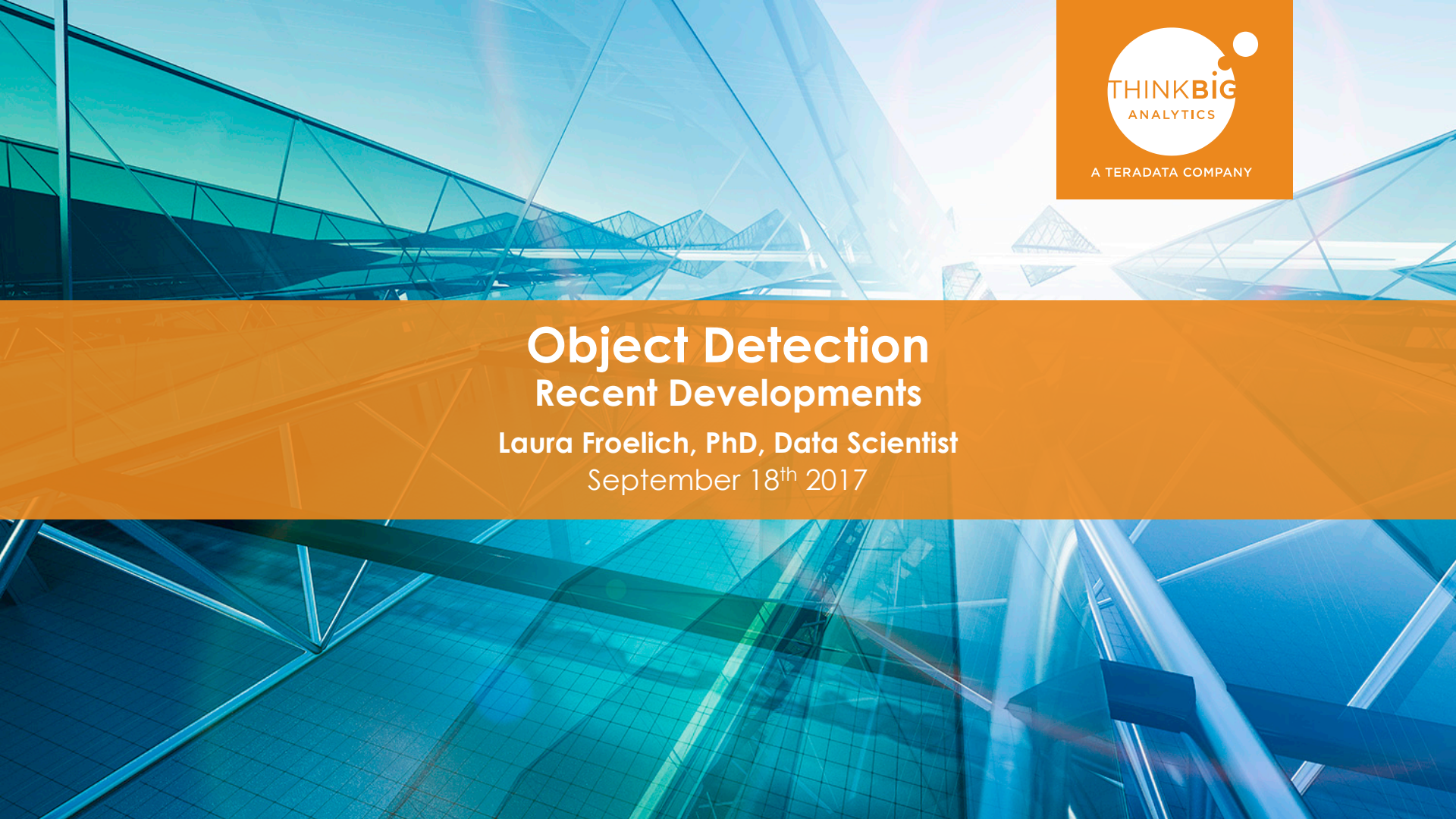


Object Detection
Recent Developments

Laura Froelich, PhD, Data Scientist

September 18th 2017

Agenda

THINKBIG ANALYTICS
A TERADATA COMPANY

1. What is object detection?
2. Recent history of object detection
    1. Region proposal approach
    2. Single forward pass approach
    3. Direct classification
    4. Refined classification
3. Quantification of object recognition
4. Single Shot Multibox Detector

