ROI Pooling

- ***ROI***

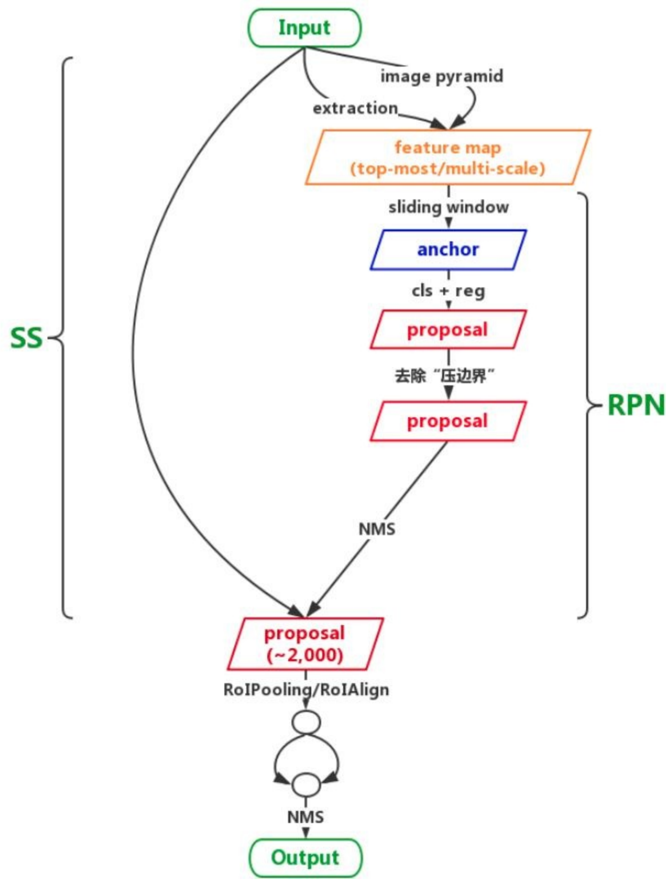
```
å…¨ç§°Region of Interestï¼Œç‰¹å¾å›¾ä¸Šçš„æ¡†;
1) åœ¨Fast RCNNä¸ï¼ŒROIæ˜¯æŒ‡Selective Searchå®ŒæˆåŽå¾—åˆ°çš„â€œå€™é€‰æ¡†â€åœ¨ç‰¹å¾å›¾ä¸Šçš„æ˜ å°„
2) åœ¨Faster RCNNä¸ï¼Œå€™é€‰æ¡†æ˜¯ç»è¿‡RPNäº§ç”Ÿçš„ï¼Œç„¶åŽå†æŠŠå„ä¸ªâ€œå€™é€‰æ¡†â€æ˜ å°„åˆ°ç‰¹å¾å›¾ä¸Šï¼Œå¾—åˆ°ROIS
```

- ***ROI Poolingçš„è¾“å…¥***
  è¾“å…¥ç”±ä¸¤éƒ¨åˆ†ç»„æˆï¼š
  1) ç‰¹å¾å›¾ï¼šåœ¨Fast RCNNä¸ï¼Œå®ƒä½äºŽROI Poolingä¹‹å‰ï¼Œåœ¨Faster RCNNä¸ï¼Œå®ƒæ˜¯ä¸ŽRPNå…±äº«é‚£ä¸ªç‰¹å¾å›¾ï¼Œé€šå¸¸æˆ‘ä»¬ç§°ä¹‹ä¸ºâ€œshare_convâ€;
  2) rois:åœ¨Fast RCNNä¸ï¼ŒæŒ‡çš„æ˜¯Selective Searchçš„è¾“å‡º;åœ¨Faster RCNNä¸æŒ‡çš„æ˜¯RPNçš„è¾“å‡ºï¼Œä¸€å †çŸ©å½¢å€™é€‰æ¡†ï¼Œå½¢çŠ¶ä¸º1Ã—5Ã—1Ã—1(4ä¸ªåæ ‡+ç´¢å¼•index)ï¼Œå…¶ä¸å€¼å¾—æ³¨æ„çš„æ˜¯ï¼šåæ ‡çš„å‚è€ƒç³»ä¸æ˜¯é’ˆå¯¹feature mapè¿™å¼ å›¾çš„ï¼Œè€Œæ˜¯é’ˆå¯¹åŽŸå›¾çš„ã€‚
- ***ROI Poolingçš„è¾“å‡º***
  è¾“å‡ºæ˜¯batchä¸ªvectorï¼Œå…¶ä¸batchçš„å€¼ç‰äºŽROIçš„ä¸ªæ•°ï¼Œvectorçš„å¤§å°ä¸ºchannelÃ—wÃ—h;ROI Poolingçš„è¿‡ç¨‹å°±æ˜¯ä¸€ä¸ªä¸ªå¤§å°ä¸åŒçš„boxçŸ©å½¢æ¡†ï¼Œéƒ½æ˜ å°„æˆå¤§å°å›ºå®š(wÃ—h)çš„çŸ©å½¢æ¡†;

RPN

Region Proposal Network
æœ¬è´¨æ˜¯åŸºäºŽæ»‘çª—çš„æ— ç±»åˆ«objectæ£€æµ‹å™¨

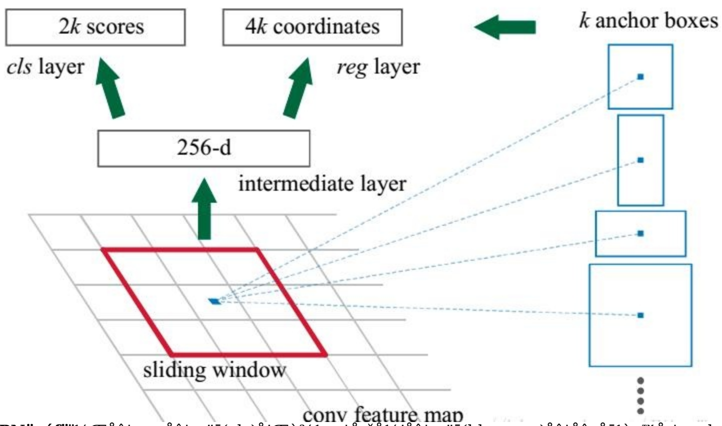
- **RPNæ‰€åœ¨ä½ç½®**



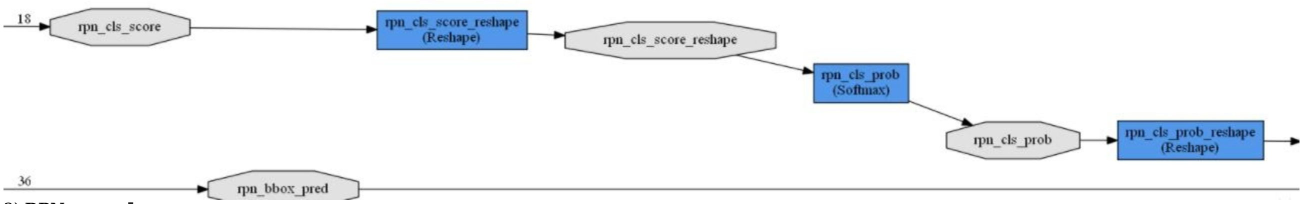
- **tip:**
  1) åªæœ‰åœ¨trainæ—¶ï¼Œcls+regæ‰èƒ½å¾—åˆ°å¼ºç›‘ç£ä¿¡æ¯(æ¥æºäºŽground truth)ã€‚å³ground truthä¼šå‘Šè¯‰cls+regç»“æžœï¼Œå“ªäº›æ˜¯çœŸçš„å‰æ™¯ï¼Œä»Žè€Œå¼•å¯¼cls+regç»“æžœå¦‚å¾—æ£ç¡®åŒºåˆ†å‰åŽæ™¯çš„èƒ½åŠ›ï¼›åœ¨referenceé˜¶æ®µï¼Œå°±è¦é cls+regè‡ªåŠ›æ›´ç”Ÿäº†ã€‚
  2) åœ¨trainé˜¶æ®µï¼Œä¼šè¾“å‡ºçº¦2000ä¸ªproposalï¼Œä½†åªä¼šæŠ½å–å…¶ä¸256ä¸ªproposalæ¥è®ç»ƒRPNçš„cls+regç»“æžœï¼›åˆ°äº†referenceé˜¶æ®µï¼Œåˆ™ç›´æŽ¥è¾“å‡ºæœ€é«˜scoreçš„300ä¸ªproposalã€‚æ¤æ—¶ç”±äºŽæ²¡æœ‰äº†ç›‘ç£ä¿¡æ¯ï¼Œæ‰€æœ‰RPN**å¹¶ä¸çŸ¥é“è¿™äº›proposalæ˜¯å‰æ™¯**ï¼Œæ•´ä¸ªè¿‡ç¨‹åªæ˜¯æƒ¯æ€§åœ°æŽ¨é€ä¸€æ³¢æ— tagçš„proposalç»™åŽé¢çš„Fast R-CNNã€‚
  3) RPNçš„è¿ç”¨ä½¿å¾—region proposalçš„é¢å¤–å¼€é”€å°±åªæœ‰ä¸€ä¸ªä¸¤å±‚ç½‘ç»œã€‚

- **RPNçš„ç»“æž„**

1) **RPNå¤´éƒ¨**ï¼Œé€šè¿‡ä»¥ä¸‹ç»“æž„ç”Ÿæˆanchor(å…¶å®žå°±æ˜¯ä¸€å †æœ‰ç¼–å·æœ‰åæ ‡çš„bbox)



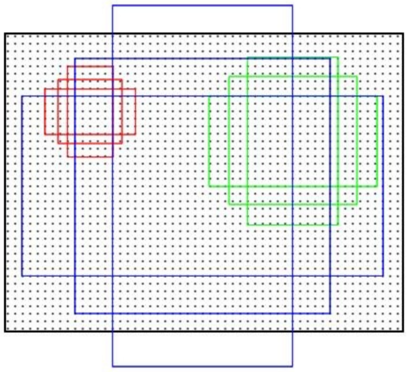
2) **RPNä¸éƒ¨**ï¼Œåˆ†ç±»åˆ†æ”¯(cls)å'Œè¾¹æ¡†å›žå½'åˆ†æ”¯(bbox reg)åˆ†åˆ«å¯¹è¿™å †anchorè¿›è¡Œè®¡ç®—



