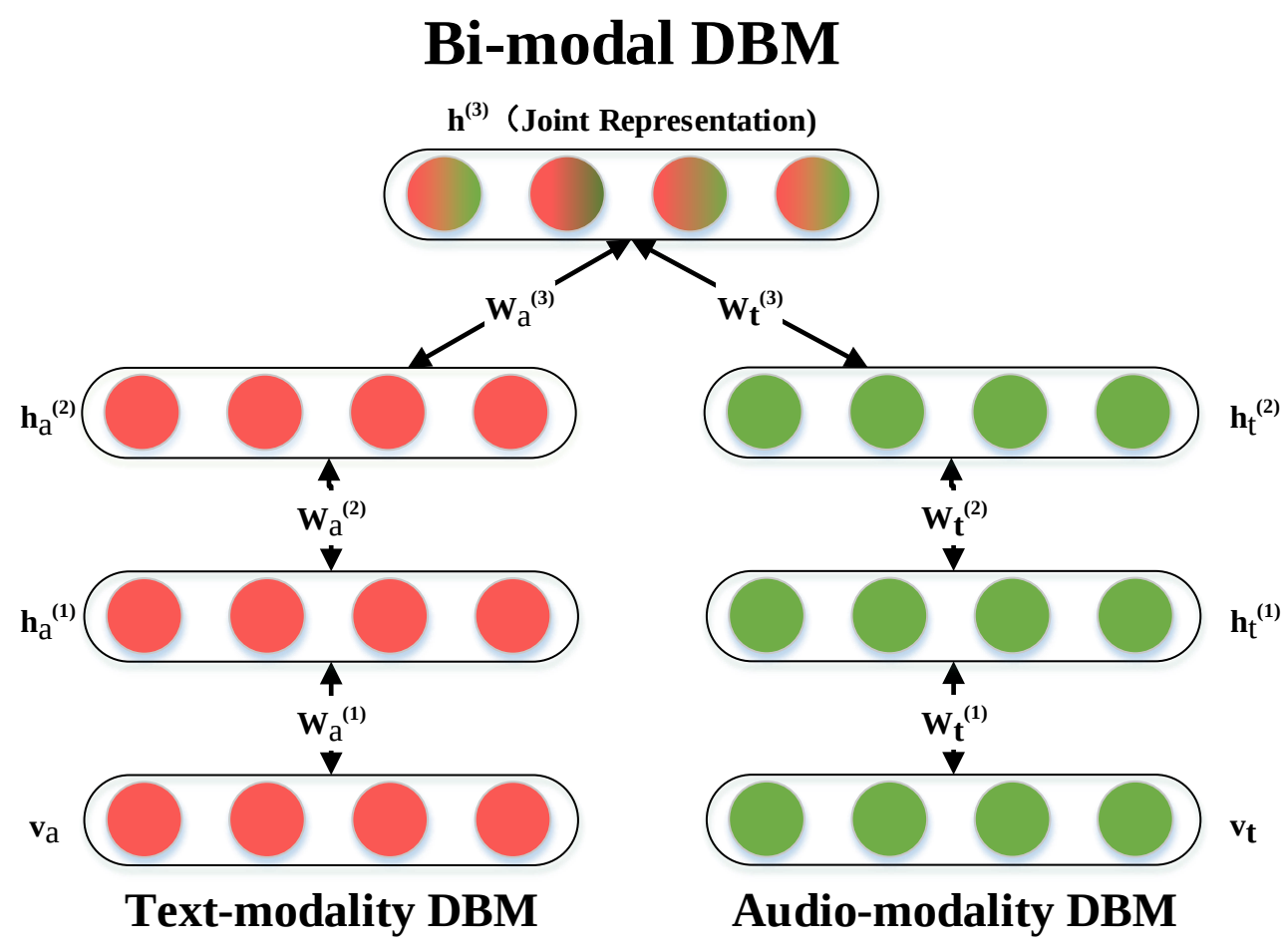
Bi-modal DBM
h(3)（Joint Representation)
Wa(3)
Wt(3)
ha(2)
ht(2)
Wa(2)
Wt(2)
ha(1)
ht(1)
Wa(1)
Wt(1)
va
vt
Text-modality DBM
Audio-modality DBM