LIMITLESS LABELS IN A LABELLESS WORLD: WEAK SUPERVISION WITH NOISY LABELS

Arash Vahdat, August 23rd 2020

ECCV Tutorial: New Frontiers for Learning with Limited Labels or Data

TRAINING WITH LABEL NOISE

The Inevitable Problem

Deep learning is data hungry: massive datasets are required for the best results

Collecting data can be done cheaply, but collecting annotation is costly:

- Train annotators, hire experts in a field, ensure consistency using multiple annotators, etc.

Annotation cost can be reduced with **cheaply collected but noisy labels**

Label Noise in an inevitable problem:

- We can only reduce the label noise but we cannot eliminate it completely (ambiguity and subjectivity)
- Many well-known datasets contain some labels noise: CIFAR10 or ImageNet

Robust deep learning: learn better models from noisy labeled data

TODAY'S AGENDA

A Quick Tour

- A quick tour of the prior arts
- A probabilistic framework for learning from noisy labels
  - Image labeling
  - Object segmentation
  - Object detection
- Training from text as noisy labels with contrastive learning
- Conclusions and open questions

PRIOR WORK

Training from Noisy Labels*

(1) Loss Correction

Noise Transition Model [1]

Forward and Backward Losses [2]

Label Smoothing Regularization [3]

$l(T^T f_\theta(x_i), \tilde{y}_i)$ where $T_{i,j} = p(\tilde{y} = j | y = i)$

(2) Robust Loss Functions

Mean Absolute Error [7]

Generalized Cross Entropy [8]

Normalized Loss Functions [9]

$l(f_\theta(x_i), \tilde{y}_i) = ||f_\theta(x_i) - \tilde{y}_i||_1$

(3) Curriculum Learning

Sample Reweighting [10]

MentorNet [11]

Decoupling [12], Co-teaching [13]

$$\frac{1}{N}\sum_{i=1}^{N} \beta(x_i, y_i) l(f_\theta(x_i), \tilde{y}_i))$$

(4) Label Correction

Graphical Models [4]

Neural Networks [5]

Knowledge Graphs [6]

(5) Regularization

Dropout [14]

Mixup [15]

* A recent survey is available [26]

TOWARD ROBUSTNESS AGAINST LABEL NOISE IN TRAINING DEEP DISCRIMINATIVE NEURAL NETWORKS

Arash Vahdat,

NeurIPS 2017

IMAGE LABELING

Problem Definition

predict a set of binary labels in images

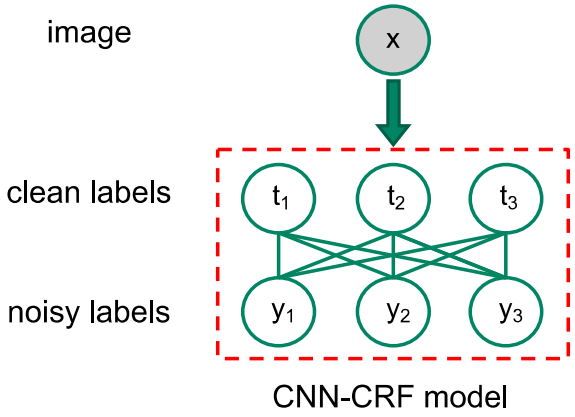
Labels	cat	cooker	kitchen	plant	person	appliance
Noisy Annotation (**y**)	1	1	0	0	0	0
Clean Annotation (**t**)	1	1	1	1	0	1

Text: my cat is sleeping in the cooker!

DEEP STRUCTURE PREDICTION

Problem Formulation

- Define a joint distribution over clean and noisy labels using Conditional Random Fields (CRFs)
- CNN to compute the parameters of the CRF



CNN-CRF model

the joint distribution
over clean and noisy labels

$$p_\theta(y,t|x) \propto \exp(-E_\theta(y,t,;x))$$

$p_\theta(t|x,y)$ can be formed easily!

TRUE LABELS AS LATENT VARIABLES

A Latent Variable Problem

Maximize the marginal log-likelihood of the data:

$$\max_{\theta} \sum_{(x^{(n)},y^{(n)})} \log\, p_\theta\left(y^{(n)}\middle|x^{(n)}\right)$$

This term is formed by marginalizing the conditional over the latent variables:

$$\log\left(p_\theta(y|x)\right) = \log\left(\sum_t p_\theta(y,t|x)\right)$$

Intractable summation!

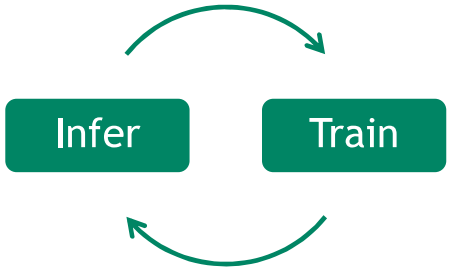
Expectation-Maximization (EM) can be used!

