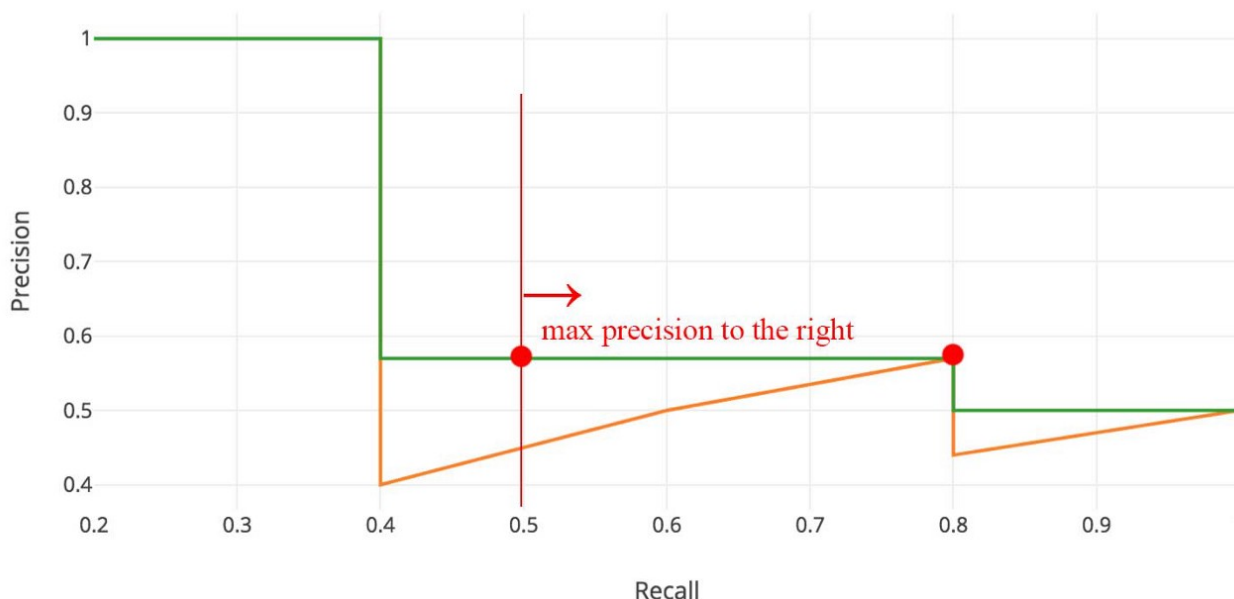


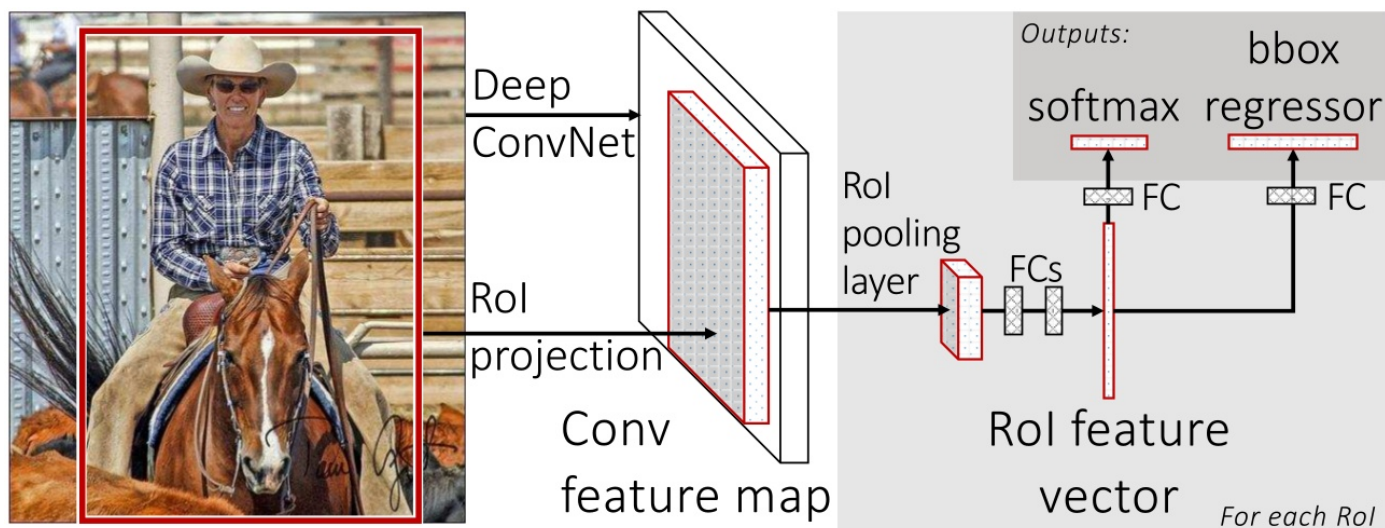
Lekce 7 [40]

