Visuospatial relationship detection network

We trained a recurrent neural network GRU (Gate Recurrent Unit) to process the word vector generated by word2vec model to capture the semantic correlation between entity classes.

The network structure is as follows.

Detailed structure of VRD-REI

Index	Input	Operation	Details
(0)	Depth data	3D relative pos encode	1) Linear(3, 256), ReLU()
(1)	2D spatial data	2D spatial distribution encode	2) (conv1_p): Conv2d(2, 32, kernel_size=(5, 5), stride=(2, 2), padding=(2, 2)), BatchNorm2d (32), ReLU() 3) (conv2_p): Conv2d(32, 64,kernel_size=(3, 3), stride=(1,1),padding=(1,1)),BatchNorm2d(64), ReLU(), Maxpool2d(2) 4) Hourglass(8, 64), BatchNorm2d(64), ReLU(), Maxpool2d(2) 5) (conv3_p): Conv2d(64, 256, kernel_size=(4, 4), stride=(1, 1)), BatchNorm2d(64), ReLU()
(2)	Image with bbox	Visual fea. encode	1) ResNet18 2) Linear(512, 256), ReLU()
(3)	Subject /object label	Subject /object class encode	1) GRU(300, 150, batch_first=True, bidirectional= True), weights=((450, 300), (450, 150), (450,), (450,), (450, 300), (450, 150), (450,), (450,)) 2) (sub_encode): Linear(300, 512), ReLU() 3) (obj_encode): Lin VRD-REI ear(300, 512), ReLU()
(4)	(0), (1), (2)	Features combination	1) Linear(768, 512), BatchNorm1d(512)
(5)	(3), (4)	Relations prediction	1) Linear(512, 12), Softmax(), dropout=0.5

Detailed structure of DR-NET

Index	Input	Operation	Details
(0)	2D spatial data	2D spatial distribution encode	1) Linear(8, 256), ReLU()
(1)	Image with bbox	Visual fea. encode	1) ResNet18 2) Linear(512, 256), ReLU()
(2)	Subject /object label	Subject /object class encode	1) GRU(300, 150, batch_first=True, bidirectional= True), weights=((450, 300), (450, 150), (450,), (450,), (450, 300), (450, 150), (450,), (450,))

			2) (sub_encode): Linear(300, 512), ReLU() 3) (obj_encode):Linear(300, 512), ReLU()
(3)	(0), (1)	Features combination	1) Linear(768, 512), BatchNorm1d(512)
(4)	(2), (3)	Relations prediction	1) Linear(512, 9), Softmax(), dropout=0.5

FIG. 1 shows the loss and accuracy change curves of each model in Table II in the test process. It can be seen that with the continuous training of the sample data, the loss value decreases while the accuracy value increases, and the accuracy of the VRD-REI model increases. After the 50th iteration, the oscillation tends to converge.