3) **RPNæœ«ç«¯**
é€šè¿‡å¯¹ä¸¤ä¸ªåˆ†æ”¯çš„ç»“æžœè¿›è¡Œæ±‡æ€»ï¼Œæ¥å®žçŽ°å¯¹anchorçš„ åˆæ¥ç›é™¤ï¼ˆå…ˆå‰”é™¤è¶Šç•Œçš„anchorï¼Œå†æ ¹æ®clsç»“æžœé€šè¿‡NMSç®—æ³•åŽ»é‡ï¼‰å'Œ åˆæ¥å®šä½ï¼ˆæ ¹æ®bbox regç»“æžœï¼‰ï¼Œæ¤æ—¶è¾“å‡ºçš„éƒ½æ”¹å¤´æ¢é¢å« Proposal äº†
add:RPNä¹‹åŽï¼Œproposalæˆä¸ºROIï¼Œè¢«è¾“å…¥ROIPoolingæˆ–ROIAlignä¸è¿›è¡Œsizeä¸Šçš„å½'ä¸€åŒ–ï¼Œ**ä¸å±žäºŽRPNçš„èŒƒå›´**
tip:ä½†æ˜¯å¦‚æžœåªåœ¨æœ€åŽä¸€å±‚feature mapä¸Šæ˜ å°„å›žåŽŸå›¾åƒçš„è¯ï¼Œä¸"å®šä¼šé€ æˆå°å›¾ç‰‡å¯¹åº"æ˜ å°„å›žåŽŸå›¾åƒçš„anchorè¢«é™å®šäº†å°ºå¯¸ä¸‹é™ï¼Œé£ä¹ˆä½ŽäºŽæœ€å°anchorå°ºå¯¸çš„å°ç›®æ ‡è™½ç„¶è¢«anchoråœˆå…¥ï¼Œåœ¨åŽé¢çš„è¿‡ç¨‹ä¸ä¾ç„¶å®¹æ˜"è¢«æ¼æŽ‰ã€‚**FPN**çš„å‡ºçŽ°å¤§å¤§é™ä½ŽäºŽå°ç›®æ ‡çš„æ¼æ£€çŽ‡ï¼Œä½¿å¾—RPNå¦‚è™Žæ·»ç¿¼ã€‚

- anchoræœºåˆ¶
rpnç½'ç»œçš„æ ¸å¿ƒï¼Œanchorç»™å‡ºä¸€ä¸ªåŸºå‡†çª—å¤§å°ï¼ŒæŒ‰ç…§å€æ•°å'Œé•¿å®½å¾—åˆ°ä¸åŒå¤§å°çš„çª—ï¼Œè®ºæ–‡ä¸åŸºå‡†çª—çš„å¤§å°ä¸º16,ç»™äº†(8,16,32)ä¸‰ç§å€æ•°å'Œ(0.5,1,2)ä¸‰ç§æ¯"ä¾‹ï¼Œè¿™æ ·èƒ½å¤Ÿå¾—åˆ°ä¸€å…±9ç§å°ºåº¦çš„anchorï¼Œåœ¨å¯¹60Ã—40çš„mapè¿›è¡Œæ»'çª—æ—¶(æ¥é•¿ä¸º1)ï¼Œä»¥ä¸å¿ƒåƒç´ ä¸ºåŸºç‚¹æž„é€ 9ç§anchoræ˜ å°„åˆ°åŽŸæ¥çš„1000Ã—600åŽŸå›¾ä¸ï¼Œæ˜ å°„æ¯"ä¾‹ä¸º16å€ã€‚é£ä¹ˆæ€»å…±å¯ä»¥å¾—åˆ°60Ã—40Ã—9å¤§çº¦2ä¸‡ä¸ªanchorã€‚

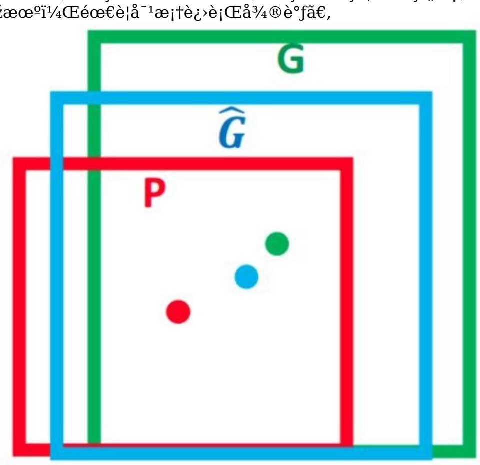
- è®ç»ƒ
RPNç½'ç»œè®ç»ƒï¼Œé£ä¹ˆå°±æ¶‰åŠground truthå'Œloss functionçš„é—®é¢˜ã€‚å¯¹äºŽå·¦æ"¯è·¯ï¼Œground truthä¸ºanchoræ˜¯å¦ä¸ºç›®æ ‡ï¼Œç"¨0/1è¡¨ç¤ºã€‚é£ä¹ˆæ€Žä¹ˆåˆ¤å®šä¸€ä¸ªanchorå†…æ˜¯å¦æœ‰ç›®æ ‡å'¢ï¼Ÿè®ºæ–‡ä¸é‡‡ç"¨äº†è¿™æ ·çš„è§„åˆ™ï¼š1ï¼‰å‡å¦‚æŸanchorä¸Žä»»ä¸€ç›®æ ‡åŒºåŸŸçš„IoUæœ€å¤§ï¼Œåˆ™è¯¥anchoråˆ¤å®šä¸ºæœ‰ç›®æ ‡ï¼›2ï¼‰å‡å¦‚æŸanchorä¸Žä»»ä¸€ç›®æ ‡åŒºåŸŸçš„IoU>0.7ï¼Œåˆ™åˆ¤å®šä¸ºæœ‰ç›®æ ‡ï¼›3ï¼‰å‡å¦‚æŸanchorä¸Žä»»ä¸€ç›®æ ‡åŒºåŸŸçš„IoU<0.3ï¼Œåˆ™åˆ¤å®šä¸ºèƒŒæ™¯ã€‚æ‰€è°"IoUï¼Œå°±æ˜¯é¢„æµ‹boxå'ŒçœŸå®žboxçš„è¦†ç›–çŽ‡ï¼Œå…¶å€¼ç‰äºŽä¸¤ä¸ªboxçš„äº¤é›†é™¤ä»¥ä¸¤ä¸ªboxçš„å¹¶é›†ã€‚å…¶å®ƒçš„anchorä¸å‚ä¸Žè®ç»ƒã€‚
- è"åˆè®ç»ƒ
å››æ¥è®ç»ƒæ³•ï¼š
1) å•ç‹¬è®ç»ƒRPNç½'ç»œï¼Œç½'ç»œå‚æ•°ç"±é¢„è®ç»ƒæ¨¡åž‹è½½å…¥
2) å•ç‹¬è®ç»ƒFast-RCNNç½'ç»œï¼Œå°†ç¬¬ä¸€æ¥RPNçš„è¾"å‡ºå€™é€‰åŒºåŸŸä½œä¸ºæ£€æµ‹ç½'ç»œçš„è¾"å…¥ã€‚å…·ä½"è€Œè¨€ï¼ŒRPNè¾"å‡ºä¸€ä¸ªå€™é€‰æ¡†ï¼Œé€šè¿‡å€™é€‰æ¡†æˆªå–åŽŸå›¾åƒï¼Œå¹¶å°†æˆªå–åŽçš„å›¾åƒé€šè¿‡å‡ æ¬¡conv-poolï¼Œç„¶åŽå†é€šè¿‡roi-poolingå'Œfcå†è¾"å‡ºä¸¤æ¡æ"¯è·¯ï¼Œä¸€æ¡æ˜¯ç›®æ ‡åˆ†ç±»softmaxï¼Œå¦ä¸€æ¡æ˜¯bboxå›žå½'ã€‚æˆªæ¢åˆ°çŽ°åœ¨ï¼Œä¸¤ä¸ªç½'ç»œå¹¶æ²¡æœ‰å…±äº«å‚æ•°ï¼Œåªæ˜¯åˆ†å¼€è®ç»ƒäº†;
3) å†æ¬¡è®ç»ƒRPNï¼Œæ¤æ—¶å›ºå®šç½'ç»œå…¬å…±éƒ¨åˆ†çš„å‚æ•°ï¼Œåªæ›´æ–°RPNç‹¬æœ‰éƒ¨åˆ†çš„å‚æ•°ï¼›
4) æ‹¿RPNçš„ç»"æžœå†æ¬¡å¾®è°ƒFast-RCNNç½'ç»œï¼Œå›ºå®šç½'ç»œå…¬å…±éƒ¨åˆ†çš„å‚æ•°ï¼Œåªæ›´æ–°Fast-RCNNç‹¬æœ‰éƒ¨åˆ†çš„å‚æ•°ã€‚



BBR

Bounding Box Regression
è¾¹æ¡†å›žå½’

å¯¹äºŽä¸Šå›¾ï¼Œç»¿è‰²çš„æ¡†ä»£è¡¨Ground Truthï¼Œçº¢è‰²çš„æ¡†ä¸ºSelective Searchæå–çš„Region Proposalã€‚å³ä¾¿çº¢è‰²çš„æ¡†è¢«åˆ†ç±»å™¨è¯†åˆ«ä¸ºé£žæœºï¼Œä½†æ˜¯ç”±äºŽçº¢è‰²çš„æ¡†å®šä½ä¸å‡†(IOU<0.5)ï¼Œé‚£ä¹ˆè¿™å¼ å›¾ç‰‡ç›¸å½“äºŽæ²¡æœ‰æ£€æµ‹åˆ°é£žæœºï¼Œéœ€è¦å¯¹æ¡†è¿›è¡Œå¾®è°ƒã€‚