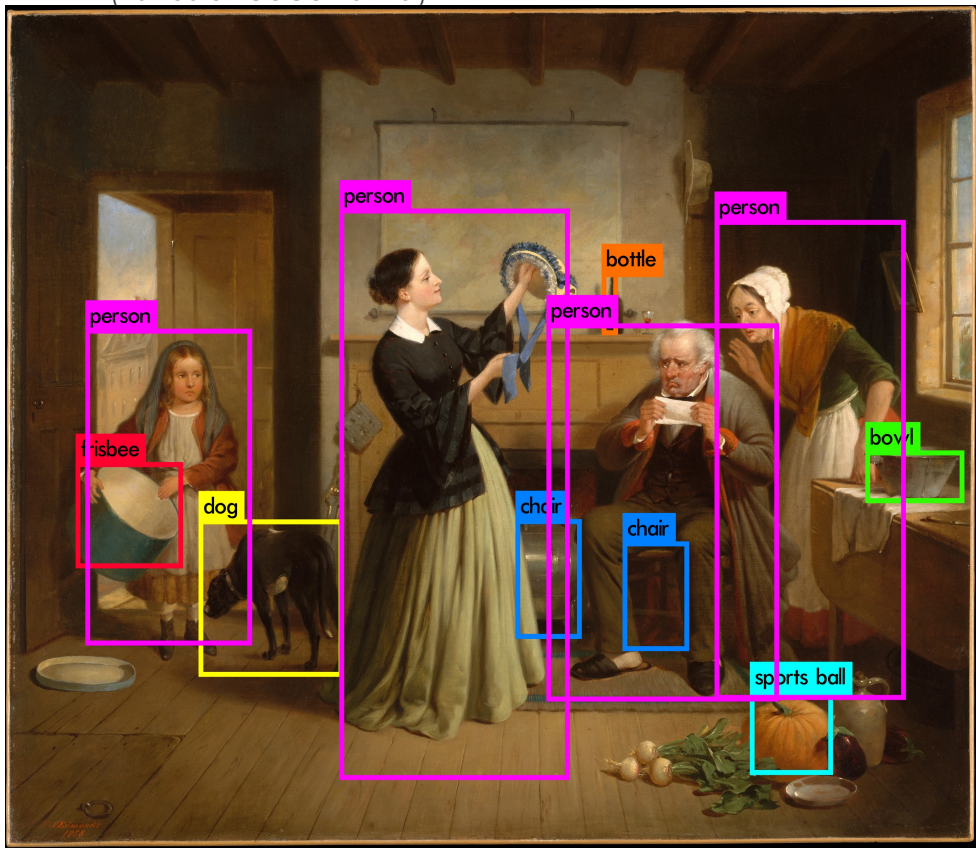
Public domain, http://images.metmuseum.org/CRDImages/ap/original/DT77.jpg

What is object detection?

THINKBIG ANALYTICS
A TERADATA COMPANY

Predictions from You Only Look Once v2 https://arxiv.org/abs/1612.08242
(trained on COCO trainval)

- Detect locations of objects in image
  - Object locations are defined by boxes surrounding objects.
  - Boxes surrounding objects are referred to as Bounding Boxes
- Classify detected objects

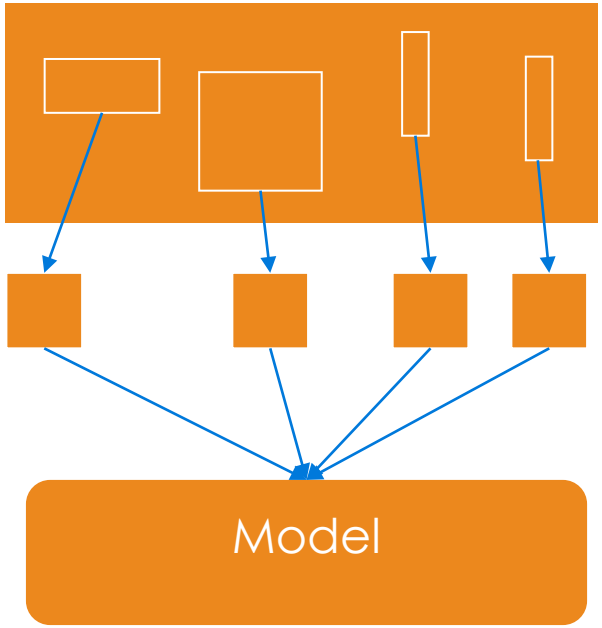


Heuristic Region Proposal Approach

Approx. 2014-2015

THINKBIG ANALYTICS
A TERADATA COMPANY

- Idea
  - Quickly run through image to detect regions likely to contain objects.
    - All three methods listed below use Selective Search.
  - Process proposed regions more extensively.
    - Refine proposed regions (bounding boxes).
    - Predict class of object.
- Methods
  - Region-based Convolutional Neural Networks (R-CNN), Girschick et al. (**2014**), Rich feature hierarchies for accurate object detection and semantic segmentation
  - Spatial Pyramid Pooling Net (SPPNet), He et al. (**2014**), Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition
  - Fast R-CNN, Girschick (**2015**), Fast R-CNN



