Module 4: 2D Object Detection

内容
- Define the 2D object detection problem.
- Apply ConvNets to 2D object detection.
- Challenges of object detection.
- 2D object tracking problem.

Lesson 1: The Object Detection Problem

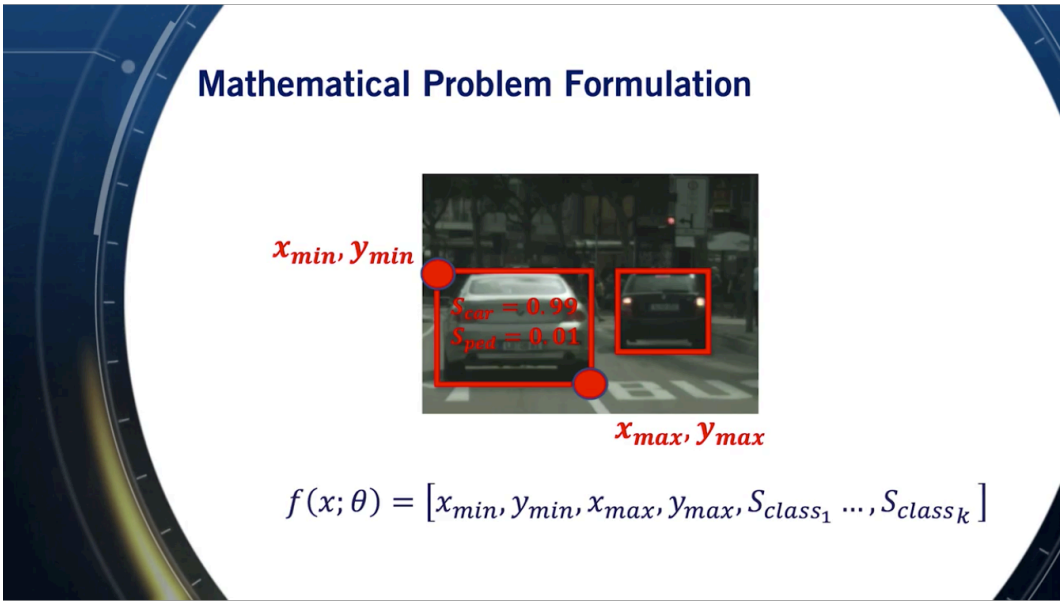
単語
- extent: The extent of something is its length, area, or size.
- hallucinate: If you hallucinate, you see things that are not really there, either because you are ill or because you have taken a drug.
- precipitation: precipitation is rain, snow or hail.

内容
- The 2D object detection task problem formulation.
- Determine how good a 2D object detector is through evaluating performance measures.

Brief History of Object Detection
- 2001 - Viola, Jones - Viola Jones Object Detection Framework.
  - face detection.
- 2005 - Dalal, Triggs - Histogram of Oriented Gradients. これはSIFTも使っている方法?
  - pedestrian detection.
- 2012 - Krizhevsky, Sutskever, Hinton - Alexnet.
  - in 2012, it was the only deep learning-based entry in the challenge.

The Object Detection Problem
Given a 2D image input, to estimate the location defined by a bounding box and the class of all objects in the scene.
- For self-driving cars, we usually are most interested in object classes that are dynamic.
- These include vehicles and their subclasses, pedestrians, and cyclists.

Object Detection Is Not Trivial!
- Extent of objects is not fully observed!
  - Occlusion: Background objects covered by foreground objects.

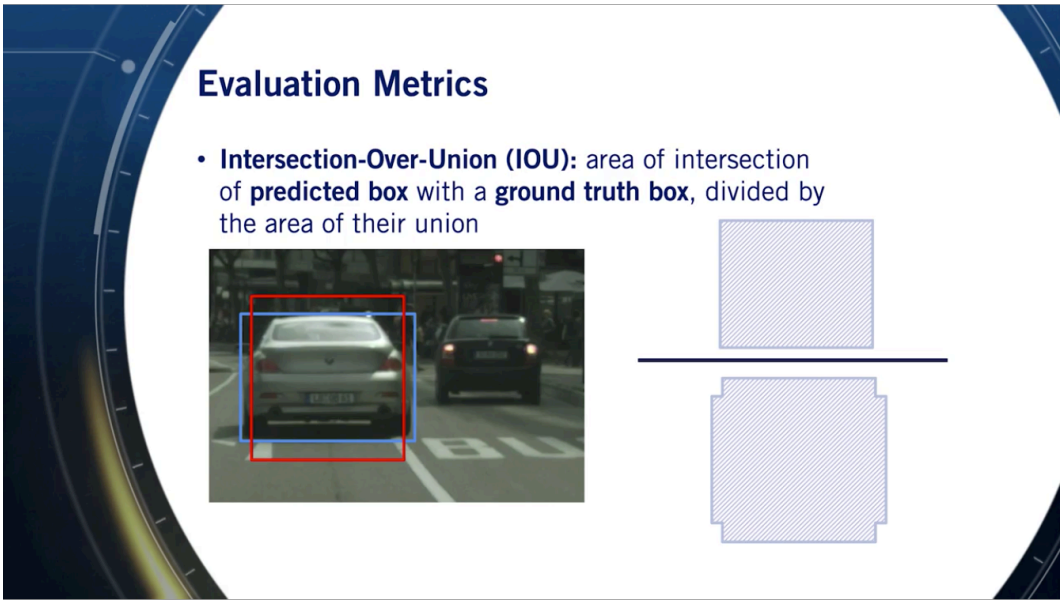


    - This requires the algorithm to be able to hallucinate the extent of objects to properly detect them.
  - Truncation: Objects are out of image boundaries.
    - This phenomenon creates huge variability in the bounding box sizes.
- Scale: Object size gets smaller as the object moves farther away.
  - The algorithm is expected to determine the class of these objects at variable scales.
- Illumination Changes:
  - whole image illumination variations from bright sun to night driving.
  - partial variations due to reflections, shadows, precipitation, and other nuisance effects.
  - Too bright.
  - Too dark.

Mathematical Problem Formulation

$f(x;\theta) = [x_{min}, y_{min}, x_{max}, y_{max}, S_{class_1}, \ldots, S_{class_k}]$.

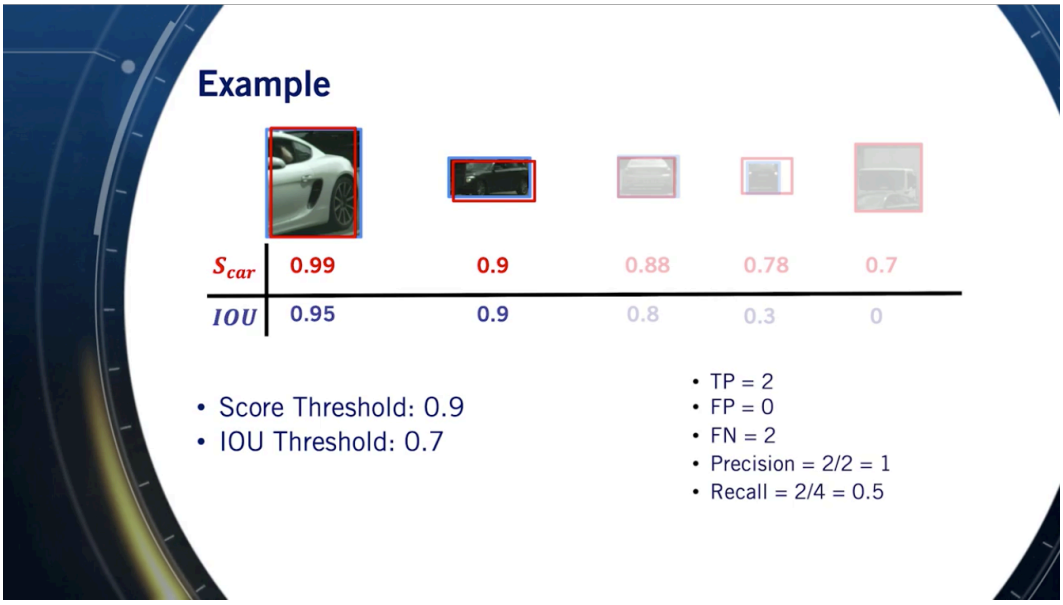
- x_{min}, y_{min}: coordinates of the top-left of the box.
- x_{max}, y_{max}: coordinates of the lower right corner of the box.
- $S_{class_1}, \ldots, S_{class_k}$: class scores.
- k: the number of classes of interest.



