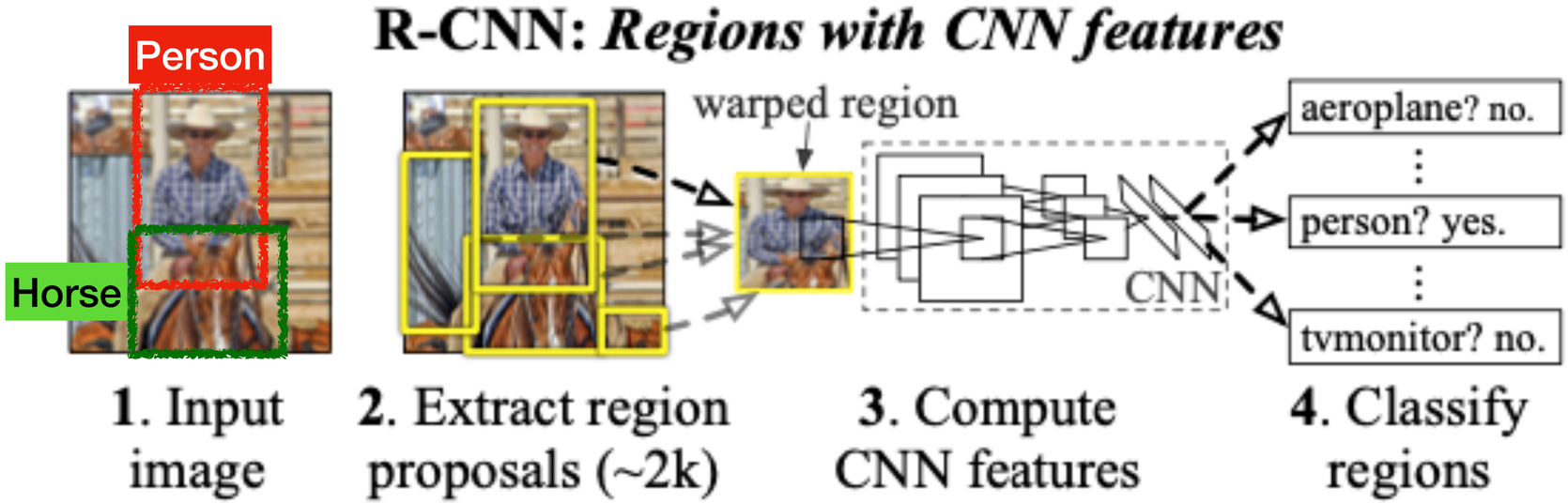
R-CNN

Rich feature hierarchies for accurate object detection and semantic segmentation

R-CNN

Region proposal + CNN



Region Proposal → **Deep CNN** → **SVM**

Region Proposal

localizing objects with a deep network

? localization as a regression problem

→ Szegedy에 의해 잘 작동하지 않음을 증명

? sliding-window detector

→ 5개의 convolutional layers 모델

기술적 문제 발생

Region Proposal

“recognition using regions”

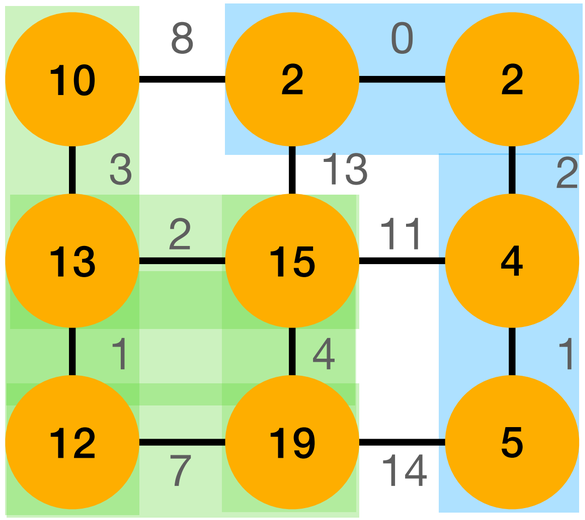
successful for both object detection and semantic segmentation