çª—å£ä½¿ç”¨å››ç»´å'é‡(x,y,w,h)æ¥è¡¨ç¤ºï¼Œç›®æ ‡æ˜¯å¯»æ‰¾ä¸€ç§å…³ç³»ä½¿å¾—è¾“å…¥åŽŸå§‹çš„çª—å£Pç»è¿‡æ˜ å°„å¾—åˆ°ä¸€ä¸ªè·ŸçœŸå®žçª—å£Gæ›´æŽ¥è¿'çš„å›žå½'çª—å£G^ã€‚è¾¹æ¡†å›žå½'çš„ç›®çš„æ—¢æ˜¯ï¼šç»™å®š(Px,Py,Pw,Ph)å¯»æ‰¾ä¸€ç§æ˜ å°„fï¼Œä½¿å¾—
$$f(Px,Py,Pw,Ph) = (GÌ‚x,GÌ‚y,GÌ‚w, GÌ‚h)$$
å¹¶ä¸"$$(GÌ‚x,GÌ‚y,GÌ‚w, GÌ‚h) â‰ˆ (Gx,Gy,Gw, Gh)$$

- ***æ¥éª¤***
  å¹³ç§»+å°ºåº¦æ"¾ç¼©
  1) å…ˆåšå¹³ç§»(Î"x,Î"y)ï¼Œ Î"x=Pwdx(P), Î"y=Phdy(P) è¿™æ˜¯R-CNNè®ºæ–‡çš„ï¼š
  $$GÌ‚x=Pwdx(P)+Px,(1) $$
  $$GÌ‚y=Phdy(P)+Py,(2) $$
  2) ç„¶åŽå†åšå°ºåº¦ç¼©æ"¾(Sw,Sh), Sw=exp(dw(P)), Sh=exp(dh(P)), å¯¹åº"è®ºæ–‡ä¸ï¼š
  $$GÌ‚w=Pwexp(dw(P)),(3) $$
  $$GÌ‚h=Phexp(dh(P)),(4) $$
  è¾¹æ¡†å›žå½'å¦ä¹ çš„å°±æ˜¯dx(P),dy(P),dw(P),dh(P)è¿™å››ä¸ªå˜æ¢
- Input
  RegionProposalâ†'P=(Px,Py,Pw,Ph)RegionProposalâ†'P=(Px,Py,Pw,Ph) ï¼Œè¿™ä¸ªæ˜¯ä»€ä¹ˆï¼Ÿ è¾"å…¥å°±æ˜¯è¿™å››ä¸ªæ•°å€¼å—ï¼Ÿå…¶å®žçœŸæ£çš„è¾"å…¥æ˜¯è¿™ä¸ªçª—å£å¯¹åº"çš„ CNN ç‰¹å¾ï¼Œä¹Ÿå°±æ˜¯ R-CNN ä¸çš„ Pool5 featureï¼ˆç‰¹å¾å'é‡ï¼‰ã€‚(æ³¨ï¼šè®ç»ƒé˜¶æ®µè¾"å…¥è¿˜åŒ…æ‹¬ Ground Truthï¼Œ ä¹Ÿå°±æ˜¯ä¸‹è¾¹æåˆ°çš„tâˆ—=(tx,ty,tw,th)tâˆ—=(tx,ty,tw,th))
- output
  éœ€è¦è¿›è¡Œçš„å¹³ç§»å˜æ¢å'Œå°ºåº¦ç¼©æ"¾ dx(P),dy(P),dw(P),dh(P)dx(P),dy(P),dw(P),dh(P)ï¼Œ æˆ–è€…è¯´æ˜¯Î"x,Î"y,Sw,ShÎ"x,Î"y,Sw,Sh ã€‚æˆ'ä»¬çš„æœ€ç»ˆè¾"å‡ºä¸åº"è¯¥æ˜¯ Ground Truth å—ï¼Ÿ æ˜¯çš„ï¼Œ ä½†æ˜¯æœ‰äº†è¿™å››ä¸ªå˜æ¢æˆ'ä»¬å°±å¯ä»¥ç›´æŽ¥å¾—åˆ° Ground Truthï¼Œ è¿™é‡Œè¿˜æœ‰ä¸ªé—®é¢˜ï¼Œ æ ¹æ®(1)(4)æˆ'ä»¬å¯ä»¥çŸ¥é"ï¼Œ P ç»è¿‡ dx(P),dy(P),dw(P),dh(P)dx(P),dy(P),dw(P),dh(P) å¾—åˆ°çš„å¹¶ä¸æ˜¯çœŸå®žå€¼ Gï¼Œ è€Œæ˜¯é¢

„æµ‹å€¼GÌ‚ G^ã€, çš„ç¡®ï¼Œ è¿™å››ä¸ªå€¼å°”è¯¥æ˜¯ç»è¿‡ Ground Truth å’Œ Proposal è®¡ç®—å¾—åˆ°çš„çœŸæ£€æµ‹æ¡†çš„å¹³ç§»é‡(tx,ty)(tx,ty) å’Œå°ºåº¦ç¼©æ”¾(tw,th)(tw,th) ã€, è¿™ä¹Ÿå°±æ˜¯ R-CNN ä¸çš„(6)(9)ï¼š