Evaluation Metrics

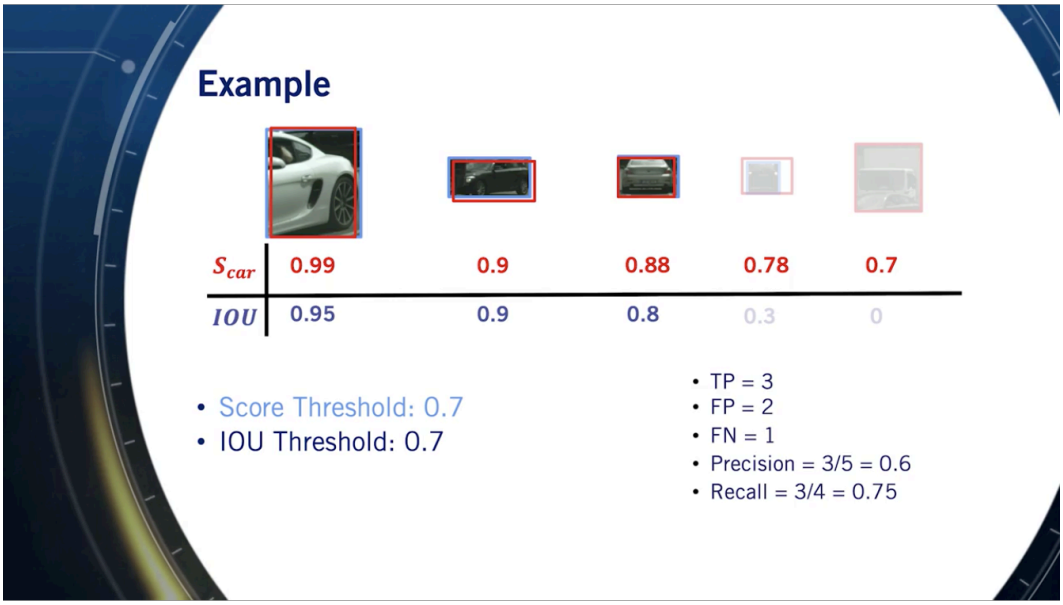
- Intersection-Over-Union (IOU): area of intersection of predicted box with a ground truth box, divided by the area of their union.
  - $x_{min}, y_{min}, x_{max}, y_{max}$を評価する。
- True Positive (TP): Object class score > score threshold, and IOU > IOU threshold.
- False Positive (FP): Object class score > score threshold, and IOU < IOU threshold.
- False Negative (FN): Number of ground truth objects not detected by the algorithm.
- Precision: $\frac{TP}{TP + FP}$.
- Recall: $\frac{TP}{TP + FN}$.
  - divided by the total number of ground truth objects.
- Precision Recall Curve (PR-Curve): Use multiple object class score thresholds to compute precision and recall.
  - Plot the values with precision on y-axis, and recall on x-axis.

- Average Precision (AP): Area under PR-Curve for a single class.
  - Usually approximated using 11 recall points. （下記のExampleでわかる）



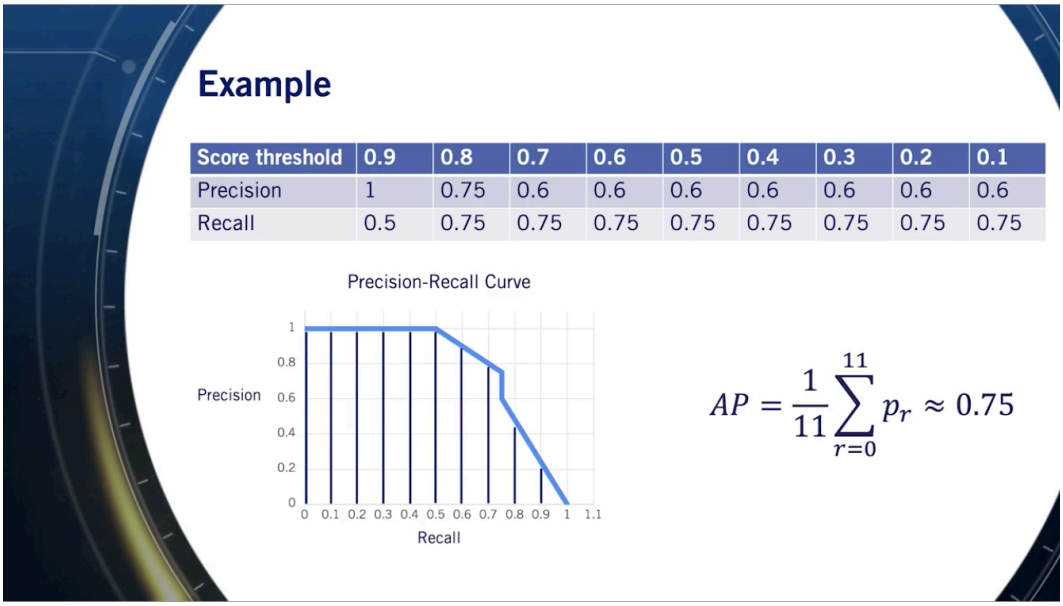
Example: Car Detector（上記の図を参考）（大事）

- $FN = 2$: 3番目、4番目.
- $TP = 2$: 1番目、2番目.
- 5番目は誤検知、このground truth object（car）は存在しない。これはtruck。
  - 明らかに誤検知は評価していないでしょう。いいえ、違う。S_{car}はまだ低いから。もし5番目の誤検知のS_{car}がScore Threshold以上であれば、FPに入るので、評価される。
  - なぜ先にScoreを見るの?
- The detector in this case is a high precision, low recall detector.
  - This means that the detector misses some objects in the scene, but when it does detect an object, it makes very few mistakes in category classification and bounding box location.



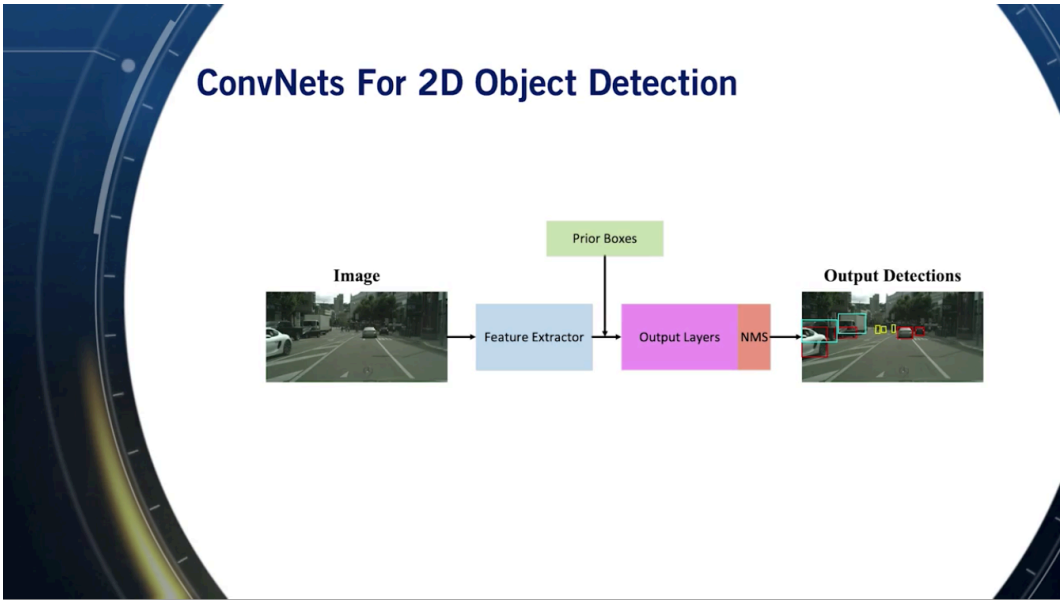
Example: Car Detector（同じ例、違うScore Threshold）（大事）

- $FP = 2$: 4番目、5番目.
- $FN = 1$: 4番目.
- We can conclude that the effect of lowering the score threshold is less accurate detection results at the expense of detecting more objects in the scene.