Write down how is AP_{50} computed. [5]

1. Pro každou ze tříd narankujeme bounding boxy podle confidence levelu
2. V rámci jedné třídy postupujeme v pořadí z bodu (1) a kreslíme si precision/recall curve. Box bereme jako match pokud má $IoU > 0.5$
3. Upravíme křivku tak, aby byla monotónní
4. AP pro jednu třídu je průměrná precision v recallu 0, 0.1, 0.2, ..., 1.0, a AP pro celý dataset je průměr AP jednotlivých tříd



Considering a Fast-RCNN architecture, draw overall network architecture, explain what a RoI-pooling layer is, show how the network parametrizes bounding boxes and write down the loss. Finally, describe non-maximum suppression and how is the Fast-RCNN prediction performed. [10]



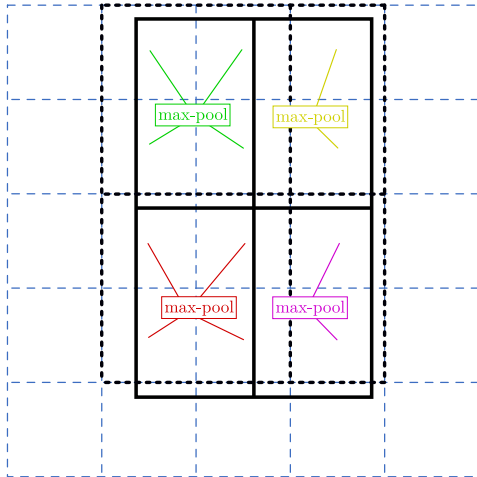
1. Začátek jako VGG, získáme 14x14 reprezentaci obrázku

2. Místo max pooling na vyrobíme 7x7 reprezentaci pomocí RoI pooling (viz níže)
3. Získám tím 7x7 reprezentaci každého RoI, kterou proženu FC a softmaxem (jako ve VGG) abych získal její classu (kterých je $K + 1$, abychom mohli říct "nic"), a bbox regresorem, abych získal její pozici (viz níže)

RoI Pooling

#

RoI rozdělím na 7x7 binů, každý z nich zaokrouhlím na původní 14x14 reprezentaci (viz obrázek níže) a jen z nich udělám max pooling (čímž jeden ze 7x7 binů).



Parametrizace bounding box

#

Pozice je dána relativně k RoI. Konkrétně

$$t_x = (x - x_r) / w_r, \quad t_y = (y - y_r) / h_r \\ t_w = \log(w / w_r), \quad t_h = \log(h / h_r)$$

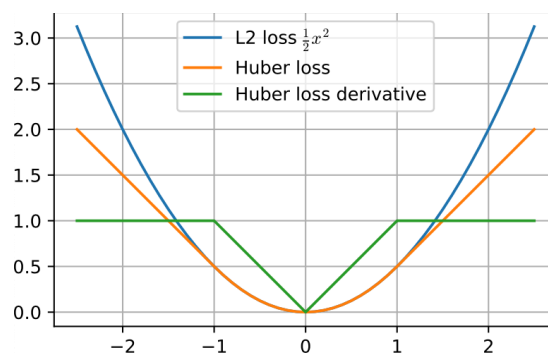
Logaritmy jsou ve w a h proto, že zmenšují range generovaných čísel (což je pro síť vždycky fajn).

Loss

#

Pro bbox se používá tzv. Huber loss, která se stará o gradient clipping.