$$ tx=(Gxâˆ'Px)/Pw,(6) $$

$$ ty=(Gyâˆ'Py)/Ph,(7) $$

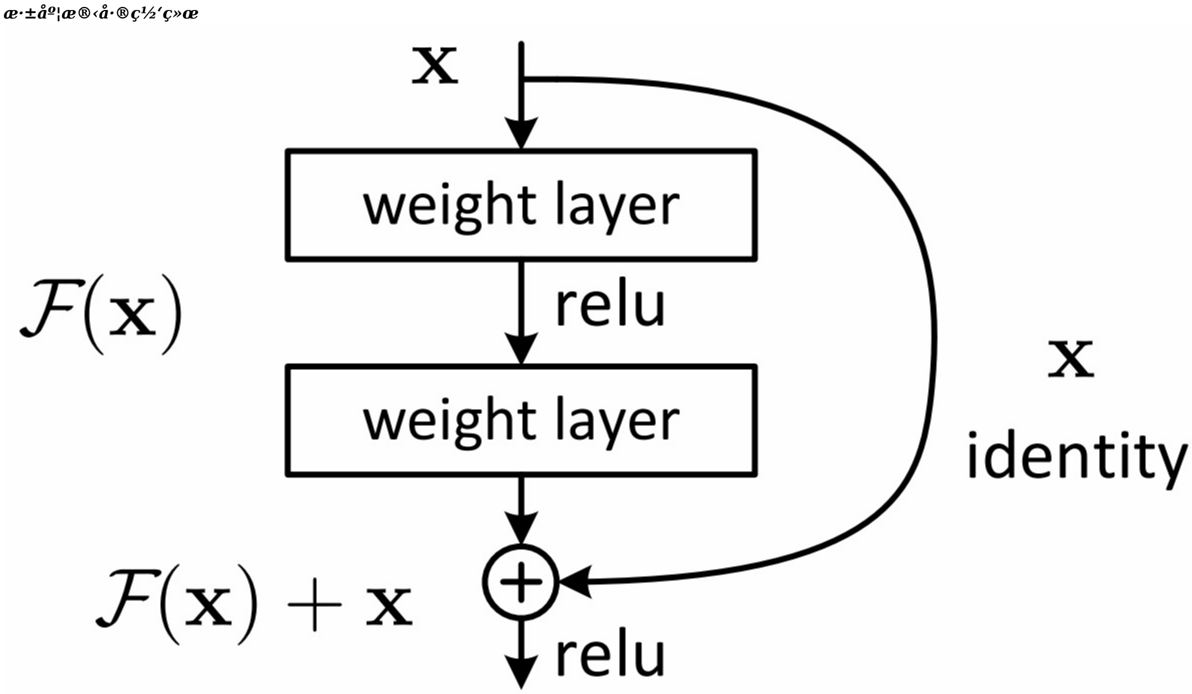
$$ tw=log(Gw/Pw),(8) $$

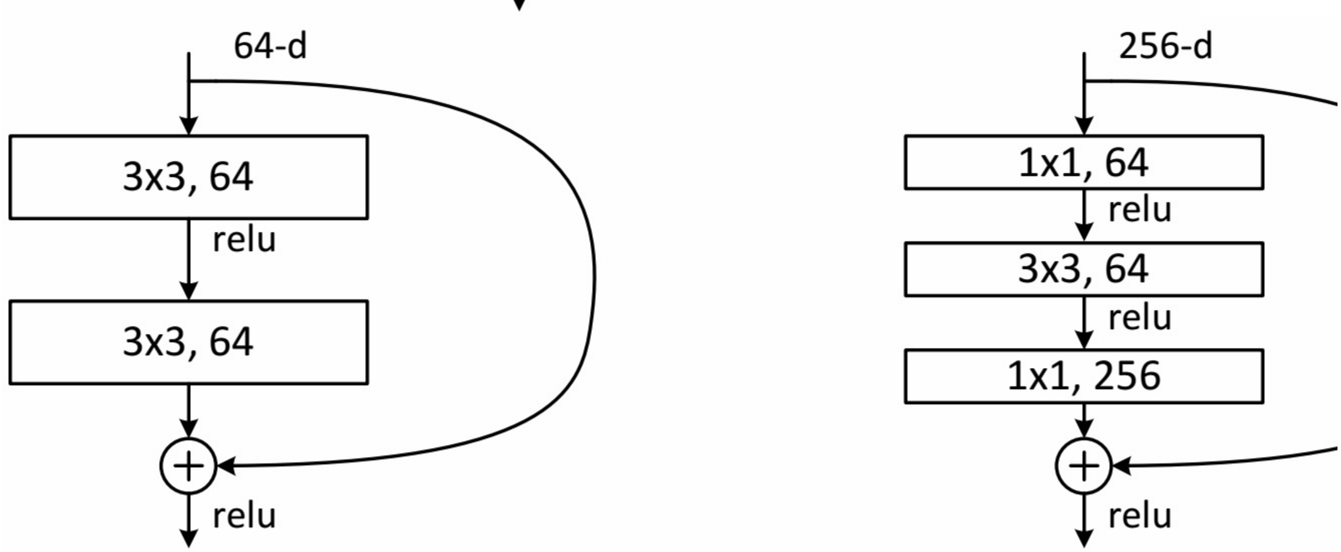
$$ th=log(Gh/Ph),(9) $$

é‚£ä¹ˆç›®æ ‡å‡½æ•°å¯ä»¥è¡¨ç¤ºä¸º dâˆ—(P)=wTâˆ—Î¦5(P)dâˆ—(P)=wâˆ—TÎ¦5(P)ï¼Œ Î¦5(P)Î¦5(P)æ˜¯è¾“å…¥ Proposal çš„ç‰¹å¾å'é‡ï¼Œwâˆ—wâˆ—æ˜¯è¦å¦ä¹ çš„å‚æ•°ï¼ˆ*è¡¨ç¤º x,y,w,hï¼Œ ä¹Ÿå°±æ˜¯æ¯ä¸€ä¸ªå˜æ¢å¯¹åº”ä¸€ä¸ªç›®æ ‡å‡½æ•°ï¼‰ , dâˆ—(P)dâˆ—(P) æ˜¯å¾—åˆ°çš„é¢„æµ‹å€¼ã€, æˆ'ä»¬è¦è®©é¢„æµ‹å€¼è·ŸçœŸå®žå€¼tâˆ—=(tx,ty,tw,th)tâˆ—=(tx,ty,tw,th)å·®è·æœ€å°ã€,

ResNet

æ·±åº¦æ®‹å·®ç½'ç»œ



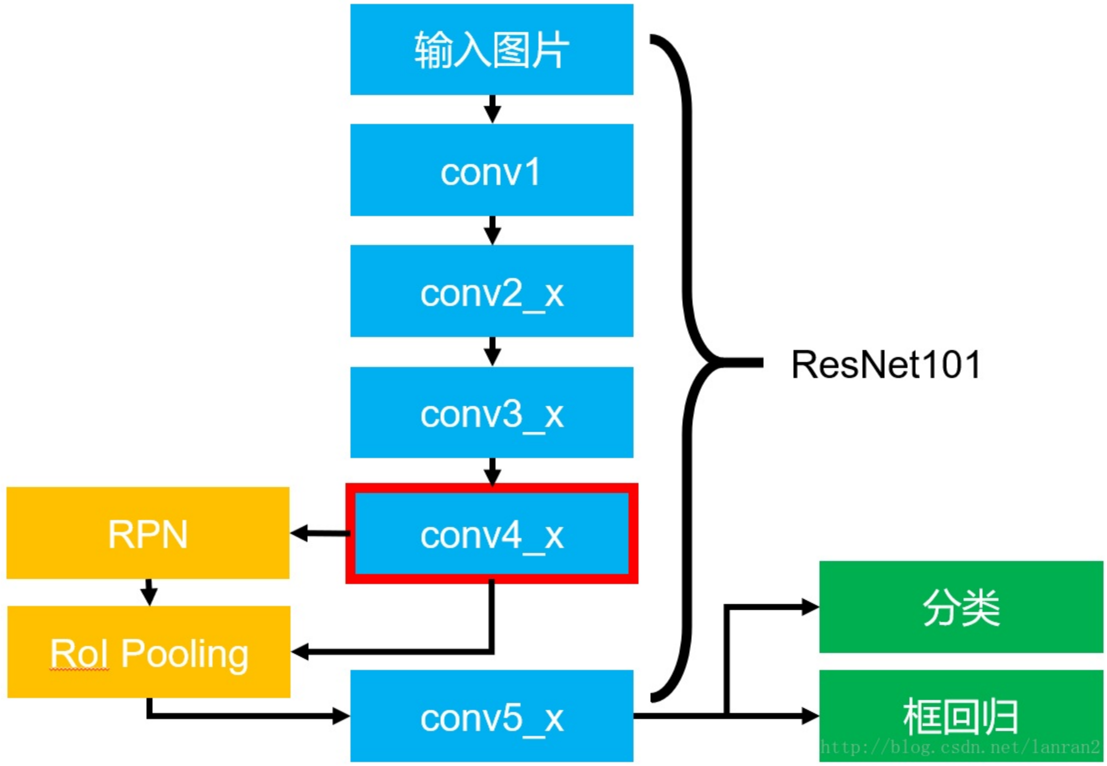


identity mapping "å¼¯å¼¯çš„æ›²çº¿" residual mapping â€œé™¤æ›²çº¿é‚£éƒ¨åˆ†â€ã€,

- ä¸¤ç§shortcut Connectionæ–¹å¼;
- 5ç§æ·±åº¦çš„ResNetï¼Œåˆ†åˆ«ä¸º18,34,50,101,152

```
æ‰€æœ‰çš„ç½'ç»œéƒ½åˆ†æˆ5éƒ¨åˆ†ï¼Œåˆ†åˆ«ä¸º:conv1,conv2_x,conv3_x,conv4_x,conv5_x
ä»¥101-layerä¸ºä¾‹ï¼Œé¦–å…ˆè¾“å…¥7Ã—7Ã—64çš„å·ç§¯ï¼Œç„¶åŽç»è¿‡3+4+23+3 = 33 ä¸ªbuilding blockï¼Œæ¯ä¸ªblockä¸º3å±‚ï¼Œæ‰€ä»¥33Ã—3 = 99 å±‚ï¼Œæœ€åŽæœ‰ä¸ªfcå±‚(ç"¨äºŽåˆ†ç±»)ï¼
tipï¼š101ä»…æŒ‡å·ç§¯æˆ–è€…å…¨è¿žæŽ¥å±‚ï¼Œè€Œæ¿€æ´»å±‚æˆ–è€…Poolingå±‚å¹¶æ²¡æœ‰è®¡ç®—åœ¨å†…ã€,
50-layerå'Œ101-layerä¸åŒåœ¨äºŽconv4_xï¼ŒResNet50æœ‰6ä¸ªblockï¼Œè€ŒResNet101æœ‰23ä¸ªblockï¼Œå·®äº†17ä¸ªblockï¼Œä¹Ÿå°±æ˜¯17Ã—3 = 51 å±‚ã€,
```

- åŸºäºŽResNet101çš„Faster RCNN



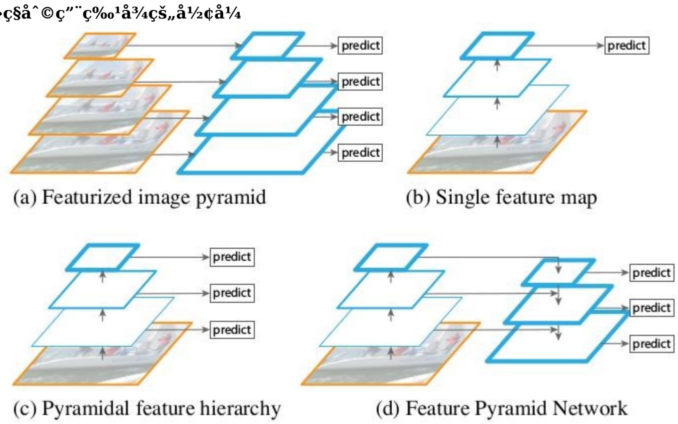
è"è‰²çš„éƒ¨åˆ†ä¸ºResNet101ï¼Œå¯ä»¥å‘çŽ°con4_xçš„æœ€åŽçš„è¾"å‡ºä¸ºRPNå'ŒROI Poolingçš„å…±äº«éƒ¨åˆ†ï¼Œè€Œconv5_x(å…±9å±‚ç½'ç»œ)éƒ½ä½œç"¨äºŽROI Poolingä¹‹åŽçš„ä¸€å †ç‰¹å¾å›¾(14Ã—14Ã—1024)ï¼Œç‰¹å¾å›¾çš„å¤§å°ç»´åº¦ä¹Ÿåˆšå¥½ç¬¦åˆåŽŸæœ¬çš„ResNet101ä¸conv5_xçš„è¾"å…¥;ä¸€å®šè¦è®°å¾—æœ€åŽæŽ¥ä¸€ä¸ªaverage poolingï¼Œå¾—åˆ°2048ç»´ç‰¹å¾ï¼Œåˆ†åˆ«ç"¨äºŽåˆ†ç±»å'Œæ¡†å›žå½'ã€‚

FPN

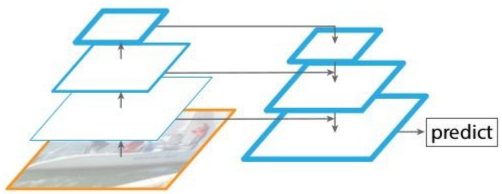
Feature Pyramid Networks

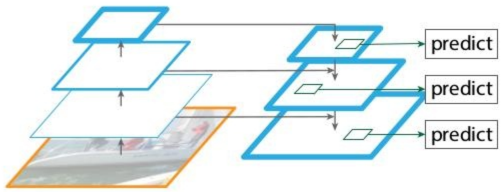
ä½Žå±‚çš„ç‰¹å¾è¯ä¹‰ä¿¡æ¯æ¯"è¾ƒå°'ï¼Œä½†æ˜¯ç›®æ ‡ä½ç½®å‡†ç¡®ï¼›é«˜å±‚çš„ç‰¹å¾è¯ä¹‰ä¿¡æ¯æ¯"è¾ƒä¸°å¯Œï¼Œä½†æ˜¯ç›®æ ‡ä½ç½®æ¯"è¾ƒç²—ç•¥ã€‚å¦å¤–è™½ç„¶ä¹Ÿæœ‰äº›ç®—æ³•é‡‡ç"¨å¤šå°ºåº¦ç‰¹å¾èžåˆçš„æ–¹å¼ï¼Œä½†æ˜¯ä¸€èˆ¬æ˜¯é‡‡ç"¨èžåˆåŽçš„ç‰¹å¾åšé¢„æµ‹ï¼Œè€Œæœ¬æ–‡ä¸ä¸€æ ·çš„åœ°æ–¹åœ¨äºŽé¢„æµ‹æ˜¯åœ¨ä¸åŒç‰¹å¾å±‚ç‹¬ç«‹è¿›è¡Œçš„ã€‚