Example: Precision-Recall Curve

- Note that we also add the precision, recall points of $(1,0)$ as the first point in the plot, and $(0,1)$ as the final point in the plot.
- APの意味：The value of the average precision of the detector can be thought of as an average of performance over all score thresholds, allowing objective comparison of the performance of detectors without having to consider the exact score threshold that generated those detections.
  - しかしIOU thresholdは固定でしょう。でも常に先にScoreを見るので。



- 2D object detection comprises localizing an object and determining what the object is.
- 2D object detectors are evaluated using the Average Precision metric, at a specific IOU threshold.
- Implementation Resources from Google: https://github.com/tensorflow/models/tree/master/research/object_detection

Lesson 2: 2D Object Detection with Convolutional Neural Networks

単語
- intricacy: Intricacy is the state of being made up of many small parts or details.
- eke out: If you eke out something, for example, a victory, you obtain it with difficulty.
- inception: The inception of an institution or activity is the start of it.

内容
- Build standard single stage (??) architecture for 2D object detection.
- Common neural network design choices for performing 2D object detection using the proposed architecture.

ConvNets For 2D Object Detection
- The architecture takes as an input an image and a set of manually crafted prior boxes.
  - 僕の理解だと、prior boxesはbounding boxの可能なサイズ? つまりbounding boxのサイズは自動的に推測した物ではない?
- a set of (??) fully connected layers take as an input, the output of the feature extractor and provide a location refinement of each 2D prior box (??) as well as a classification. まだ分かっていない。
- non maximum suppression is performed on the output of the fully connected layers to generate the final detections.

The Feature Extractor
- Feature extractors are the most computationally expensive component of the 2D object detector.
- The output of feature extractors usually has much lower width and height than those of the input image, but much greater depth.
  - the output's depth is usually two to three orders of magnitude greater than that of the input image.
- Very active area of research, with new extractors proposed on regular basis.

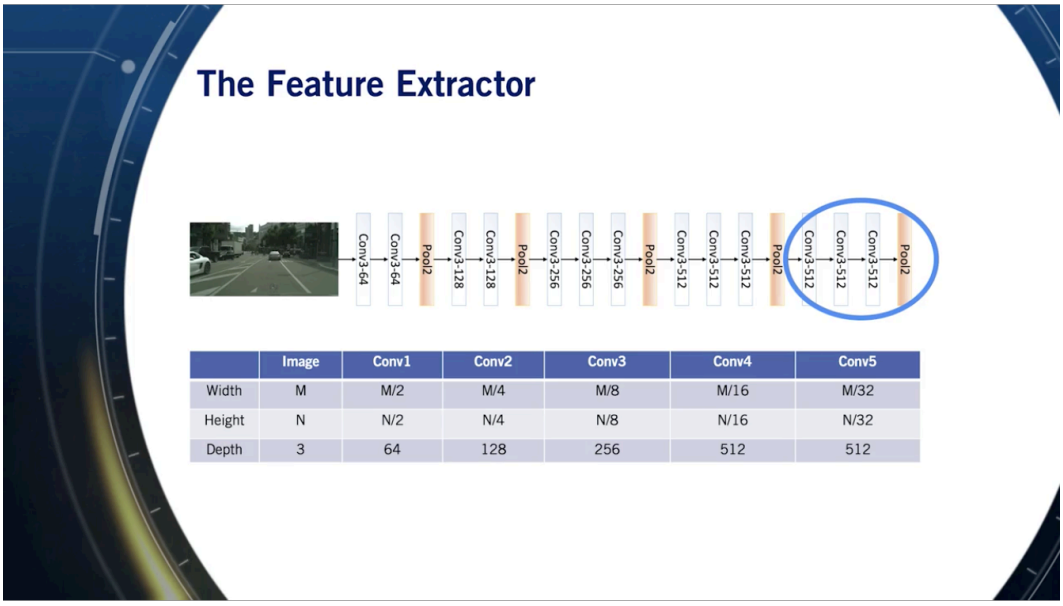
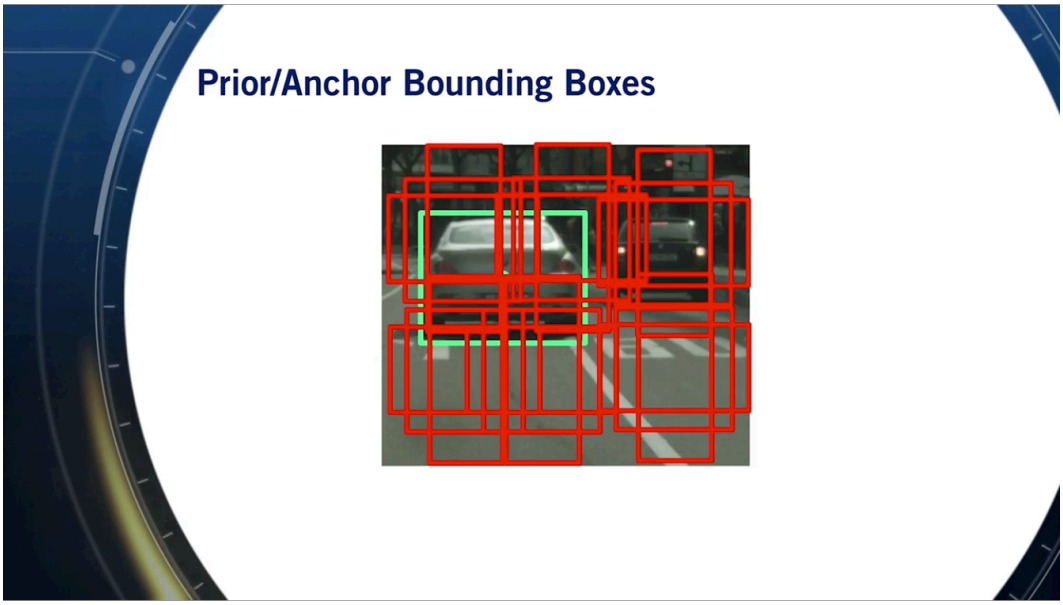