Single pass of multiple pre-defined regions

Approx. 2014

THINKBIG ANALYTICS
A TERADATA COMPANY

- Idea
  - Extract features from evenly spaced squares on differently scaled copies of image.
  - Predict bounding box corrections and class probabilities for each square.
- Methods
  - OverFeat, Sermanet et al. (**2014**), OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks

Combine predictions from each window

Direct Classification

Approx. 2015-2016

THINKBIG ANALYTICS
A TERADATA COMPANY

- Idea
  - Use a region proposal neural network to propose regions instead of heuristic method. Using a neural network to propose regions makes it possible to improve the proposed regions through training.
    - Either propose only center of object's bounding box or center as well as size.
  - Predict corrections to bounding box proposals and class probabilities for the proposed regions.
- Methods
  - You Only Look Once (YOLO)
    - Version 1: Redmon et al. (**2015**), You Only Look Once: Unified, Real-Time Object Detection
    - Version 2: Redmon et al. (**2016**), YOLO9000: Better, Faster, Stronger
  - Single Shot Multibox Detector (SSD), Liu et al. (**2015**), SSD: Single Shot MultiBox Detector
  - Faster R-CNN, Ren et al. (**2015**), Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks

Refined Classification

2017

THINKBIG ANALYTICS
A TERADATA COMPANY

- Idea
  - Add another branch, in addition to the bounding box regression and classification branches, that predicts pixel-wise classes.
  - Multi-task training should improve results.
  - The pixel-wise predictions should refine the predicted region.
- Methods
  - Mask R-CNN, He et al. (**2017**), Mask R-CNN

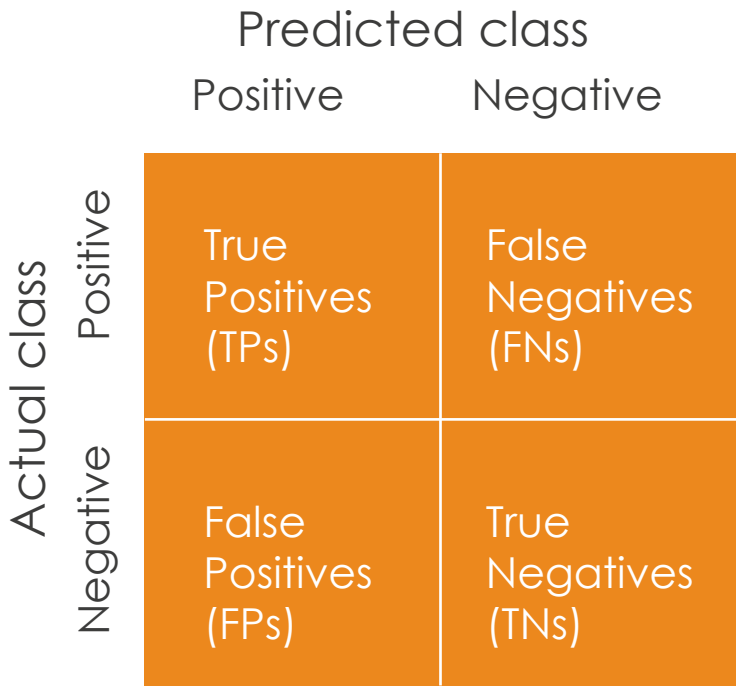
Quantifying overlap of bounding boxes

THINKBIG ANALYTICS
A TERADATA COMPANY

- Jaccard similarity, also known as Intersection over Union (IoU)
- Definition: Let A and B denote sets of pixels from two bounding boxes
  - IoU = | A intersection B | / | A union B |
- Applications
  - Define matches between predicted and ground truth bounding boxes.
  - Non-maximum suppression:
    - For predicted bounding boxes with IoU above pre-specified threshold, only retain the predicted bounding box that predicts the highest class confidence for any one class.