- **å››ç§å^©ç"¨ç‰¹å¾çš„å½¢å¼**

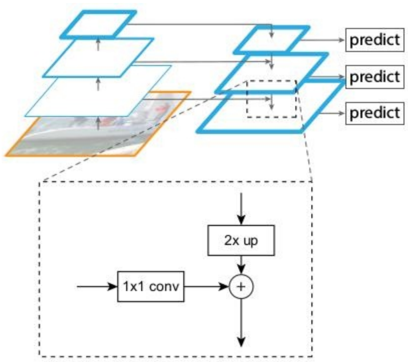



```
(a) å›¾åƒé‡'å—å¡"ï¼Œå³å°†å›¾åƒåšæˆä¸åŒçš„scaleï¼Œç„¶åŽä¸åŒscaleçš„å›¾åƒç"Ÿæˆå¯¹åº"çš„ä¸åŒscaleçš„ç‰¹å¾ã€‚
(b) åƒSPP netï¼ŒFast RCNNï¼ŒFaster RCNNæ˜¯é‡‡ç"¨è¿™ç§æ–¹å¼ï¼Œå³ä»…é‡‡ç"¨ç½'ç»œæœ€åŽä¸€å±‚çš„ç‰¹å¾ã€‚
(c) åƒSSDï¼ˆSingle Shot Detectorï¼‰é‡‡ç"¨è¿™ç§å¤šå°ºåº¦ç‰¹å¾èžåˆçš„æ–¹å¼ï¼Œæ²¡æœ‰ä¸Šé‡‡æ ·è¿‡ç¨‹ï¼Œå³ä»Žç½'ç»œä¸åŒå±‚æŠ½å–ä¸åŒå°ºåº¦çš„ç‰¹å¾åšé¢„æµ‹ï¼Œè¿™ç§æ–¹å¼ä¸ä¼šå¢žåŠ
(d) æœ¬æ–‡ä½œè€…æ˜¯é‡‡ç"¨è¿™ç§æ–¹å¼ï¼Œé¡¶å±‚ç‰¹å¾é€šè¿‡ä¸Šé‡‡æ ·å'Œä½Žå±‚ç‰¹å¾åšèžåˆï¼Œè€Œä¸"æ¯å±‚éƒ½æ˜¯ç‹¬ç«‹é¢„æµ‹çš„ã€‚
```






```
ä¸Šé¢ä¸€ä¸ªå¸¦æœ‰skip connectionçš„ç½‘ç»œç»“æž„åœ¨é¢„æµ‹çš„æ—¶å€™æ˜¯åœ¨finest levelï¼ˆè‡ªé¡¶å'ä¸‹çš„æœ€åŽä¸€å±‚ï¼‰è¿›è¡Œçš„ï¼Œç®€å•è®²å°±æ˜¯ç»è¿‡å¤šæ¬¡ä¸Šé‡‡æ ·å¹¶èžåˆç‰¹å
```

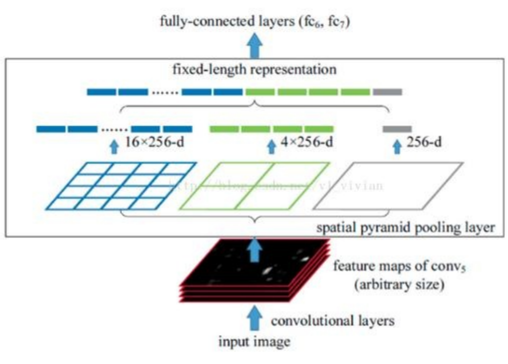



```
ä½œè€…çš„ä¸»ç½‘ç»œé‡‡ç"¨ResNetï¼Œç®—æ³•å¤§è‡´ç»"æž„å¦‚ä¸Šå›¾ï¼š**ä¸€ä¸ªè‡ªä½Žå'ä¸Šçš„çº¿è·¯ï¼Œä¸€ä¸ªè‡ªé¡¶å'ä¸‹çš„çº¿è·¯ï¼Œæ¨ªå'è¿žæŽ¥(lateral connection)**ï¼Œå›¾ä¸æ"¾å¤§ç
```

è‡ªåº•å'ä¸Šå…¶å®žå°±æ˜¯ç½'ç»œçš„å‰å'è¿‡ç¨‹ã€,

SppNet

Spatial Pyramid Pooling ç©ºé—´é‡'å—å¡"æ± åŒ–
å…¨è¿žæŽ¥å±‚éœ€è¦æŒ‡å®šè¾"å…¥å±‚ç¥žç»å…ƒä¸ªæ•°ï¼Œæ‰€ä»¥éœ€è¦è§„å®šè¾"å…¥çš„featureçš„å¤§å°ï¼Œå¯¼è‡´ç¥žç»ç½'ç»œéœ€è¦è¾"å…¥ç»è¿‡cropã€warpæ"ä½œçš„å›ºå®šå°ºå¯¸çš„å›¾ç‰‡ã€,
ORI: images -> crop/warp -> conv layers -> fc layers -> output
SPP: images -> conv layers -> Spatial Pyramid Pooling -> fc layers -> output
- æ•´ä½"è¿‡ç¨‹




```
é»'è‰²å›¾ç‰‡ä»£è¡¨å·ç§¯ä¹‹åŽçš„ç‰¹å¾å›¾ï¼ŒæŽ¥ç€æˆ'ä»¬ä»¥ä¸åŒå¤§å°çš„å—æ¥æå–ç‰¹å¾ï¼Œåˆ†åˆ«æ˜¯4Ã—4,2Ã—2,1Ã—1ï¼Œæ"¾åˆ°ç‰¹å¾å›¾ä¸Šå°±å¯ä»¥å¾—åˆ°16+4+1=21ç§ä¸åŒå¤§å°çš„å—(Spati
è¾"å‡ºå'é‡å¤§å°ä¸ºMKï¼ŒM=#binsï¼ŒK=#filtersï¼Œä½œä¸ºå…¨è¿žæŽ¥å±‚çš„è¾"å…¥ï¼Œä¾‹å¦‚ä¸Šå›¾ï¼Œconv5è®¡ç®—å‡ºçš„feature mapæ˜¯ä»»æ„å¤§å°çš„ï¼Œç»è¿‡SPPåŽï¼Œå°±å¯ä»¥å˜æˆå›ºå®šå¤
```

RetinaNet

detectorä¸»è¦åˆ†ä¸ºä»¥ä¸‹ä¸¤å¤§é—¨æ´¾ï¼š
one stage–> YOLOV1–3,SSD ç²¾åº¦ä½Žä½†é€Ÿåº¦å¿«
two stage–> R-CNN,SPPNet,Fast/Faster R-CNN ç²¾åº¦é«˜ä½†é€Ÿåº¦æ…¢

```
1) one stageå—å ¶äºŽâ€˜ç±»åˆ«ä¸å¹³è¡¡â€™(æ£€æµ‹ç®—æ³•æ©æœŸä¼šç"Ÿæˆä¸€å¤§æ³¢bboxï¼Œç»å¤§å¤šæ•°å±žäºŽè´Ÿæ ·æœ¬backgroundï¼Œå³ä½¿å†ç±»å™¨æ— è„'åœ°æŠŠæ‰€æœ‰çš„bboxåŒä¸€å½'ç±
2) two stageæœ‰RPNï¼Œç¬¬ä¸€é˜¶æ®µçš„RPNä¼šå¯¹anchorè¿›è¡Œç®€å•çš„äºŒåˆ†ç±»(åªæ˜¯ç®€å•çš„åŒºåˆ†æ˜¯å‰æ™¯è¿˜æ˜¯èƒŒæ™¯ï¼Œå¹¶ä¸åŒºåˆ«ç©¶ç«Ÿå±žäºŽé‚£ä¸ªç±»)ï¼›ç¬¬äºŒé˜¶æ®µåˆ†ç±»
```

- focal loss

focal loss的标准公式非常简单：

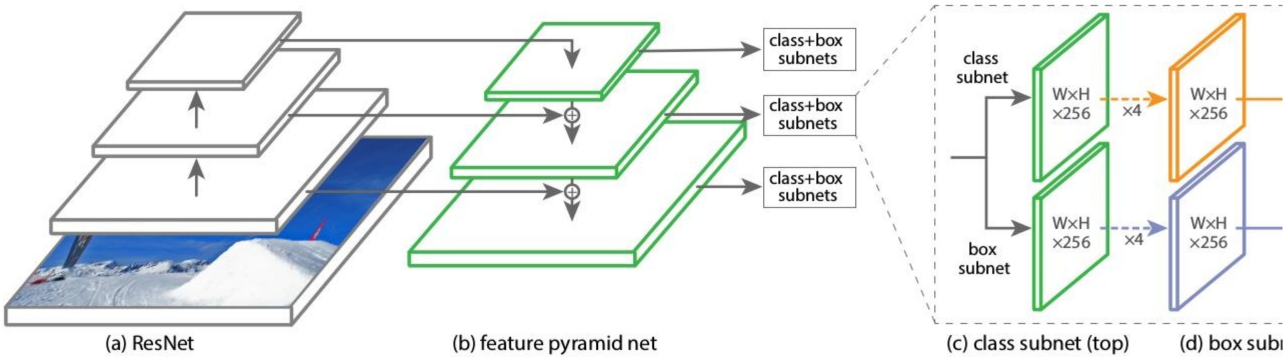
$$FL(p_t) = -(1-p_t)^{\gamma}\log(p_t) = (1-p_t)^{\gamma}CE(\hat{y})_i$$