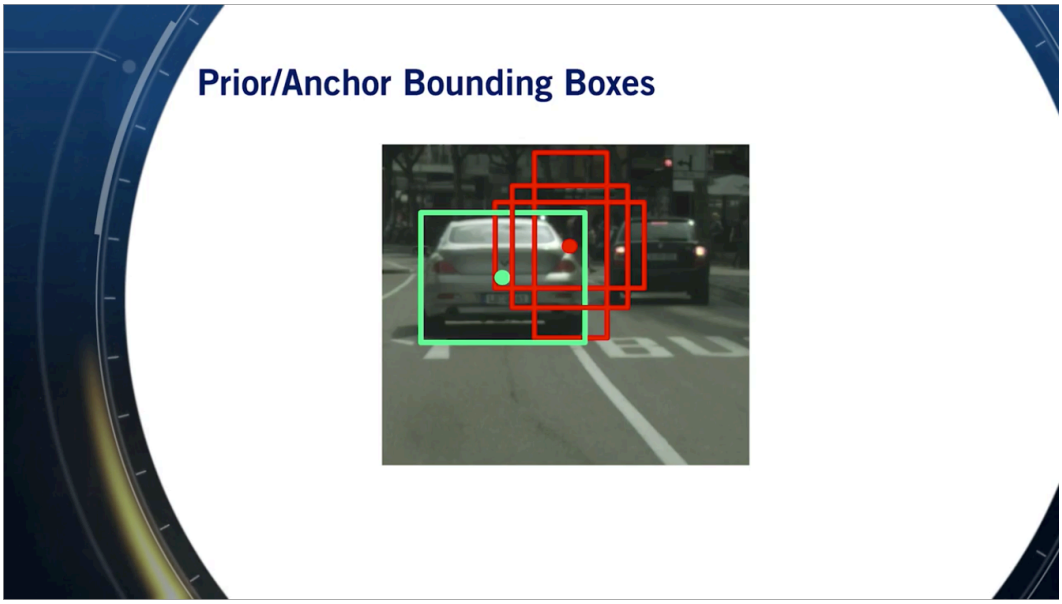
	Image	Conv1	Conv2	Conv3	Conv4	Conv5
Width	M	M/2	M/4	M/8	M/16	M/32
Height	N	N/2	N/4	N/8	N/16	N/32
Depth	3	64	128	256	512	512

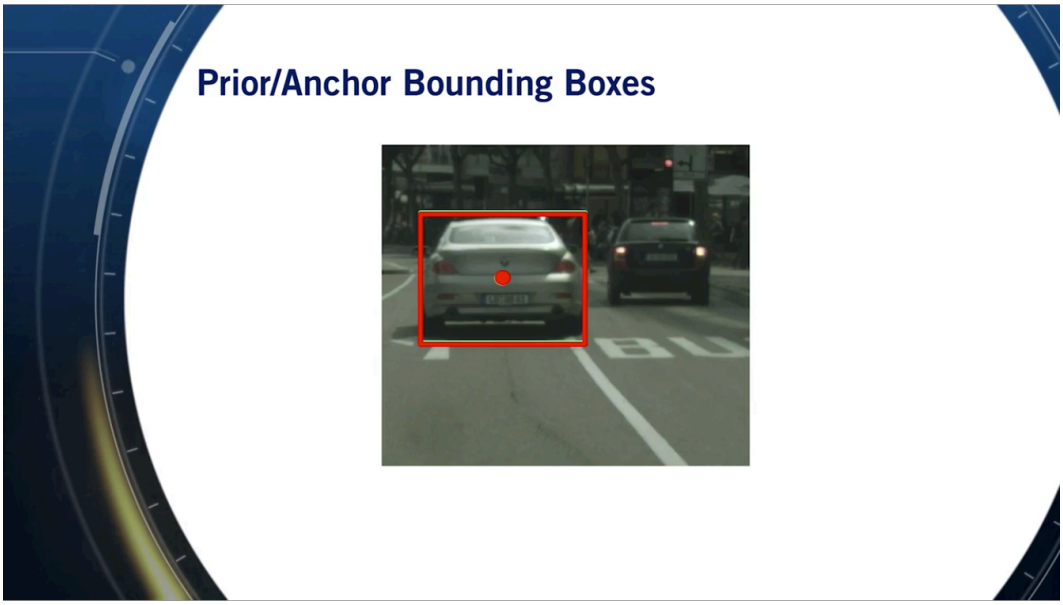
- Most common extractors are: VGG, ResNet, and Inception.

VGG Feature Extractor
- Alternating convolutional and pooling layers.
- All convolutional layers are of size $3 \times 3 \times K$, with stride 1 and 1 zero-padding.
  - つまりK個filter。channelの数は特に書いていない。勿論input imageのchannel数によってkernelのchannelも一致する。
  - $W_{out} = \frac{W_{in} - m + 2 \times P}{S} + 1 = \frac{W_{in} - 3 + 2 \times 1}{1} + 1 = W_{in}.$
  - $H_{out} = \frac{H_{in} - m + 2 \times P}{S} + 1 = \frac{H_{in} - 3 + 2 \times 1}{1} + 1 = H_{in}.$
  - $D_{out} = K.$
- All pooling layers use the max function, and are of size 2×2, with stride 2 and no padding.
  - $W_{out} = \frac{W_{in} - n}{S} + 1 = \frac{W_{in} - 2}{2} + 1 = \frac{W_{in}}{2}.$
  - $H_{out} = \frac{H_{in} - n}{S} + 1 = \frac{H_{in} - 2}{2} + 1 = \frac{H_{in}}{2}.$
  - $D_{out} = D_{in}.$
- These particular hyperparameters were arrived at through intensive experimentation and have performed extremely well in a wide range of problems, making VGG an extremely popular extractor.
- VGGのoutput size: $\frac{Width}{32} \times \frac{Height}{32} \times 512.$
- 例：input size: $1280 \times 960 \times 3$. output size: $40 \times 30 \times 512$.



Prior/Anchor Bounding Boxes
- To generate 2D bounding boxes, we usually do not start from scratch and estimate the corners of the bounding box without any prior.
- We assume that we do have a prior on where the boxes are in image space and how large these boxes should be.
- Anchor boxes are manually defined over the whole image usually on an equally-spaced grid.