EM ALGORITHM

Iterating between Expectation and Maximization Steps

Expectation Step

$q = p_\theta(t|x, y)$



Maximization Step

$\max_\theta \mathbb{E}_{t\sim q}[\log p_\theta(y, t|x)]$

The maximization steps is fitting the model to incorrect labels inferred in the expectation steps!

REGULARIZED EM

Guiding the Maximization Step

Assume we have access to an auxiliary source of information that can help us infer true labels

We denote the auxiliary distribution by $p_{aux}(t|x,y)$

We can modify the EM objective such that the inference model is guided towards $p_{aux}(t|x,y)$:

$$\min_{q} \underbrace{\mathrm{KL}(q(t|x,y) \,||\, p_\theta(t|x,y))}_{\text{true posterior}} + \alpha \underbrace{\mathrm{KL}(q(t|x,y) \,||\, p_{aux}(t|x,y))}_{\text{auxiliary distribution}}$$

Modified E-step in regularized EM: $q(t|x,y) \propto \left(p_\theta(t|x,y)\, p_{aux}^{\alpha}(t|x,y)\right)^{\frac{1}{\alpha+1}}$

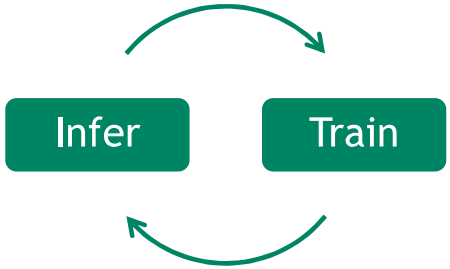
The geometric mean is equivalent to weighted average in the log probability space. It can be implemented by weighted averaging of logits before applying softmax/sigmoid.

REGULARIZED EM ALGORITHM

Iterating between Expectation and Maximization

Expectation Step

$$q(t|x,y) \propto \left(p_\theta(t|x,y)\, p_{aux}^{\alpha}(t|x,y)\right)^{\frac{1}{\alpha+1}}$$



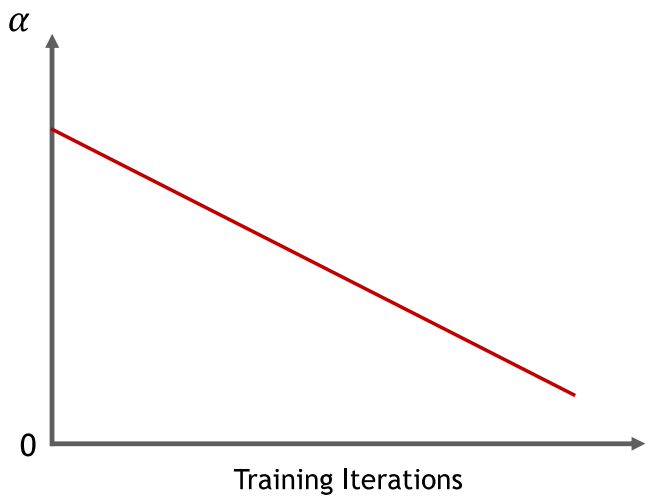
Maximization Step

$$\max_\theta \mathbb{E}_{t\sim q}[\log p_\theta(y,t|x)]$$

HOW TO CHOOSE α

The Balancing Coefficient

$$q(t|x,y) \propto \left(p_\theta(t|x,y)\, p_{aux}^{\alpha}(t|x,y)\right)^{\frac{1}{\alpha+1}}$$



DO WE NEED STRUCTURE PREDICTION MODEL?

The Joint Distribution on Noisy and Clean Labels

- Recall that we derived our framework for the joint distribution $p_\theta(y, t|x)$
- Assume that we train a classification model only for clean labels in the form $p_\theta(t|x)$
- Similar to pseudo labels but in a probabilistic setting

Expectation Step

$$q(t|x,y) \propto \left(p_\theta(t|x)\, p_{aux}^{\alpha}(t|x,y)\right)^{\frac{1}{\alpha+1}}$$