也可以更复杂一点（论文中的实验即采用此公式）：

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t)$$

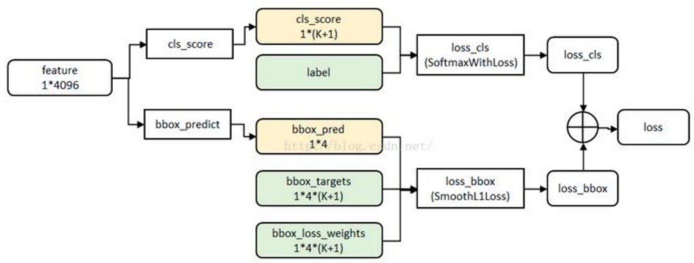
åœ¨åŽŸæœ¬çš„äº¤å‰ç†µæŸå¤±å‡½æ•°å‰ä¹˜ä¸Šä¸€ä¸ªæƒé‡ï¼Œå®žéªŒä¸å‘çŽ°Î³=2ï¼ŒÎ±=0.25çš„å–å€¼ç»„åˆæ•ˆæžœæœ€å¥½ã€‚

- RetinaNet:



FasterRCNNä¸çš„SmoothL1Loss

- **Fast RCNNæŸå¤±å‡½æ•°ï¼Œå¤šä»»åŠ¡æŸå¤±**



multi-task数据结构

fast rcnnç½'ç»œæœ‰ä¸¤ä¸ªåŒçº§è¾"å‡ºå±‚(cls scoreå'Œbbox_predictå±‚)ï¼Œéƒ½æ˜¯å…¨è¿žæŽ¥å±‚ï¼Œç§°ä¸ºmulti-task
1) clsscoreå±‚ï¼Œç"¨äºŽåˆ†ç±»ï¼Œè¾"å‡ºk+1ç»´æ•°ç»„pï¼Œè¡¨ç¤ºå±žäºŽkç±»å'ŒèƒŒæ™¯çš„æ¦‚çŽ‡ã€‚å¯¹æ¯ä¸ªROIè¾"å‡ºç¦»æ•£åž‹æ¦‚çŽ‡åˆ†å¸ƒï¼š
$p = (p_0, p_1, \cdots, p_k)$ ï¼Œé€šå¸¸ï¼Œpç"±k+1ç±»çš„å…¨è¿žæŽ¥å±‚åˆ©ç"¨softmaxè®¡ç®—å¾—å‡ºã€‚
2) bbox_predictå±‚ï¼Œç"¨äºŽè°ƒæ•´å€™é€‰åŒºåŸŸä½ç½®ï¼Œè¾"å‡ºbounding boxå›žå½'çš„ä½ç§»ï¼Œè¾"å‡º4*kç»´æ•°ç»„tï¼Œè¡¨ç¤ºåˆ†åˆ«å±žäºŽkç±»æ—¶ï¼Œåº"è¯¥å¹³ç§»ç¼©æ"¾çš„å‚æ•°ï¼š
$t^k = (t^k_x, t^k_y, t^k_w, t^k_h)$

3) loss_clså±‚è¯„ä¼°åˆ†ç±»æŸå¤±å‡½æ•°ï¼Œloss_bboxè¯„ä¼°æ£€æµ‹æ¡†å®šä½çš„æŸå¤±å‡½æ•°ï¼Œæ¯"è¾ƒçœŸå®žåˆ†ç±»å¯¹åº"çš„é¢„æµ‹å¹³ç§»ç¼©æ"¾å‚æ•°å'ŒçœŸå®žå¹³ç§»ç¼©æ"¾å‚æ•°çš„å·®åˆ«ï¼š

$$L_{\text{loc}}(t^u, v) = \sum_{i \in \{\text{x,y,w,h}\}} \text{smooth}_{L_1}(t^u_i - v_i), \qquad (2)$$

in which

$$\text{smooth}_{L_1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise,} \end{cases} \qquad (3)$$

ä½¿å¾—lossåˆ¤äºŽç¦»ç¾¤ç‚¹æ›´åŠ é²æ£'ã€‚
æœ€åŽæ€»æŸå¤±ä¸ºä¸¤è€…åŠ æƒå'Œï¼Œå¦‚æžœåˆ†ç±»ä¸ºèƒŒæ™¯åˆ™ä¸è€ƒè™'å®šä½æŸå¤±ã€‚

$$L(p, u, t^u, v) = L_{\text{cls}}(p, u) + \lambda[u \geq 1] L_{\text{loc}}(t^u, v), \qquad (1)$$

è‰¾å¼—æ£®æ‹¬å·æŒ‡æ•°å‡½æ•°[uâ‰¥1]è¡¨ç¤ºèƒŒæ™¯å€™é€‰åŒºåŸŸå³è´Ÿæ ·æœ¬ä¸å‚ä¸Žå›žå½'æŸå¤±ï¼Œä¸éœ€è¦å¯¹å€™é€‰åŒºåŸŸè¿›è¡Œå›žå½'æ"ä½œã€‚Î»æŽ§åˆ¶åˆ†ç±»æŸå¤±å'Œå›žå½'æŸå¤±çš„å¹³è¡¡ã€‚Fast R-CNNè®ºæ–‡ä¸ï¼Œæ‰€æœ‰å®žéªŒÎ»=1ã€‚

- **Faster RCNNæŸå¤±å‡½æ•°**

$$L(\{p_i\}, \{t_i\}) = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(t_i, t_i^*). \qquad (1)$$

å…¶ä¸ï¼š

p_i为 anchor 预测为目标的概率;

GT标签：$p_i^* = \begin{cases} 0 & negative\ label \\ 1 & positive\ label \end{cases}$;

$t_i = \{t_x, t_y, t_w, t_h\}$是一个向量，表示预测的 bounding box 包围盒的 4 个参数化坐标;

t_i^*是与 positive anchor 对应的 ground truth 包围盒的坐标向量;

$L_{cls}(p_i, p_i^*)$是两个类别（目标 vs.非目标）的对数损失:

$$L_{cls}(p_i, p_i^*) = -\log\left[p_i^* p_i + (1-p_i^*)(1-p_i)\right]$$

$L_{reg}(t_i, t_i^*)$是回归损失，用$L_{reg}(t_i, t_i^*) = R(t_i - t_i^*)$来计算，R 是 smooth L1 函数。

$p_i^* L_{reg}$这一项意味着只有前景 anchor（$p_i^* = 1$）才有回归损失，其他情况就没有（$p_i^* = 0$）。cls 层和 reg 层的输出分别由$\{p_i\}$和$\{u_i\}$组成，这两项分别由N_{cls}和N_{reg}以及一个平衡权重λ归一化（早期实现及公开的代码中，$\lambda = 10$，cls 项的归一化值为 mini-batch 的大小，即$N_{cls} = 256$，reg 项的归一化值为 anchor 位置的数量，即$N_{reg} \sim 2400$（$40 * 60$），这样 cls 和 reg 项差不多是等权重的。

SmoothL1LossLayer 计算一张图片的损失函数

$$\lambda \frac{1}{N_{reg}} \sum_i p_i{}^* L_r eg(t_i, t_i{}^*) \tag{1}$$

- i ： mini-batch 的 anchor 的索引。
- p_i ：目标的预测概率。
- $p_i{}^*$ ：target二分类是否有物体，有物体为1，否则为0。
- t_i 是一个四点向量，预测坐标
- $t_i{}^*$ 是一个四点向量，是ground truth boungding box的坐标（真实坐标）

$$L_{reg}(t, t_i^*) = R(t_i - t_i *) \tag{2}$$

bottom[0]预测坐标，即 t_i
bottom[1]target坐标，即$t_i{}^*$
bottom[2]inside，有物体，即有前景(foreground)时为1，否则为0，即$p_i{}^*$
bottom[3]outside，没有前景（fg）也没有后景（bg）的为0，其他为1/（bg+fg），对应于加号右边的系数部分。

Lreg的公式如下，其中 $x = t_i - t_i *$，

$$smooth_{L1}(x) = \begin{cases} 0.5x^2, if\ |x| < 1 \\ |x| - 0.5, otherwise \end{cases} \tag{3}$$

$p_i * L_{reg}(t_i, t_i{}^*)$表明只有有fg（20个物体类别）的才有regression loss.

SmoothL1LossLayer 计算一张图片的损失函数

$$\lambda \frac{1}{N_{reg}} \sum_i p_i{}^* L_r eg(t_i, t_i{}^*) \tag{1}$$

- i ： mini-batch 的 anchor 的索引。
- p_i ：目标的预测概率。
- $p_i{}^*$ ：target二分类是否有物体，有物体为1，否则为0。
- t_i 是一个四点向量，预测坐标
- $t_i{}^*$ 是一个四点向量，是ground truth boungding box的坐标（真实坐标）

$$L_{reg}(t, t_i^*) = R(t_i - t_i *) \tag{2}$$

bottom[0]预测坐标，即 t_i
bottom[1]target坐标，即$t_i{}^*$
bottom[2]inside，有物体，即有前景(foreground)时为1，否则为0，即$p_i{}^*$
bottom[3]outside，没有前景（fg）也没有后景（bg）的为0，其他为1/（bg+fg），对应于加号右边的系数部分。

Lreg的公式如下，其中 $x = t_i - t_i *$ ，

$$smooth_{L1}(x) = \begin{cases} 0.5x^2, if\,|x| < 1 \\ |x| - 0.5, otherwise \end{cases} \tag{3}$$

$p_i * L_{reg}(t_i, t_i{}^*)$ 表明只有有fg（20个物体类别）的才有regression loss.