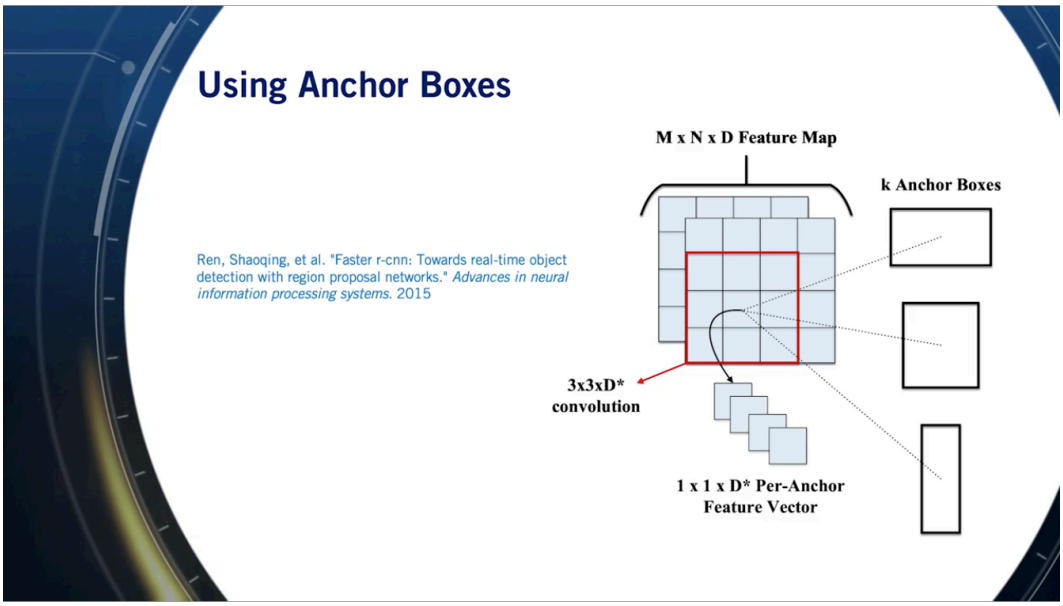




Prior/Anchor Bounding Boxes

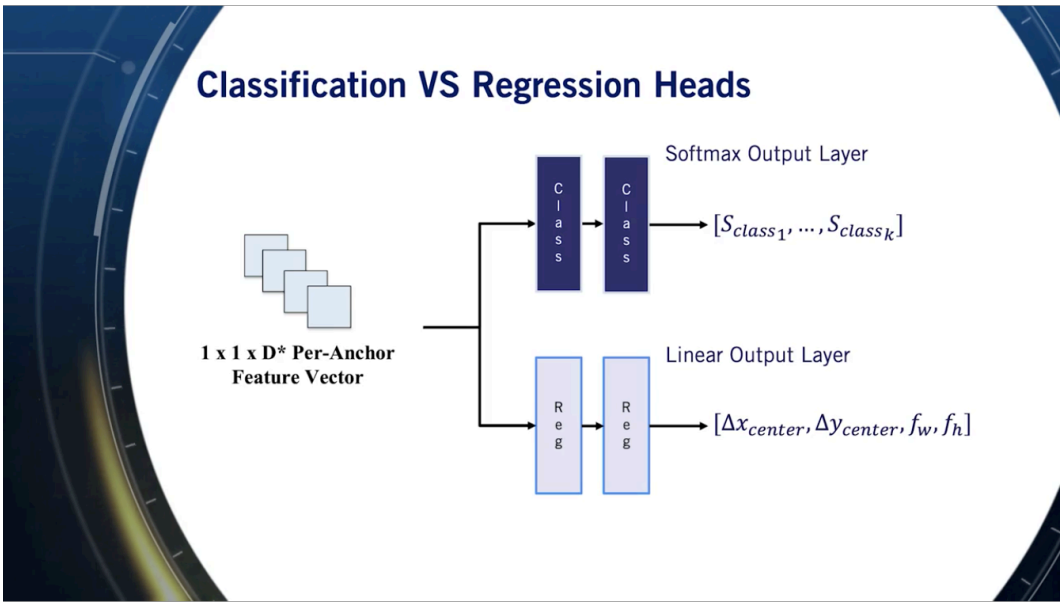
- Let's assume that we have a set of anchors close to our ground-truth boxes.
  - 全部移動するの?
- During training, the network learns to take each of these anchors and tries to move it as close as possible to the ground-truth bounding box in both the centroid location and box dimensions.
  - "box dimensions"の意味は長さも合わせること?
  - なぜ"each of these anchors"? 何かrobustnessの考慮?
- This is termed residual learning and it takes advantage of the notion that it is easier to nudge an existing box a small amount to improve it rather than to search the entire image for possible object locations.（大事）
  - In practice, residual learning has proven to provide much better results than attempting to directly estimate bounding boxes without any prior.

- Many different methods have been proposed in the literature on how to use the anchor bounding boxes to generate the final prediction.
- Usually, the anchors interact with the feature map to generate fixed sized feature vectors for every anchor. まだ分かっていない。



Using Anchor Boxes（[2015] Shaoqing Ren, et al. *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*の方法）（https://arxiv.org/abs/1506.01497）（大事）

- $M \times N \times D$ Feature Map.
  - M, Nはfeature mapのサイズ！ input imageのサイズじゃない!
- For every pixel in the feature map, we associate k anchor boxes.
  - k anchor boxes.
  - kは固定？ 多分固定。
- Then perform a $3 \times 3 \times D^*$ convolution operation on that pixel's neighborhood.
  - $3 \times 3 \times D^*$ convolutionの意味はまだ分かっていない。普通に考えればいいでしょう。
- This results in a $1 \times 1 \times D^*$ feature vector for that pixel.



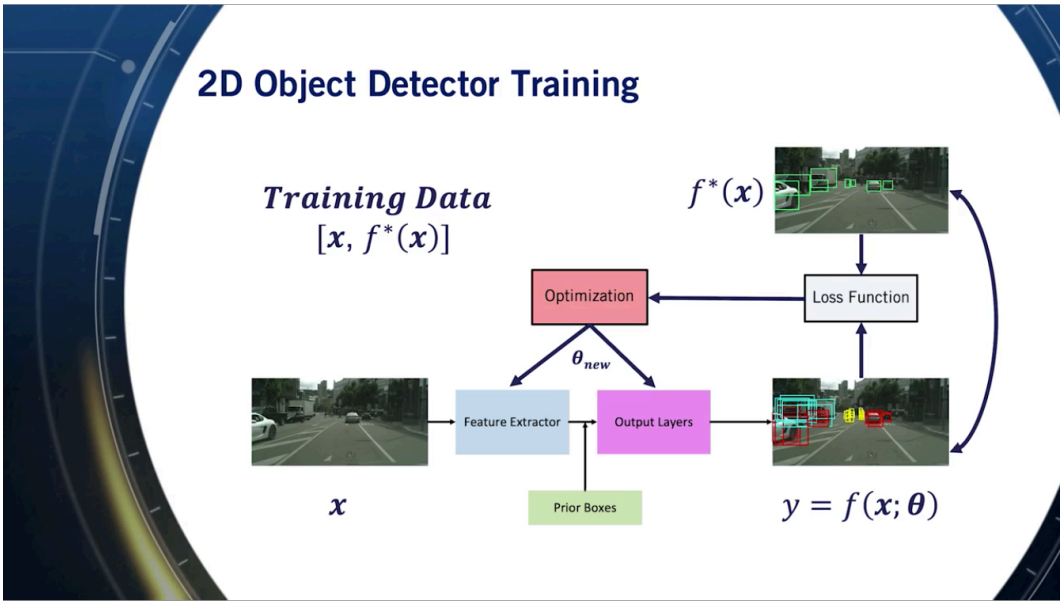
- We use this $1 \times 1 \times D^*$ feature vector as the feature vector of every one of the k anchors associated with that pixel.
- つまり1つpixelはk個anchorsとk個$1 \times 1 \times D^*$ feature vectorを持つ。
  - これらのfeature vectorsは全部同じ。
- つまり出力は$M \times N$個pixel、k固定だったら
  - $M \times N \times k$個$1 \times 1 \times D^*$ feature vectors。

Output Layers
- We then proceed to feed the extracted feature vector to the output layers in the neural network.

Classification VS Regression Heads（Output Layerの話）（大事大事）
- The output layers of a 2D object detector usually comprise of a regression head and a classification head.
- The regression head usually includes multiple fully-connected hidden layers with a linear output layer.
  - The regressed output is typically a vector of residuals that need to be added to the anchor at hand to get the ground truth bounding box.
    - $[\Delta x_{min}, \Delta y_{min}, \Delta x_{max}, \Delta y_{max}]$.
    - つまり、anchor boxesのサイズはまだ調整していない。まだpriorのまま。
    - このresidualを使ってbounding boxのサイズを調整するんだ!
    - つまりこの方法はanchor boxのcentroidを気にしなく、ただanchor boxの4つ点を移動するんだ。
  - Another method to update the dimension of the anchors is to regress a residual from the center of the anchor to the center of the ground truth bounding box in addition to two scale factors that correct the ground truth bounding box width and height when multiplied with an anchor's width and height.
    - $[\Delta x_{center}, \Delta y_{center}, f_w, f_h]$.
    - 上の方法はanchor boxの4つ頂点を移動する。この方法はanchor boxの中心点を移動して、また高さと広さをscaleする。
- The classification head is also comprised of multiple fully-connected hidden layers, but with a final softmax output layer. 分類だから。
  - The softmax output is a vector with a single score per class.
  - The highest score usually defines the class of the anchor at hand.
- Redundant Bounding Boxesの解決策（大事）（まだ終わっていない、次のLessonへ）
  - We will end up with k bounding boxes per pixel of the output feature map.
  - Even if we consider boxes with high classification score for all classes of interest, we still will have many redundant detections in the image. わかる!
    - bounding boxesが多すぎるので、ほとんど重なって、高い点数を得るのはよくあるからでしょう。
  - During average precision computation, we only consider one detection per ground truth bounding box.
  - Any redundant detections are considered false positives and results in lower precision and thus a lower average precision overall.
  - In fact, this is only an issue of inference and it turns out to be beneficial to drive the network to output the best possible residuals for every anchor during training.