Recall and Precision

- Recall: TPs/(TPs+FNs)
  - "How well does the model discover objects that were there"
- Precision: TPs/(TPs+FPs)
  - "Degree to which model avoids spurious detections"
- True Negatives not included in either
  - sensible in cases with many True Negatives that are not important
- Predictions usually given with probability
  - Increasing detection threshold reduces number of positive predictions
  - Increasing detection threshold improves precision and worsens recall



Mean Average Precision (mAP)

Taking Precision and Recall into account simultaneously

THINKBIG ANALYTICS
A TERADATA COMPANY

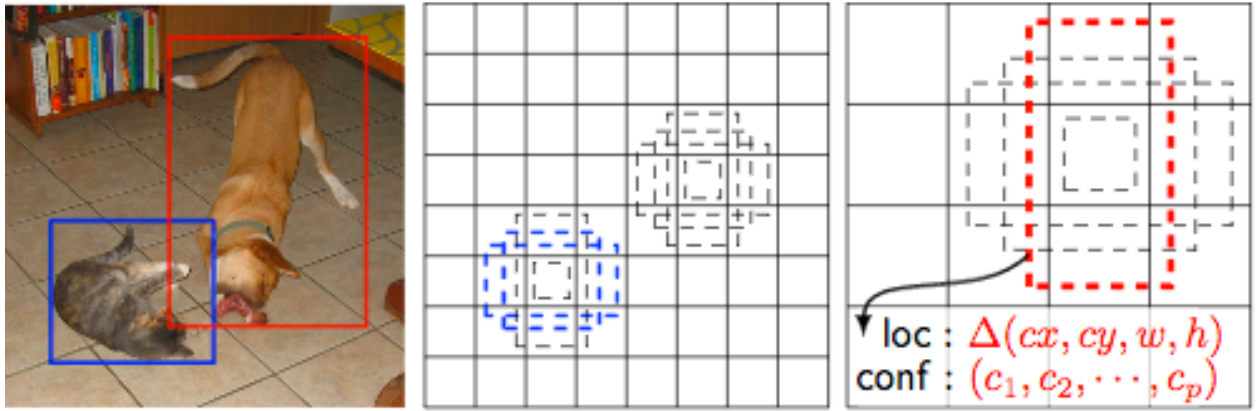
- Definitions vary in the details.
- General idea:
  - At each of a number of recall levels, retrieve the precision,
  - average these precision values,
  - this yields the mean Average Precision.

SSD: Single Shot MultiBox Detector

https://arxiv.org/pdf/1512.02325.pdf

- Single Shot
  - Only one pass of the model over each image
- MultiBox
  - Give predictions for each of several prior boxes

Figure1 in SSD: Single Shot MultiBox Detector https://arxiv.org/pdf/1512.02325.pdf



SSD, prior boxes and multiple scales

On feature maps of different scales, a number of prior boxes of different aspect ratios are placed at each location in the feature map.

a) Input for training: images with ground truth bounding boxes and labels
b) Intermediate calculation: two prior boxes match ground truth cat
c) Intermediate calculation: one prior box matches ground truth dog

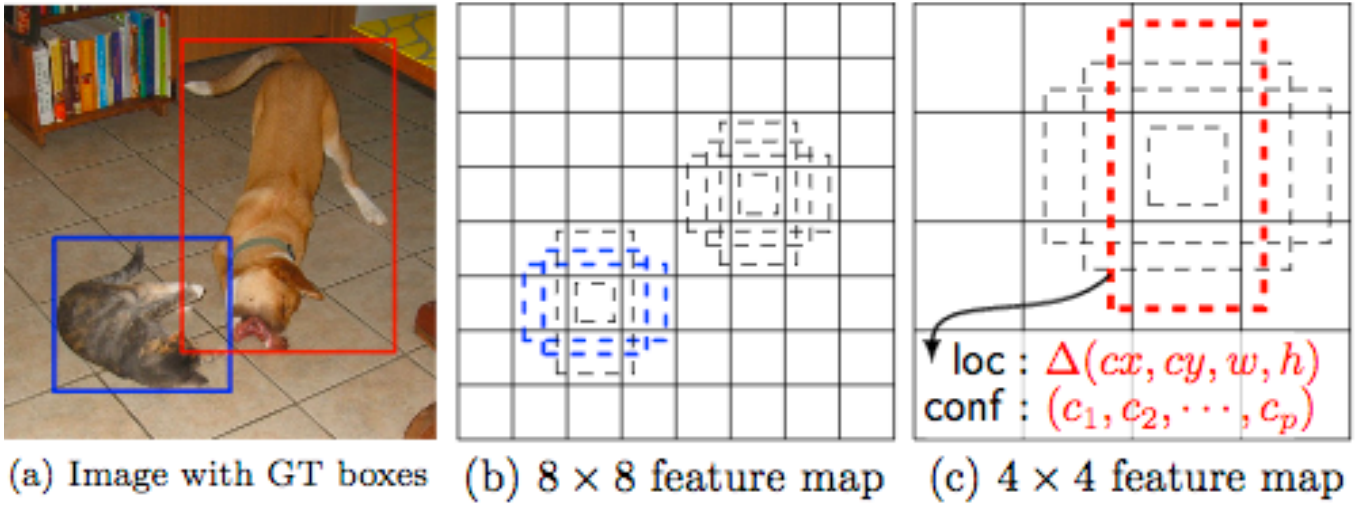


Figure1 in
SSD: Single Shot MultiBox Detector
https://arxiv.org/pdf/1512.02325.pdf