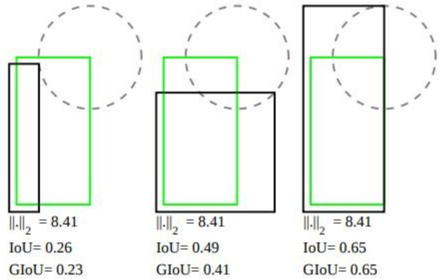
GIOU

- ***IOUä½œä¸ºæŸå¤±å‡½æ•°çš„é—®é¢˜***
  1) å¦‚æžœä¸¤ä¸ªå¯¹è±¡ä¸é‡å ï¼ŒIOUå€¼å°†ä¸ºé›¶ï¼Œä¸è®ºç›¸å·®å¤šè¿œï¼Œå³ä¸ä¼šåæ˜ ä¸¤ä¸ªå½¢çŠ¶å½¼æ¤ä¹‹é—´çš„è·ç¦»ï¼Œå¦‚æžœç”¨IOUç”¨ä½œæŸè€—ï¼Œåˆ™å…¶æ¢¯åº¦å°†ä¸ºé›¶ï¼Œæ— æ³•ä¼˜åŒ–ã€‚
  2) IOUçš„å€¼æ— æ³•åæ˜ ä¸¤ä¸ªå¯¹è±¡ä¹‹é—´å¦‚ä½•é‡å ï¼ŒåŒä¸€ä¸ªIOUå€¼ä¸¤ä¸ªå¯¹è±¡é—´æœ‰å¤šç§é‡å æ–¹å¼ã€‚
- ***GIOUè®¡ç®—***

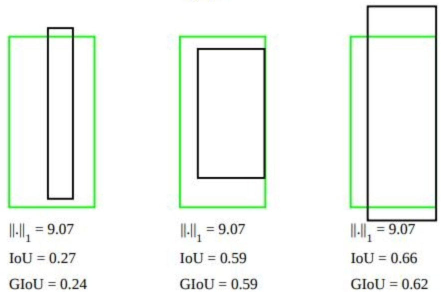
Algorithm 1: Generalized Intersection over Union

input : Two arbitrary convex shapes: $A, B \subseteq \mathbb{S} \in \mathbb{R}^n$
output: $GIoU$
1 For A and B, find the smallest enclosing convex object C, where $C \subseteq \mathbb{S} \in \mathbb{R}^n$
2 $IoU = \dfrac{|A \cap B|}{|A \cup B|}$
3 $GIoU = IoU - \dfrac{|C \backslash (A \cup B)|}{|C|}$

- ***IOUä¸ŽGIOUå¯¹æ¯”***



(a)



(b)

- ***IOUä¸ŽGIOU lossè®¡ç®—è¿‡ç¨‹***

Algorithm 2: IoU and $GIoU$ as bounding box losses

input : Predicted B^p and ground truth B^g bounding box coordinates:
$B^p = (x_1^p, y_1^p, x_2^p, y_2^p)$, $\quad B^g = (x_1^g, y_1^g, x_2^g, y_2^g)$.

output: $\mathcal{L}_{IoU}, \mathcal{L}_{GIoU}$.

1 For the predicted box B^p, ensuring $x_2^p > x_1^p$ and $y_2^p > y_1^p$:
$\hat{x}_1^p = \min(x_1^p, x_2^p)$, $\quad \hat{x}_2^p = \max(x_1^p, x_2^p)$,
$\hat{y}_1^p = \min(y_1^p, y_2^p)$, $\quad \hat{y}_2^p = \max(y_1^p, y_2^p)$.

2 Calculating area of B^g: $A^g = (x_2^g - x_1^g) \times (y_2^g - y_1^g)$.

3 Calculating area of B^p: $A^p = (\hat{x}_2^p - \hat{x}_1^p) \times (\hat{y}_2^p - \hat{y}_1^p)$.

4 Calculating intersection $\mathcal{I}$ between B^p and B^g:
$x_1^{\mathcal{I}} = \max(\hat{x}_1^p, x_1^g)$, $\quad x_2^{\mathcal{I}} = \min(\hat{x}_2^p, x_2^g)$,
$y_1^{\mathcal{I}} = \max(\hat{y}_1^p, y_1^g)$, $\quad y_2^{\mathcal{I}} = \min(\hat{y}_2^p, y_2^g)$,
$$\mathcal{I} = \begin{cases} (x_2^{\mathcal{I}} - x_1^{\mathcal{I}}) \times (y_2^{\mathcal{I}} - y_1^{\mathcal{I}}) & \text{if} \quad x_2^{\mathcal{I}} > x_1^{\mathcal{I}}, y_2^{\mathcal{I}} > y_1^{\mathcal{I}} \\ 0 & \text{otherwise.} \end{cases}$$

5 Finding the coordinate of smallest enclosing box B^c:
$x_1^c = \min(\hat{x}_1^p, x_1^g)$, $\quad x_2^c = \max(\hat{x}_2^p, x_2^g)$,
$y_1^c = \min(\hat{y}_1^p, y_1^g)$, $\quad y_2^c = \max(\hat{y}_2^p, y_2^g)$.

6 Calculating area of B^c: $A^c = (x_2^c - x_1^c) \times (y_2^c - y_1^c)$.

7 $IoU = \dfrac{\mathcal{I}}{\mathcal{U}}$, where $\mathcal{U} = A^p + A^g - \mathcal{I}$.

8 $GIoU = IoU - \dfrac{A^c - \mathcal{U}}{A^c}$.

9 $\mathcal{L}_{IoU} = 1 - IoU$, $\quad \mathcal{L}_{GIoU} = 1 - GIoU$.

RefineDet

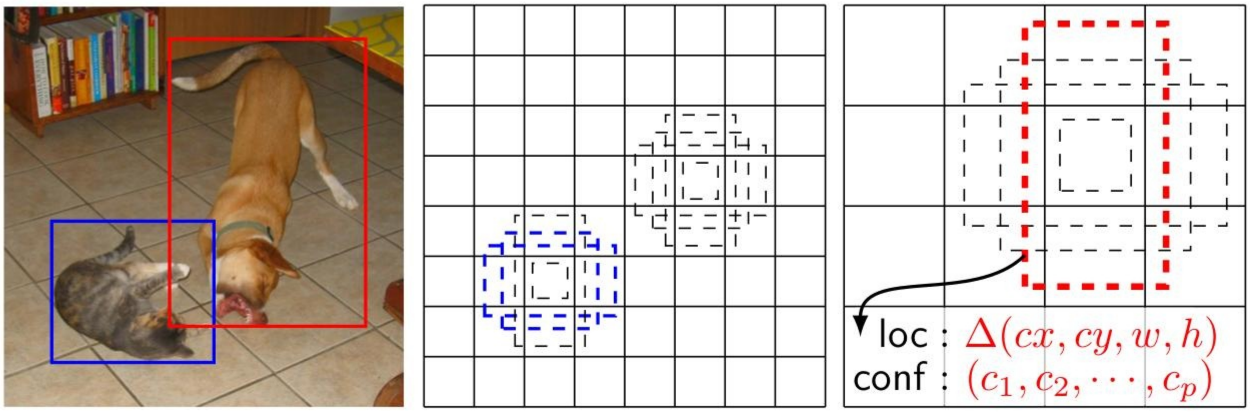
Single-Shot Refinement Neural Network for Object Detection(CVPR2018)

SSD

single shot multibox detector

- **æœ¬æ–‡æå‡ºçš„SSDç®—æ³•æ˜¯ä¸€ç§ç›´æŽ¥é¢„æµ‹bounding boxçš„åæ ‡å’Œç±»åˆ«çš„object detectionç®—æ³•ï¼Œæ²¡æœ‰ç”Ÿæˆproposalçš„è¿‡ç¨‹**
- åŸºæœ¬çš„ç½‘ç»œç»“æž„æ˜¯åŸºäºŽVGG16ï¼Œåœ¨ImageNetæ•°æ®é›†ä¸Šé¢„è®ç»ƒå®Œä»¥åŽç”¨ä¸¤ä¸ªæ–°çš„å·ç§¯å±‚ä»£æ›¿fc6å’Œfc7ï¼Œå¦å¤–å¯¹pool5ä¹Ÿåšäº†ä¸€ç‚¹æ”¹åŠ¨ï¼Œè¿˜å¢žåŠ äº†4ä¸ªå·ç§¯å±‚æž„æˆæœ¬æ–‡çš„ç½‘ç»œã€‚
1) æ–‡ç« çš„ä¸€ä¸ªæ ¸å¿ƒæ˜¯ä½œè€…åŒæ—¶é‡‡ç”¨lowerå’Œupperçš„feature mapsåšæ£€æµ‹
2) default boxï¼Œæ˜¯æŒ‡åœ¨feature mapçš„æ¯ä¸ªå°æ ¼(cell)ä¸Šéƒ½æœ‰ä¸€ç³»åˆ—å›ºå®šå¤§å°çš„boxï¼Œå¦‚ä¸‹å›¾æœ‰4ä¸ª(è™šçº¿æ¡†)ï¼Œå‡è®¾æ¯ä¸ªfeature cellæœ‰kä¸ªdefault boxï¼Œé‚£ä¹ˆå¯¹äºŽæ¯ä¸ªdefault boxéƒ½éœ€è¦é¢„æµ‹cä¸ªç±»åˆ«scoreå’Œoffsetï¼Œé‚£ä¹ˆå¦‚æžœä¸€ä¸ªfeature mapçš„å¤§å°æ˜¯mÃ—nï¼Œä¹Ÿå°±æ˜¯æœ‰mÃ—nä¸ªfeature map cellï¼Œé‚£ä¹ˆè¿™ä¸ªfeature mapå°±ä¸€å…±æœ‰kÃ—mÃ—nä¸ªdefault boxï¼Œæ¯ä¸ªdefault boxéƒ½éœ€è¦é¢„æµ‹4ä¸ªåæ ‡ç›¸å…³çš„å€¼å’Œc+1ä¸ªç±»åˆ«æ¦‚çŽ‡(**å®žé™…codeä¸åˆ†å«ç”¨ä¸åŒæ•°é‡çš„3Ã—3å·ç§¯æ ¸å¯¹è¯¥å±‚feature mapè¿›è¡Œå·ç§¯ï¼Œä¸€å¦‚å·ç§¯æ ¸æ•°é‡ä¸º(c+1)Ã—kå¯¹åº”confidenceè¾“å‡ºï¼Œè¡¨ç¤ºæ¯ä¸ªdefault boxçš„confidenceï¼Œå°±æ˜¯ç±»åˆ«ï¼›å·ç§¯æ ¸æ•°é‡ä¸º4Ã—kå¯¹åº”localizationè¾“å‡ºï¼Œè¡¨ç¤ºæ¯ä¸ªdefault**