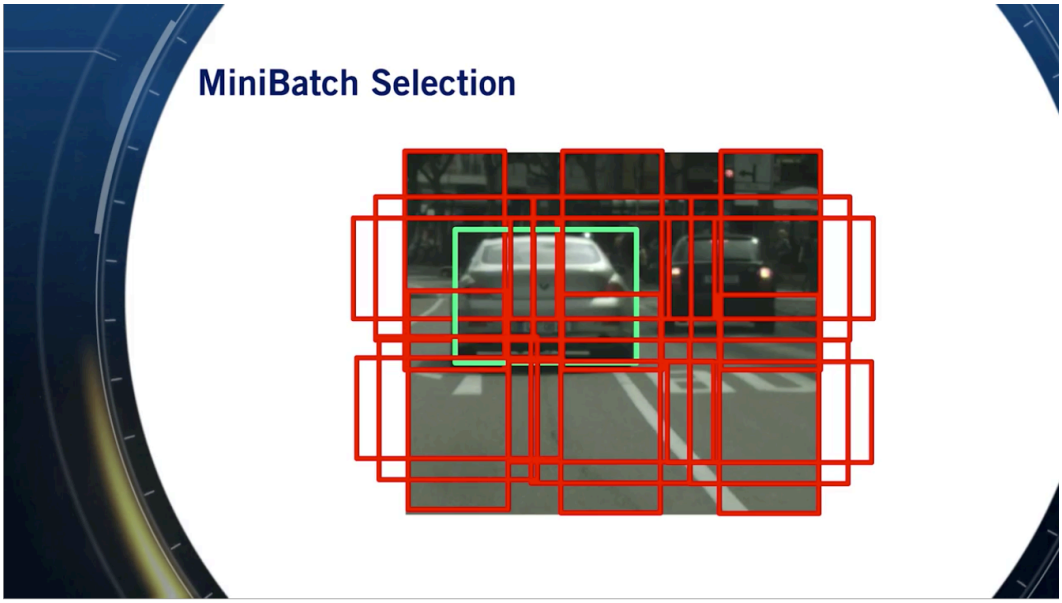
- 2D object detectors can be performed using convolutional neural networks.
- Usually, anchor boxes are used as priors for the neural network to shift around to achieve object classification and localization.

Lesson 3: Training vs. Inference

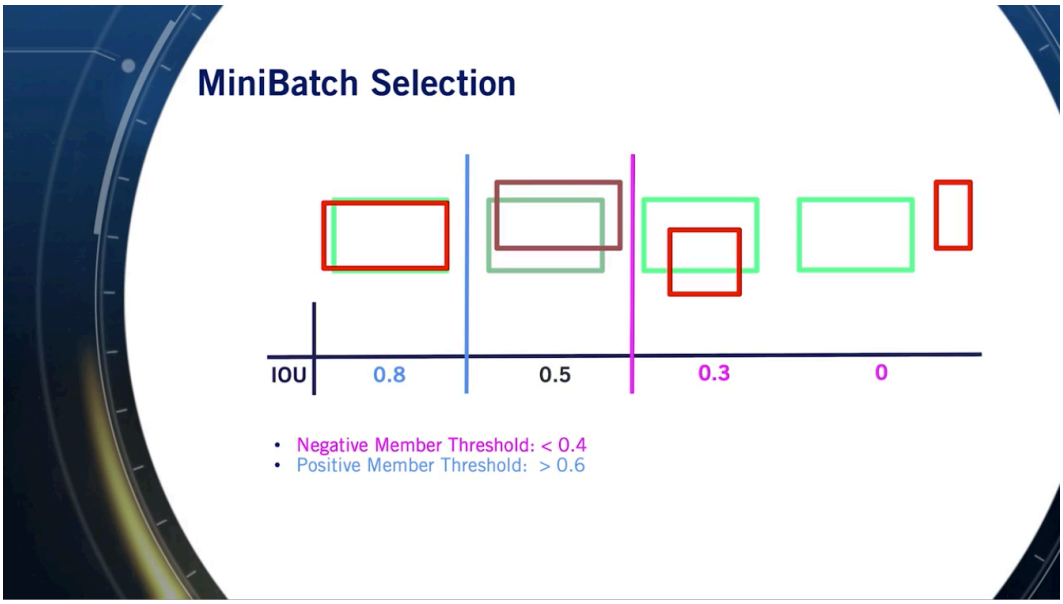


内容
- Handle multiple detections per object during training through mini batch selection.
- Handle multiple detections per object during inference through non-maximum suppression (NMS).

2D Object Detector Training
- If $f^*(x)$ and $f(x;\theta)$ are one to one, our problem would have been easy.
- However, the outputs of our network is multiple boxes that can be associated with a single ground truth box.



MiniBatch Selection（大事）
- VGG feature extractor reduces the resolution of the initial input by a factor of 32.



- That means that if we associate every pixel in the feature map with a set of anchors, these anchors will be transferred to the initial image by placing them on a grid with stride 32.
  - feature mapの1つpixelはinitial imageの32×32 grid相当だから。
- Visualize the ground-truth bounding box alongside these anchors.
  - Some anchors overlap and some don't.（上記の緑枠や赤枠の図）
  - 上記赤枠の数は$M \times N \times k$でしょう！M, Nはfeature mapのサイズ！input imageのサイズじゃない!
- Any anchor with an IOU greater than the positive anchor threshold is called a positive anchor.
- Any anchor with an IOU less than the negative anchor threshold is called a negative anchor.
- Any anchor with an IOU in between the two thresholds is fully discarded. すごい！（緑枠と赤枠以外の色）
- Negative anchors target
  - Classification: Background.
    - Background is usually a class we add to our classes of interest to describe anything not included in these classes.
  - Regression: None.
- Positive anchors target
  - Classification: Category of the ground-truth bounding box.
  - Regression: Align box parameters with highest IOU ground-truth bounding box.
- Problem: Majority of anchors are negative results in neural network will label all detections as background.（大事）
  - The proposed IOU thresholding mechanism results in most of the regressed anchors being negative anchors. 40%以下重なっているanchorが多いでしょう？まだ移動できないし、scaleできないし。ほとんど重なっていないanchorも多いでしょう!
  - When training with all these anchors, the network will be observing far more negative than positive examples leading to a bias towards the negative class.
- Solution: Sample a chosen minibatch size, with 3:1 ratio of negative to positive anchors to eliminate bias towards the negative class.
- Choose negatives with highest classification loss (online hard negative mining) to be included in the minibatch.
  - つまりIOUが低いanchorsの中にhighest確率で誤分類したnegativesを選択する?
  - This means we will be training to fix the biggest errors in negative classification.

- Example1: A minibatch of 64 examples, the negative minibatch will be the 48 negative examples with the highest classification loss, and the 16 remaining anchors will be positive anchors.

Classification Loss

$$L_{cls} = \frac{1}{N_{total}} \sum_i CrossEntropy(s_i^*, s_i).$$

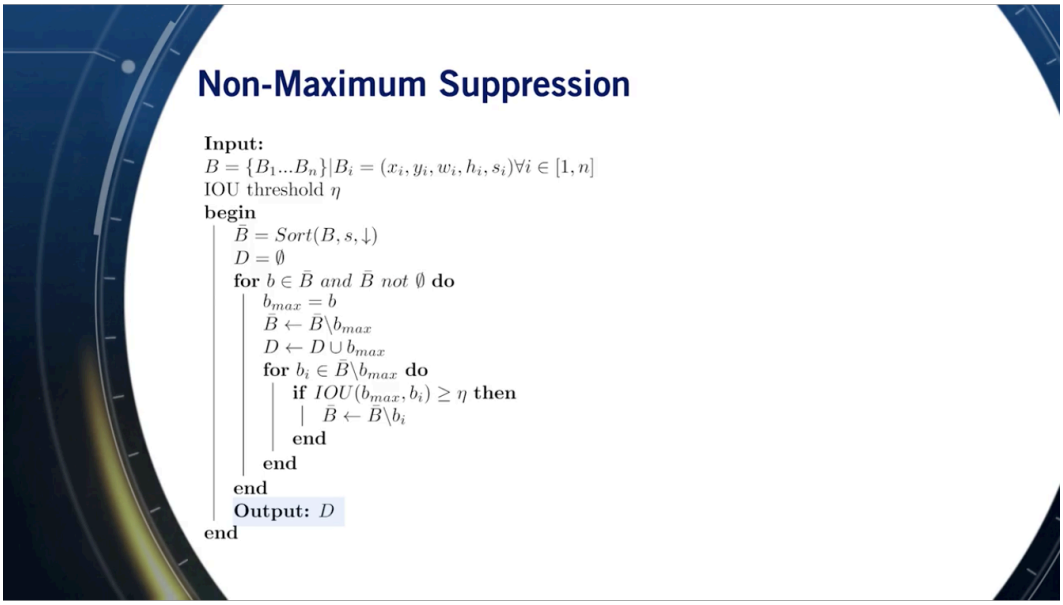
- The total classification loss is the average of the cross-entropy loss of all anchors in the minibatch.
- N_{total} is the size of our minibatch.
- s_i is the output of the neural network.
- s_i^* is the anchor classification target
  - Background if anchor is negative.
  - Ground truth box class if anchor is positive.
- 補足：CrossEntropy Loss Function: $L(\theta) = -log\left(Softmax(z_i)\right) = -z_i + log\sum_j exp(z_j)$.