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



Celková loss ještě počítá s loss za klasifikaci, s tím, že loss za bbox se počítá jen pro "opravdové" třídy (tedy nesnažíme se bbox dávat kolem třídy "nic"),

$$L(\hat{c}, \hat{t}, c, t) = L_{\text{cls}}(\hat{c}, c) + \lambda \cdot [c \geq 1] \cdot \sum_{i \in \{x, y, w, h\}} \text{smooth}_{L_1}(\hat{t}_i - t_i) \quad (2)$$

Inference

#

Non-maximum suppression se stará o to, aby se nám jeden objekt nezahlásil v několika různých RoI. Ignoruje RoI, které mají IoU nad nějakou hranicí s jiným RoI ze stejné třídy, který je lepší. Lepší jsou takové RoI, které mají vyšší pravost správné třídy.

Considering a Faster-RCNN architecture, describe the region proposal network (its architecture, what are anchors, what does the loss look like). [5]

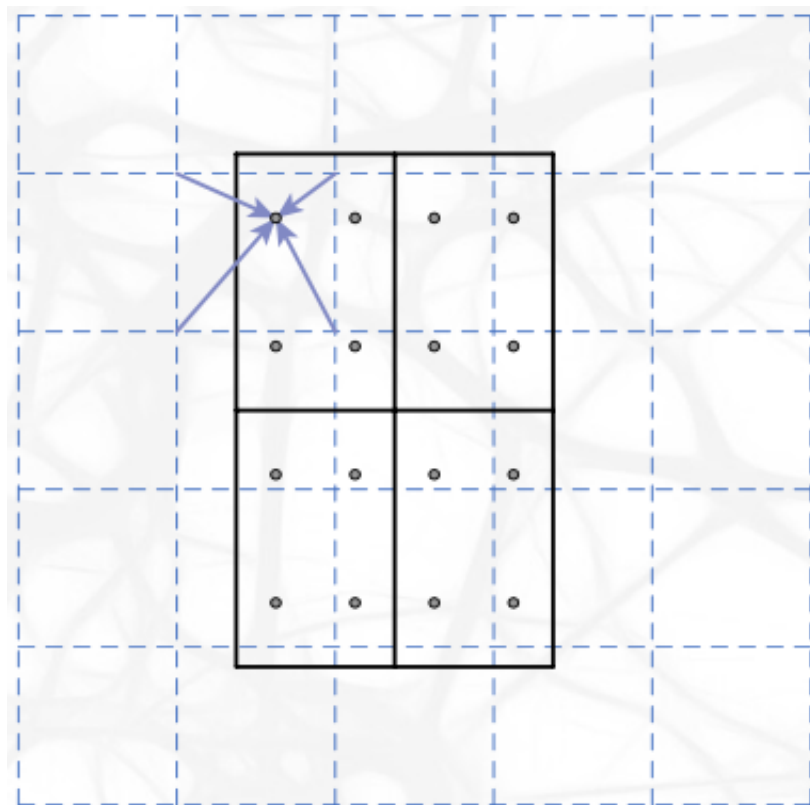
Na obrázek položíme různé druhy *anchorů* (tři různé velikosti a tři různé poměry stran). Poté vyzkoušíme všechny 3x3 okénka a na každém z nich predikujeme třídu (je/není tam objekt) a případně bbox toho objektu vzhledem ke všem možným anchorům, které jsou soustředné se současným 3x3 okénkem.

V průběhu trénování bereme anchor s největším překryvem + anchory s IoU > 70% jako pozitivní příklady (třída: objekt), anchory s IoU < 30% jako negativní příklady (třída: nic), zbytek anchorů ignorujeme.