boxçš„åæ ‡)ï¼Œdefault boxä¸ŽFaster RCNNä¸anchorå¾ˆåƒï¼ŒFaster RCNNåªç”¨åœ¨æœ€åŽä¸€ä¸ªå·ç§¯å±‚ï¼Œä½†SSDä¸ç”¨åœ¨å¤šä¸ªä¸åŒå±‚çš„feature mapä¸Šã€‚ä¸‹å›¾è¯´æ˜Žæœ‰ä¸ªé‡è¦çš„æ¦‚å¿µï¼šè®ç»ƒé˜¶æ®µï¼Œç®—æ³•åœ¨ä¸€å¼ å›¾ç‰‡é‡Œé¢…ä¸Šåˆ†è¦ä¹ˆdefault boxå'Œground truthboxè¿›è¡ŒåŒ¹é…ï¼Œæ¯”å¦‚è¿™é‡Œçš„ä¸¤ä¸ªè“è‰²å'Œä¸€ä¸ªçº¢è‰²ground truth boxåŒ¹é…ä¸Šäº†ï¼Œå³ä¸€ä¸ªground truthå¯èƒ½å¯¹åº”å¤šä¸ªdefault box;åœ¨é¢„æµ‹é˜¶æ®µï¼Œç›´æŽ¥é¢„æµ‹æ¯ä¸ªdefault boxçš„åç§»ä»¥åŠæ¯ä¸ªç±»åˆ«ç›¸åº”çš„å¾—åˆ†ï¼Œæœ€åŽé€šè¿‡NMSå¾—åˆ°æœ€ç»ˆçš„ç»“æžœ



(a) Image with GT boxes (b) 8×8 feature map (c) 4×4 feature map

- **default boxçš„scale(å¤§å°)å'Œaspect ratio(çºµæ¨ªæ¯”)**

1) scaleè®¡ç®—å…¬å¼ï¼š

$$s_k = s_{\min} + \frac{s_{\max} - s_{\min}}{m-1}(k-1), \quad k \in [1, m] \tag{4}$$

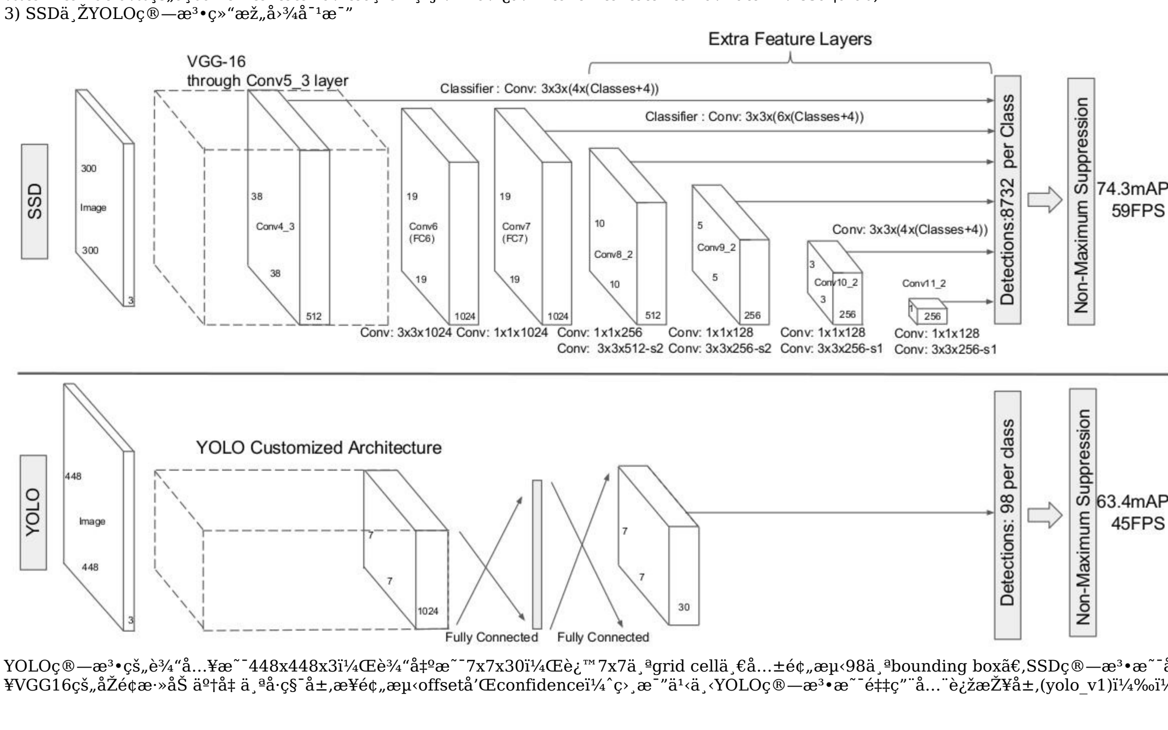
è¿™é‡Œçš„sminä¸º0.2ï¼Œè¡¨ç¤ºæœ€åº•å±‚çš„scaleå¤§å°;smaxä¸º0.9ï¼Œè¡¨ç¤ºæœ€é«˜å±‚çš„scaleå¤§å°

2) aspect ratioè®¡ç®—ï¼š

We impose different aspect ratios for the default boxes, and denote them as $a_r \in \{1, 2, 3, \frac{1}{2}, \frac{1}{3}\}$. We can compute the width ($w_k^a = s_k\sqrt{a_r}$) and height ($h_k^a = s_k/\sqrt{a_r}$) for each default box. For the aspect ratio of 1, we also add a default box whose scale is $s'_k = \sqrt{s_k s_{k+1}}$, resulting in 6 default boxes per feature map location. We set the center of each default box to $(\frac{i+0.5}{|f_k|}, \frac{j+0.5}{|f_k|})$, where $|f_k|$ is the size of the k-th square feature map, $i, j \in [0, |f_k|)$. In practice, one can also design a distribution of default boxes to best fit a specific dataset. How to design the optimal tiling is an open question as well.

å› æ¤ï¼Œå¯¹äºŽæ¯ä¸ªfeature map cellè€Œè¨€ï¼Œä¸€å…±æœ‰6ä¸ªdefault boxï¼Œå¯ä»¥çœ‹å‡ºè¿™ç§default boxåœ¨ä¸åŒçš„featureå±‚æœ‰ä¸åŒçš„scaleï¼Œåœ¨åŒä¸€ä¸ªfeatureå±‚åˆæœ‰ä¸åŒçš„aspect ratioï¼Œå› æ¤åŸºæœ¬ä¸Šå¯ä»¥è¦†ç›–è¾“å…¥å›¾åƒä¸çš„å„ç§å½¢çŠ¶å'Œå¤§å°çš„objectï¼æœ‰ä¸ªè¿™ç§æ–¹æ³•ï¼ˆè¿™é‡Œè¿˜æ²¡æœ‰æ£æ ·æœ¬çš„åŒ¹é…ï¼Œåªæ˜¯ç½®ä¿¡åº¦ï¼‰æ£è´Ÿæ ·æœ¬æ¯”ä¾‹ï¼Œå¹¶ä¸”åœ¨confidence lossè¿›è¡ŒæŽ'åºï¼Œç„¶åŽé€‰æ‹©é¢„æµ‹ç»“æžœçš„ä¸€ä¸ªè´Ÿæ ·æœ¬ä½œä¸ºè®ç»ƒæ ·æœ¬ï¼Œä½¿å¾—æ£è´Ÿæ ·æœ¬æ¯”ä¾‹åœ¨1ï¼š3å·¦å³ã€‚

3) SSDä¸ŽYOLOçš„ç½‘ç»œæž¶æž„å¯¹æ¯”



YOLOç®—æ³•çš„è¾“å…¥æ˜¯448x448x3ï¼Œè¾“å‡ºæ˜¯7x7x30ï¼Œè¿™7x7ä¸ªgrid cellä¸€å…±é¢„æµ‹98ä¸ªbounding boxã€‚SSDç®—æ³•æ˜¯åœ¨åŽŸæ¥VGG16çš„åŽé¢æ·»åŠ äº†å‡ ä¸ªå·ç§¯å±‚æ¥é¢„æµ‹offsetå'Œconfidenceï¼ˆç›¸æ¯"äºŽä¸‹YOLOç®—æ³•æ˜¯é‡‡ç"¨å…¨è¿žæŽ¥å±‚(yolo_v1)ï¼‰ï¼Œç®——

æ³•çš„è¾“å…¥æ˜¯300x300x3ï¼Œé‡‡ç”¨conv4_3ï¼Œconv7ï¼Œconv8_2ï¼Œconv9_2ï¼Œconv10_2å’Œconv11_2çš„è¾“å‡ºæ¥é¢„æµ‹locationå’Œconfidenceã
€‚

```
tip: è¯¦ç»†è®²ä¸€ä¸‹SSDçš„ç»“æž„ï¼Œå¯ä»¥ç œ‹Caffeä»£ç  ã€‚SSDçš„ç»“æž„ä¸ºconv1_1ï¼Œconv1_2ï¼Œconv2_1ï¼Œconv2_2ï¼Œconv3_1ï¼Œconv3_2ï¼Œconv3_3ï¼Œconv4_1ï¼Œconv4_2ï¼Œconv4_3ï
```