Regression Loss（大事）

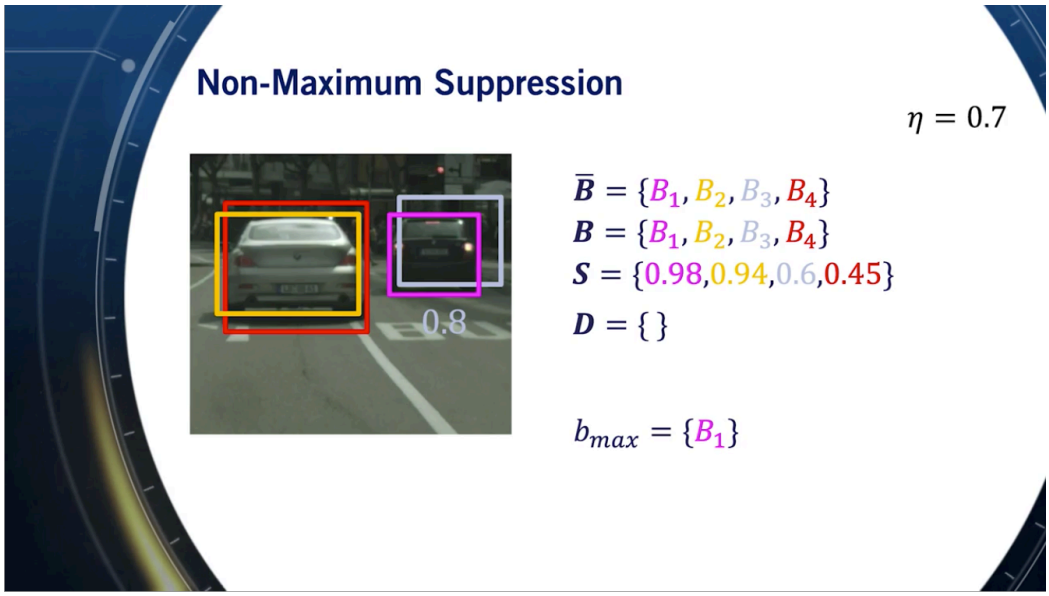
$$L_{reg} = \frac{1}{N_p} \sum_i p_i L_2(b_i^*, b_i).$$

- We only attempt to modify an anchor if it is a positive anchor.
- p_i is 0 if anchor is negative and 1 if anchor is positive.
- N_p is the number of positive anchors in the minibatch.
- b_i^* is the ground truth bounding box.
- b_i is the estimated bounding box, applying the regressed residuals to the anchor box parameters. 上記の「Classification VS Regression Heads」を参考。

疑問：1つground truth bounding boxに対して1つpositive anchorのみ出力されるという制限はやっていないでしょう？ minibatchでもただnegativeとpositive anchorの比率を制限している。
Lesson 2にこういう言葉がある：「Any redundant detections are considered false positives and results in lower precision and thus a lower average precision overall. In fact, this is only an issue



of inference and it turns out to be beneficial to drive the network to output the best possible residuals for every anchor during training.」
まずは、“any redundant detections are considered false positives”.これはどこにやっているかはまだ分かっていない。
また、“This is only an issue of inference”、つまりtrainingと関係ない? 確かにminibatchをみると、redundant detectionsは処理してなさそう。



Non-Maximum Suppression
- アルゴリズムわかりやすい。出力のbounding boxリストに、常にscoreが一番高いb_{max}を選んで、残るbboxesに、b_{max}と結構重なっているbboxを探して、リストから削除、b_{max}のみ保留。リストが空になるまでこのプロセスを繰り返す。
- D now contains a single bounding box per object.
  - いい方法ですが、この保証は本当にあるの?

- To train a neural network for 2D object detection, use minibatch selection on anchors.
  - to maintain class balance.
  - 学習中redundant bounding boxは本当にnegativeになってる?
- For inference, use Non-Maximum Suppression to get a single output bounding box per object.
- [2015] SSD: Single Shot MultiBox Detector: https://arxiv.org/abs/1512.02325
- [2016] Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks: https://arxiv.org/abs/1506.01497
- [2017] Focal Loss for Dense Object Detection: https://arxiv.org/abs/1708.02002
- 時間探して読もう!

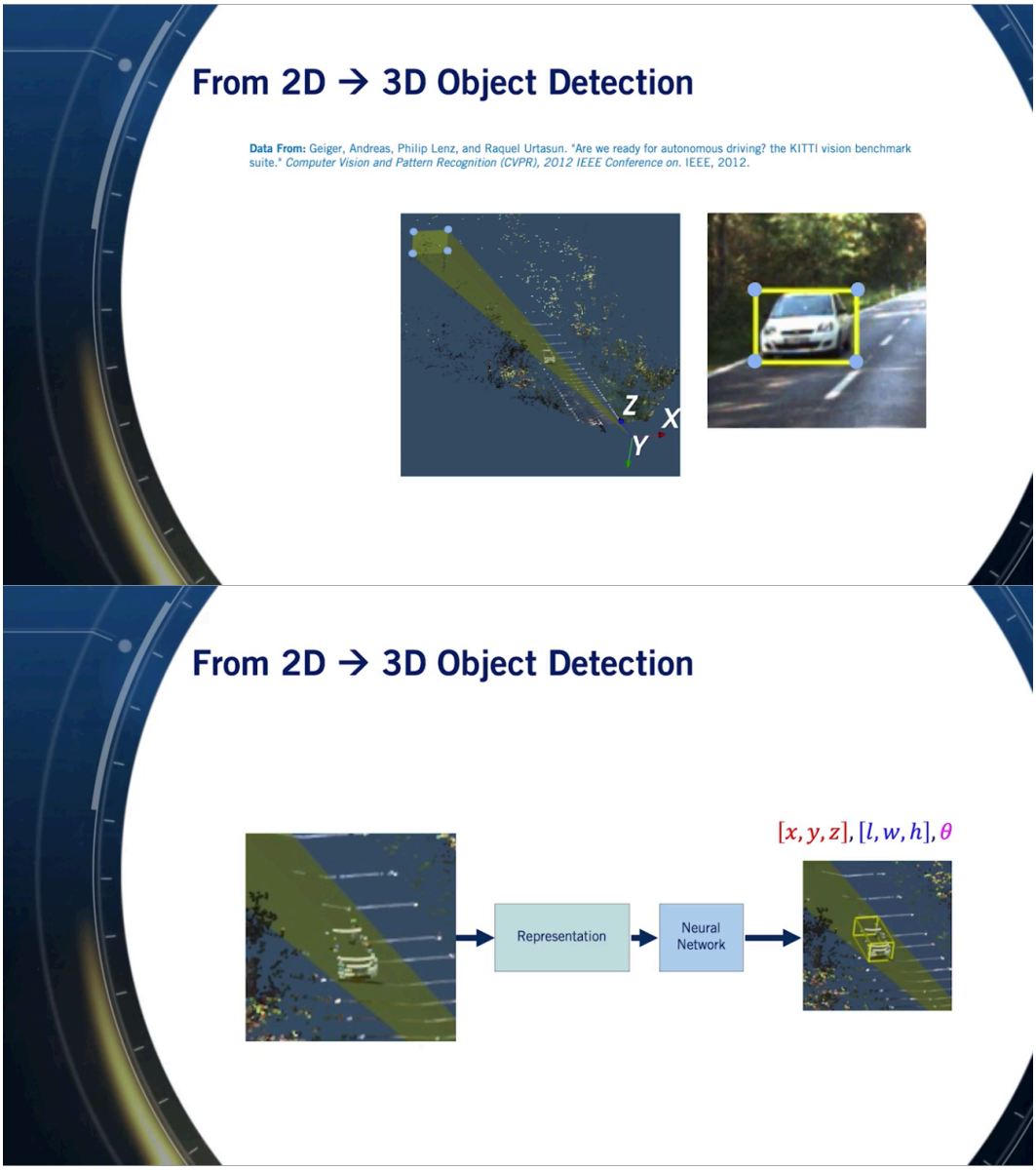
Lesson 4: Using 2D Object Detectors for Self-Driving Cars