SSD, Cost function

- The cost function consists of two terms:
  - Penalty on wrong predictions of confidence.
  - Penalty on wrong predictions of location.
- The relative importance is given by alpha. Through cross-validation, alpha was set to one.
- N is the number of matched prior boxes.

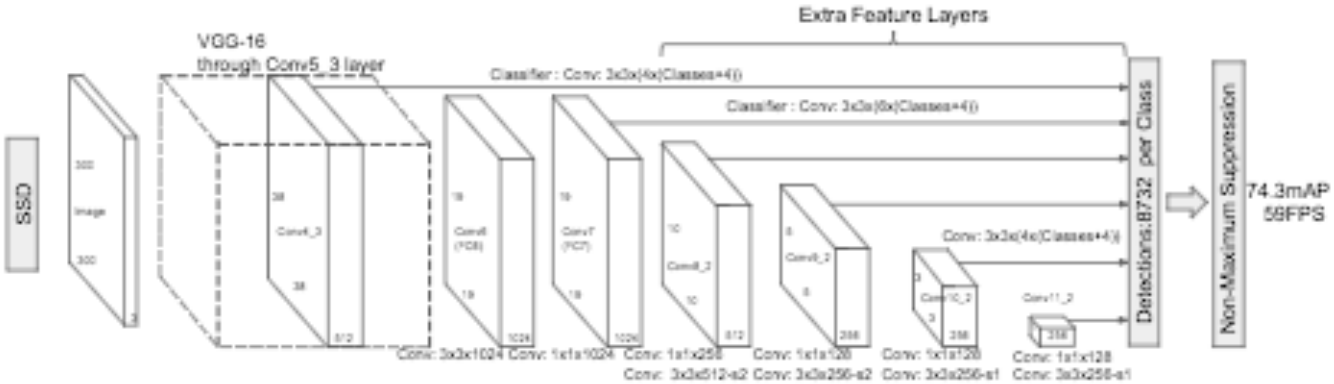
$$L(x, c, l, g) = \frac{1}{N}(L_{conf}(x, c) + \alpha L_{loc}(x, l, g))$$

Equation 1 in
SSD: Single Shot MultiBox Detector
https://arxiv.org/pdf/1512.02325.pdf

SSD, architecture

- VGG16 as basenet, without classification layer.
- VGG16 followed by convolutional layers that generate features at different scales.

From Figure 2 in
SSD: Single Shot MultiBox Detector
https://arxiv.org/pdf/1512.02325.pdf



On to some code!

SSD, cost function in more detail

- Sum loss over matched default boxes
- x^k_{ij}: "indicator for matching the i^{th} default box to the j^{th} ground truth box of category p."
- Location loss:
  - Predict relative to default boxes (d)
    - cx, cy: offsets for center of box
    - w, h: multiplicative offsets for width and height
- Confidence loss:
  - Multiple class softmax loss

$$L_{loc}(x,l,g) = \sum_{i \in Pos}^{N} \sum_{m \in \{cx,cy,w,h\}} x_{ij}^k \text{smooth}_{\text{L1}}(l_i^m - \hat{g}_j^m)$$

$$\hat{g}_j^{cx} = (g_j^{cx} - d_i^{cx})/d_i^w \qquad \hat{g}_j^{cy} = (g_j^{cy} - d_i^{cy})/d_i^h$$

$$\hat{g}_j^w = \log\left(\frac{g_j^w}{d_i^w}\right) \qquad \hat{g}_j^h = \log\left(\frac{g_j^h}{d_i^h}\right)$$

Equation 2 in
SSD: Single Shot MultiBox Detector
https://arxiv.org/pdf/1512.02325.pdf

$$L_{conf}(x,c) = -\sum_{i \in Pos}^{N} x_{ij}^p log(\hat{c}_i^p) - \sum_{i \in Neg} log(\hat{c}_i^0) \quad \text{where} \quad \hat{c}_i^p = \frac{\exp(c_i^p)}{\sum_p \exp(c_i^p)}$$

Equation 3 in
SSD: Single Shot MultiBox Detector
https://arxiv.org/pdf/1512.02325.pdf