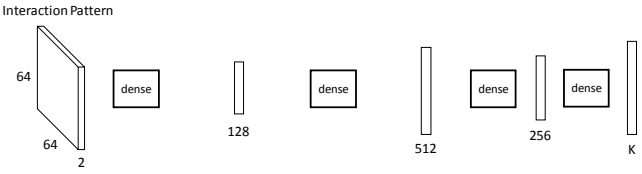
Supplementary Material: Learning to Detect Human-Object Interactions

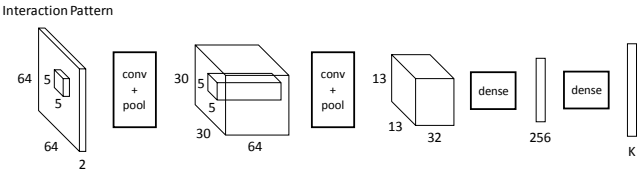
Yu-Wei Chao[1], Yunfan Liu[1], Xieyang Liu[1], Huayi Zeng[2], and Jia Deng[1]

[1]University of Michigan, Ann Arbor
{ywchao,yunfan,lxieyang,jiadeng}@umich.edu

[2]Washington University in St. Louis*
{zengh}@wustl.edu



(a) Fully-connected network (fc).



(b) Convolutional network (conv).

Figure 1: Two different architectures for the pairwise stream.

1. Pairwise Stream

In our HO-RCNN, we consider two different network architectures for the pairwise stream: a fully-connected network (fc) and a convolutional network (conv). The two architectures are illustrated in Fig. 1. Both architectures take as input an Interaction Pattern and produce as output a vector of classification scores on K HOI classes of interest. For fair comparison, both networks have approximately the same number of parameters, and are trained with identical schemes.

*Work done at the University of Michigan as a visiting student.