単語
- frustum: 錐台「すいだい」1. the part of a solid, such as a cone or pyramid, contained between the base and a plane parallel to the base that intersects the solid. 2. the part of such a solid contained between two parallel planes intersecting the solid.

- teleport: 瞬間移動 (in science fiction) to transport (a person or object) across a distance instantaneously.

内容
- Extend 2D object detection output to 3D.
- Use 2D object detection output to track objects in images and in 3D.
- Use 2D object detection to detect traffic signs and signals.

3D Object Detection
- Estimating the:
  - Category Classification: Car, pedestrian, cyclist.
  - Position of the centroid in 3D: $[x, y, z]$.
  - Extent in 3D（2Dのbounding boxと似てる）: $[l, w, h]$.
  - Orientation in 3D: $[\phi, \psi, \theta]$.
    - orientation angles. roll, pitch, yaw.
    - For road scenes, the orientation angle we are interested in is usually only the yaw angle.



From 2D -> 3D Object Detection（流れ、大事）
- The most common and successful way to extend 2D object detection results in 3D is to use LiDAR point clouds.
- Given a 2D bounding box in an image space and a 3D LiDAR point cloud, we can use the inverse of the camera projection matrix to project the corners of the bounding box as rays into the 3D space.（上記の図「From 2D -> 3D Object Detection」）
- The polygon intersection of these lines is called a frustum, and usually contains points in 3D that correspond to the object in our image.
  - つまりraysとpoint cloudのintersection.
- We then take the data in this frustum from the LiDAR, transform it to a representation of our choice, and train a small neural network to predict the seven parameters required to define our bounding box in 3D.

frustumのLiDAR pointsから3種類Representation
1. Some groups choose to directly process the raw point cloud data.
    1. X Y Z.
2. Others choose to normalize the point cloud data w.r.t. some fixed point such as the center of the frustum.
    1. Normalized X Y Z.
3. One could also preprocess the points to build fixed length representations such as a histogram of x, y, z points, making their use as an input to a ConvNet much more convenient.
    1. Histogram of X Y Z.

From 2D->3D Object Detectionのメリットやディメリット（大事）
- Advantages
  - Allows exploitation of mature 2D object detectors, with high precision and recall.
  - Class already determined from 2D detection.
  - Doest not require prior scene knowledge, such as ground plane location.
  - Searching a 3D space for possible objects is quite computationally expensive if no assumptions can be made about where the object should be found.
    - Extending 2D object detectors to 3D, usually allows us to limit the search region for object instances, keeping real-time performance manageable.
- Disadvantages
  - The performance of the 3D estimator is bounded by the performance of the 2D detector.
  - Occlusion and truncation are hard to handle from 2D only.
    - Which may not affect the LiDAR data.
  - 3D estimator needs to wait for 2D detector, inducing latency in our system.
    - If this delay is significant, the system might not be safe enough to operate as vehicle reaction time is limited by perception latency.

2D Object Tracking（コース２の考え方、Measurements、Motion Model、Kalman Filter）この流れは既にMatLabのObject Trackingソースコードから知ってた。
- Detection: We detect the object independently in each frame and can record its position over time.
- Tracking: We use image measurements to estimate position of object, but also incorporate position predicted by dynamics, i.e., out expectation of object's motion pattern.
- Tracking Assumptions:
  - Constraining how quickly a scene changes.
  - For example, we assume that our camera and the tracked objects cannot teleport to different locations within a very short time.
    - Camera is not moving instantly to new viewpoint.
    - Objects do not disappear and reappear in different places in the scene.
  - If the camera is moving, there is a gradual change in pose between camera and scene.
  - All of these assumptions are logically valid in roads scenes.

2D Object Trackingの流れ（大事）
- Given a detected object in the first frame along with their velocity vectors.

- Prediction:
  - We begin by predicting where the objects will end up in the second frame, if we model their motion using the velocity vector.
- Measurement:
  - We then get new detections in the second frame.
- Correlation:
  - We correlate each detection with a corresponding measurement.
- Measurement Update:
  - And then update our object detection using the correlated measurement.
  - つまりcorrelateされたdetectionはそのまま使うではなく、また調整するんだ!

Object Tracking: Prediction
- State: its position and velocity in image space.
- Each object will have a predefined motion model in image space.
  - That updates its state.
- Example: $p_k = p_{k-1} + v_k \Delta t + \mathcal{N}(0, \Sigma)$.
  - Constant velocity motion model.

Object Tracking: Correlation（大事）
- Get Measurement Bounding Boxes from 2D detector.
- Correlate prediction with the highest IOU measurement.
  - A measurement is correlated to a corresponding prediction if it has the highest IOU with that prediction.
  - ~~つまり1つpredictionは必ず1つmeasurementを持つ。逆は成立しない。~~
    - 違う。predictionの相手measurementが存在しなければ、tracking終了。measurementの相手predictionが存在しなければ、tracking開始。

Object Tracking: Update
- The prediction and measurement are fused as part of the Kalman Filter Framework.
- Note that we do not need to track the object sizes but can rely on the detector instead.

Object Tracking
- 課題：How to initiate tracks and how to terminate them.
- For each frame, we start new track if a measurement has no correlated prediction.
- We also terminate inconsistent tracks, if a predicted object does not correlate with a measurement for a preset number of frames.
- The same methodology can be used to track objects in 3D!
  - by defining IOU in 3D.

Traffic Sign and Traffic Signal Detection
- Usually, traffic signs and traffic signals have to be detected at long range for a car to know how to react properly in a timely manner.
- At long range, traffic signs and signals occupy a very small number of pixels in the image making the detection problem particularly challenging.
  - Traffic signs and signals appear smaller in size compared to cars, two-wheelers, and pedestrians.
- Traffic signs are highly variable with many classes to be trained on.
  - Usually including as many as 50 classes that need to be classified reliably.
- Traffic signals have different states that are required to be detected.
- In addition, traffic signals change state as the car drives!

Traffic sign and signal detection
- 2D object detectors can be used to perform traffic sign and traffic signal detection without any modifications.
- However, multi-stage hierarchical models have been shown to outperform the standard single state object detectors.
- The two stages share the output of the feature extractor to perform their respective task.

- In this example, the first stage outputs class agnostic bounding boxes that point to all traffic signs and signals in the image.
- The second state then takes all of the bounding boxes from the first stage and classifies them into categories such as red, yellow or green signals, stop signs, etc.
- Some methods also use the second stage to further refine the bounding boxes.

論文

- [2017] Frustum PointNets for 3D Object Detection from RGB-D Data: https://arxiv.org/abs/1711.08488