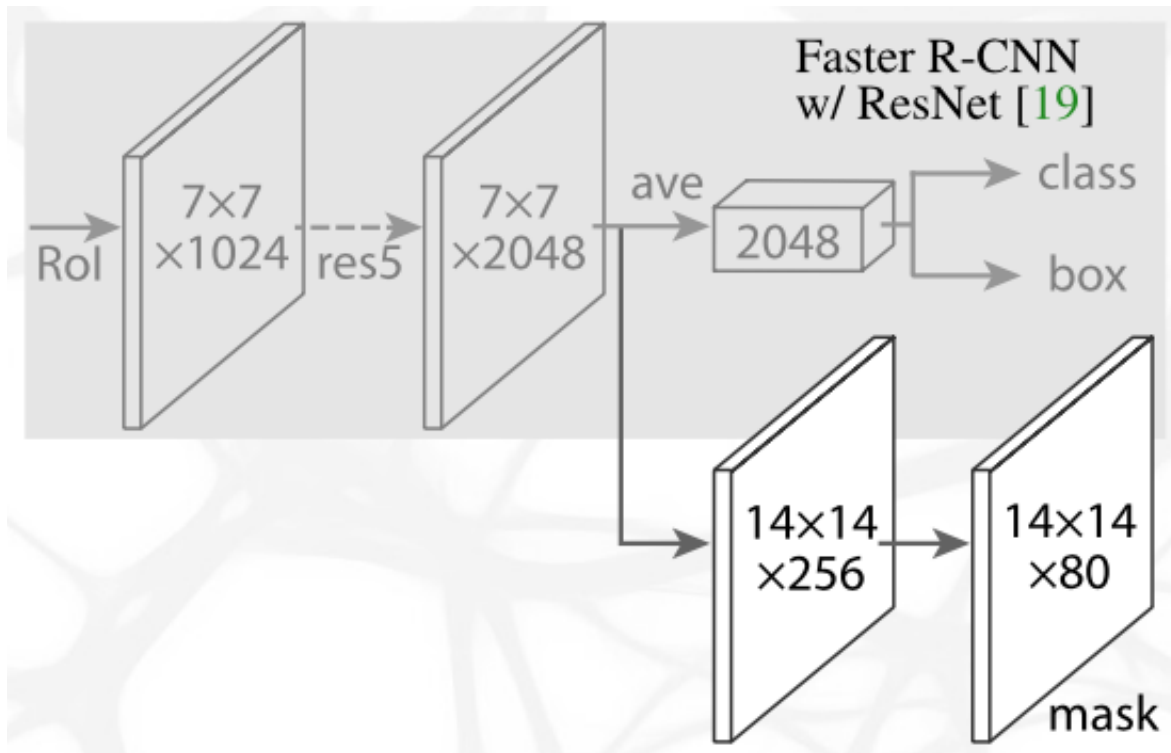
Během inference používám non-maximum suppression, abych vyhodil duplikované RoI.

Considering Mask-RCNN architecture, describe the additions to a Faster-RCNN architecture (the RoI-Align layer, the new mask-producing head). [5]

RoI Pooling je nahrazen RoI Align. Každý ze 7x7 binů si rozdělíme na 4 podbiny, a jejich hodnoty získáme bilineární interpolací hodnot z původní reprezentace 14x14. Tyto čtyři podbiny se zkombinují do finální hodnoty.



Pro vytvoření masky nejprve upscalujeme 7x7 reprezentaci zpět na 14x14, nebo 28x28. Poslední maskující konvoluce má tolik kanálů, kolik máme tříd, a masku tvoříme (a trénujeme) pro každou třídu zvlášť.



Write down the focal loss with class weighting, including the commonly used hyperparameter values. [5]

Úprava loss tak, aby se nám pozitivní příklady objektů neutopily v těch negativních.

$$\mathcal{L}_{\text{focal-loss}} = -(1 - p_{\text{model}}(y | x))^{\gamma} \cdot \log p_{\text{model}}(y | x) \quad (3)$$

Pro $\gamma = 0$ je tato loss prostě cross-entropy, pro vyšší γ vlastně snižujeme váhu loss u příkladů, u kterých si jsme hodně jistí výsledkem. Nejčastěji se používá $\gamma = 2$.

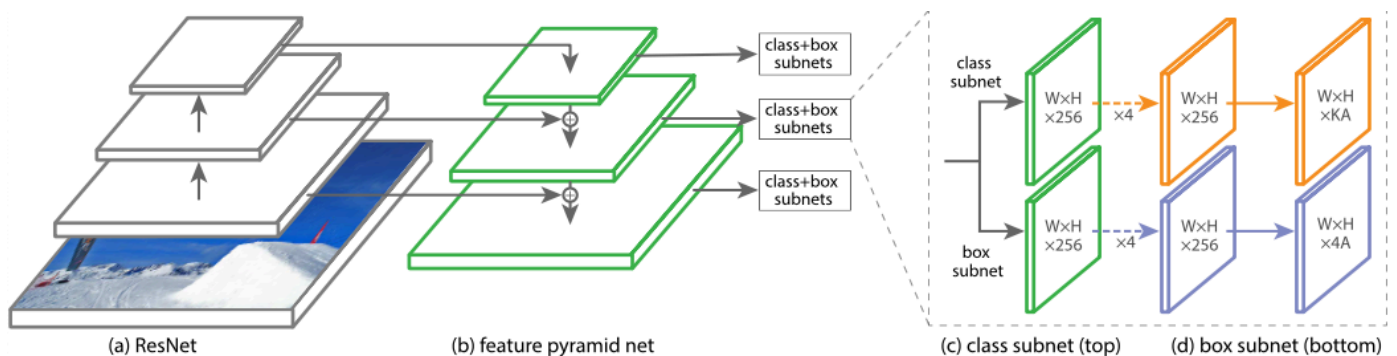
Navíc se ještě může každá třída různě navážit pevnou konstatnou,