✓ “Selective Search”

Selective Search

Initial Segmentation

efficient graph-based image segmentation



Merging the Segmentations

hierachical grouping algorithm

$$s(r_i, r_j) = a_1 s_{colour}(r_i, r_j) + a_2 s_{texture}(r_i, r_j) + a_3 s_{size}(r_i, r_j) + a_4 s_{fill}(r_i, r_j),$$

`Color, Texture, Size, Fill` 가중합

픽셀의 유사도 계산

Warp

tightest square with context → B

object proposal을 CNN input size로
isotropically(등방적) 조정

tightest square without context → C

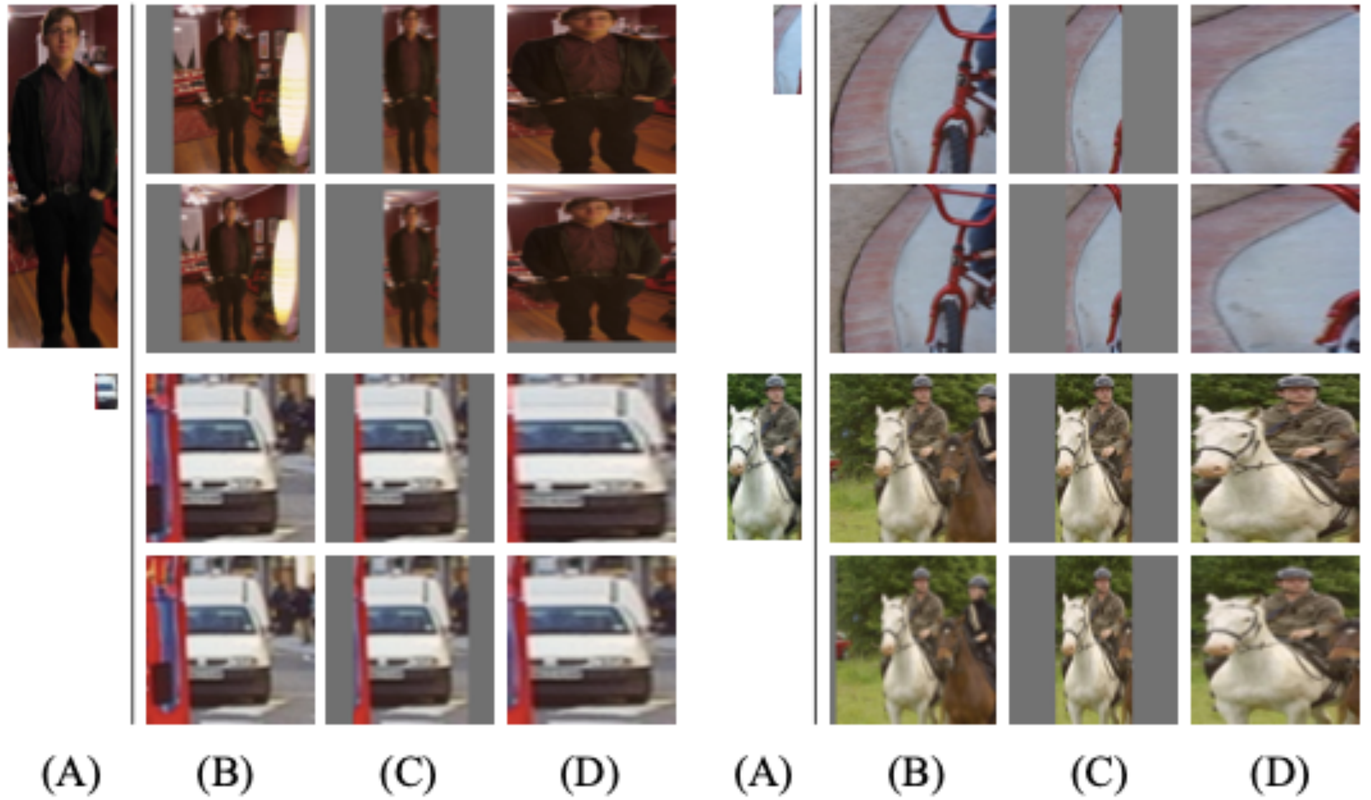
기존의 object proposal을 둘러싼
image content 제외

warp → D

object proposal을 CNN input size로
anisotropically(비등방적) 조정

additional image context 추가

context padding “p”