$$-\alpha_y \cdot (1 - p_{\text{model}}(y | x))^{\gamma} \cdot \log p_{\text{model}}(y | x) \quad (4)$$

Pro vzácné třídy bývá nejčastěji používána hodnota $\alpha = 0.25$.

Draw the overall architecture of a RetinaNet architecture (the FPN architecture including the block combining feature maps of different resolutions; the classification and bounding box generation heads, including their output size). [5]

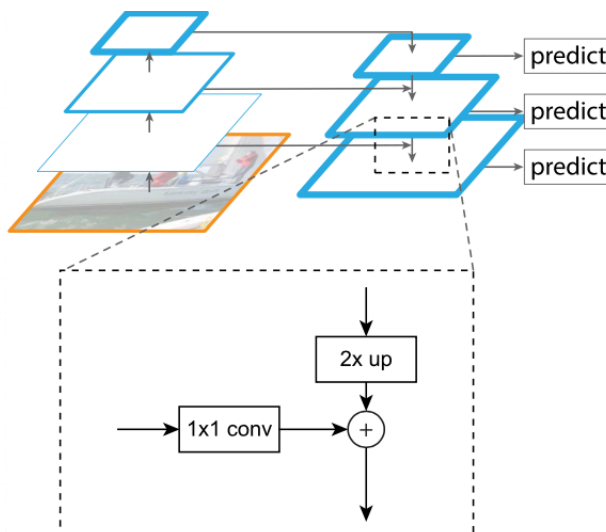
Oproti ResNetu mají navíc ještě C6 a C7, tj celkově dělají 7 max poolingů.



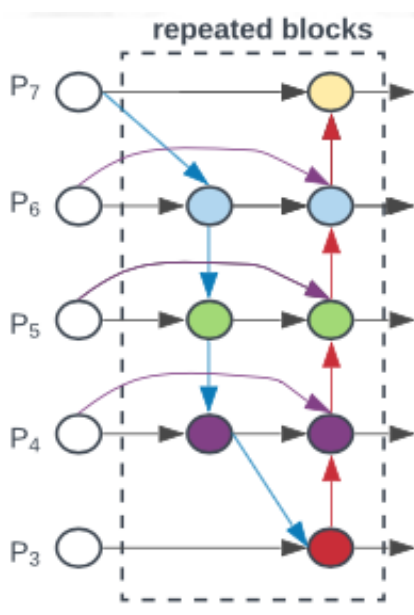
Klasifikační hlava má na výstupu $K \cdot A$ kanálů. Kolem každého "pixelu" výstupu máme A anchorů (většinou 9) a pro každý z nich potřebujeme říct pnost každé z K tříd. Klasifikace je tedy plně obstarána těmito konvolucemi, žádný pooling už nenásleduje.

Bounding boxová hlava má $4 \cdot A$ kanálů, pro každý anchor určuje hodnotu čtyřech parametrů.

V rámci ResNetu jsou mezi C vrstvami prostě max poolingy, v FPN probíhá jednoduchý 2x upscaling (doslova stávající hodnota zkopíruje na ta nová místa) a featurey zleva projdou 1x1 konvolucí, aby měly správný počet kanálů.



Draw the BiFPN block architecture, including the positions of all convolutions, BatchNorms and ReLUs. [5]



Součást EfficientDet. Všechny bloky jsou 3x3 separabilní konvoluce s BN a ReLU. Důležité jsou reziduální hrany, a dva chybějící bloky v prostředním sloupci, do kterých vedla jen jedna hrana a tak bylo možné je odstranit bez změny výsledku.

Downscaling je přes max pooling, upsampling jednoduše zkopírováním hodnot. Součet hodnot je navážený.