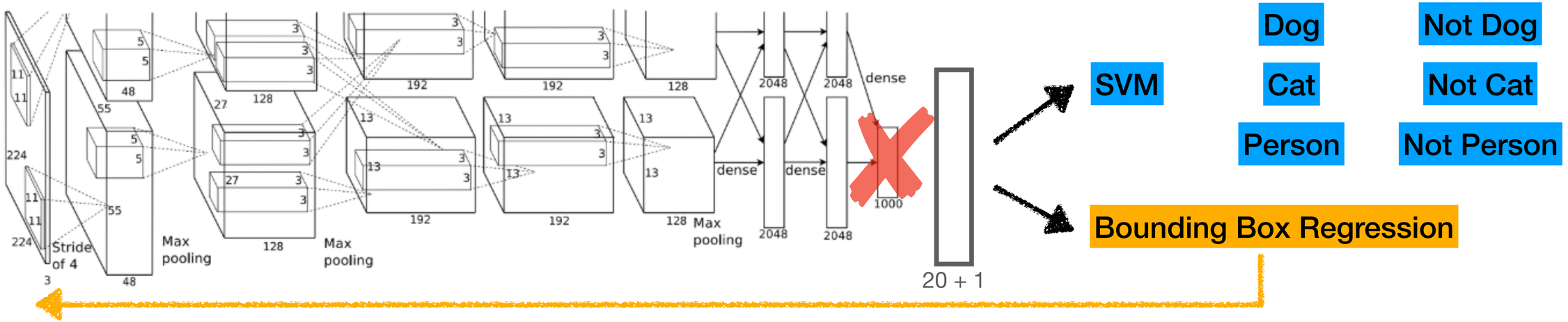
Top row: p = 0
Bottom row: p = 16

CNN

Softmax?

performance

54.2% → 50.9%



AlexNet

Supervised pre-training

new task (detection)

Domain-specific fine-tuning

pre-trained the CNN on a large dataset

(ILSVRC2012 classification)

using image-level annotations only

the CNN's ImageNet- specific 1000-way classification layer

→ randomly initialized (N + 1)-way classification layer

(N is the number of object classes, plus 1 for background)

CNN

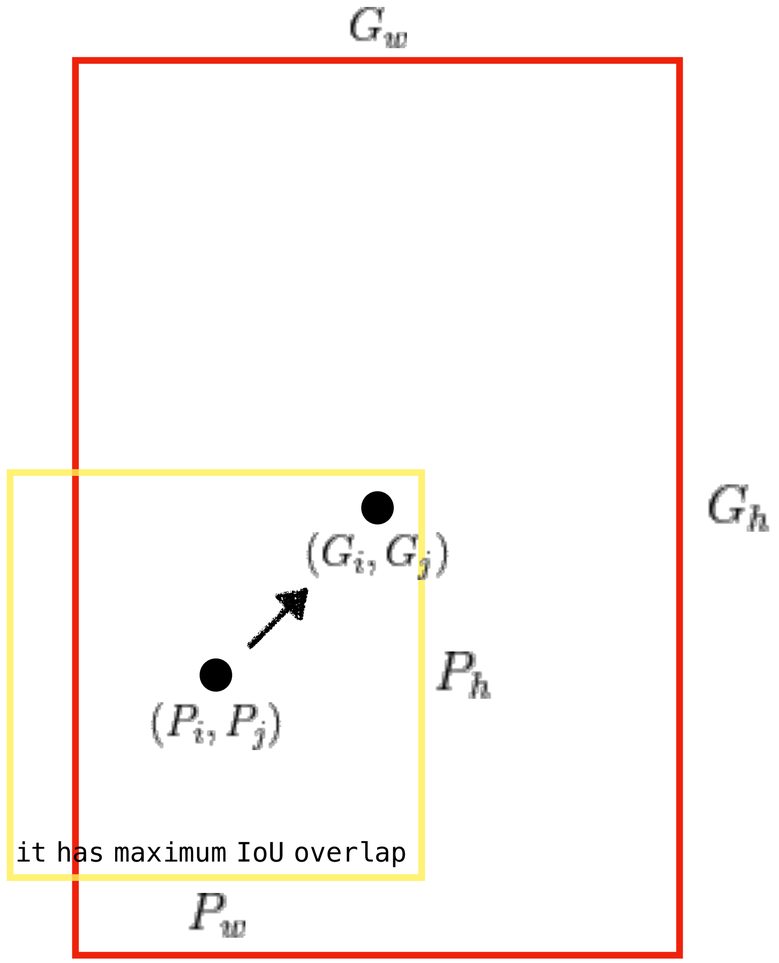
$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

region proposals → ≥ 0.5 IoU overlap with a ground-truth box

→ positives for that box's class

→ the rest as negatives → background

Bounding-Box Regression



$$P^i = (P^i_x, P^i_y, P^i_w, P^i_h) \longrightarrow G = (G_x, G_y, G_w, G_h)$$

$$\hat{G}_x = P_w d_x(P) + P_x \qquad t_x = (G_x - P_x)/P_w$$
$$\hat{G}_y = P_h d_y(P) + P_y \qquad t_y = (G_y - P_y)/P_h$$
$$\hat{G}_w = P_w \exp(d_w(P)) \qquad t_w = \log(G_w/P_w)$$
$$\hat{G}_h = P_h \exp(d_h(P)). \qquad t_h = \log(G_h/P_h).$$

$$d_*(P) = \hat{w}_*{}^T \phi_5(P^i)$$

$$\mathbf{w}_\star = \underset{\hat{\mathbf{w}}_\star}{\operatorname{argmin}} \sum_i^N (t^i_\star - \hat{\mathbf{w}}_\star^{\mathrm{T}} \phi_5(P^i))^2 + \lambda \|\hat{\mathbf{w}}_\star\|^2 .$$

$(y - \hat{y})^2$

Conclusion

bottom-up region proposals

in order to localize and segment objects,

high-capacity convolutional neural networks에 bottom-up region proposals 적용

training large CNNs

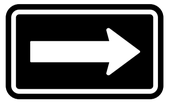
a paradigm for training large CNNs when labeled training data is scarce

pre-train the network

image classification

with abundant data

with supervision



fine-tune the network

detection

the target task

where data is scarce

Conclusion

computation time

2000개의 region proposal → 각각 CNN 수행

CNN 연산 X 2000 만큼의 수행 시간

end-to-end learning

CNN, SVM, Bounding Box Regression(3가지의 model) → multi-stage pipelines

SVM, Bounding Box Regression 에서 학습한 결과로 CNN 업데이트 불가

$$Int(C) = maxw(e)$$

$$Dif(C_i, C_j) = min\ w(v_i, v_j)$$

$$D(C_i, C_j) = \begin{cases} true & Dif(C_i, C_j) > min(Int(C_i), Int(C_j)) \\ false & otherwise \end{cases}$$

$$Dif(C_i, C_j) > min(Int(C_i), Int(C_j)) \quad \longrightarrow \quad Dif(C_i, C_j) > min(Int(C_i) + \frac{k}{|C_i|}, Int(C_j) + \frac{k}{|C_j|})$$

$$Dif(C_i, C_j) = false$$

$$if\ \ Dif(C_i, C_j) \leq min(Int(C_i) + \frac{k}{|C_i|}, Int(C_j) + \frac{k}{|C_j|})$$