Self-correcting inference

Maximization Step

$$\max_\theta \mathbb{E}_{t\sim q}[\log p_\theta(t|x)]$$

Cross Entropy Loss
with q as Soft Labels

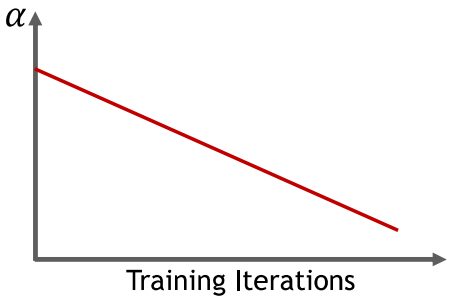
TAKE-HOME MESSAGE

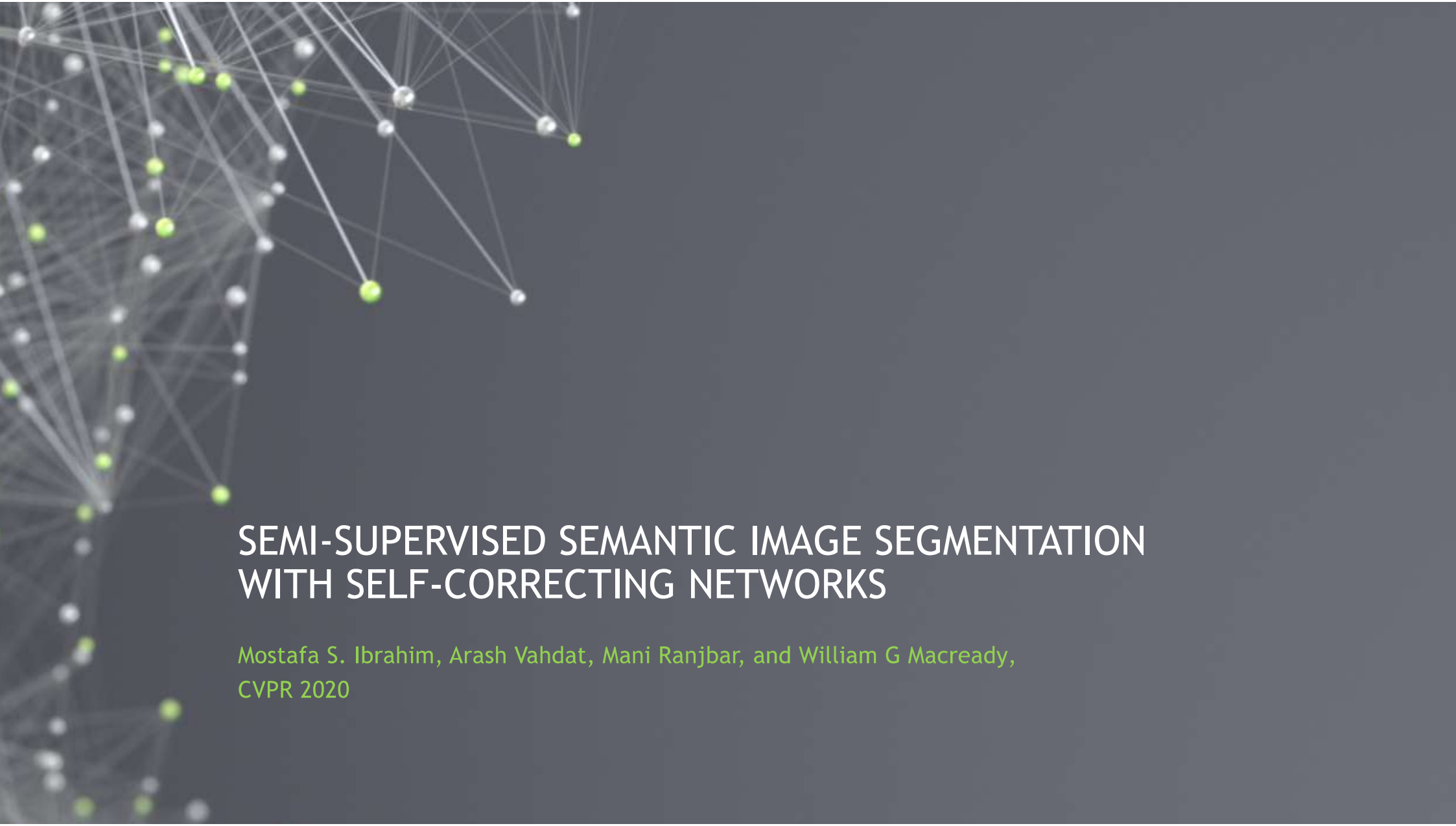
Deep Learning with Noisy Labels

- Label noise is a common problem in real-world applications
- Any auxiliary source of information can help us infer a distribution over true labels
- Our self-correction principal relies on the geometric mean of the current prediction model and an auxiliary distribution

Self-correcting Principal

$$q(t|x,y) \propto \left(p_\theta(t|x)\, p_{aux}^{\alpha}(t|x,y)\right)^{\frac{1}{\alpha+1}}$$





SEMI-SUPERVISED SEMANTIC IMAGE SEGMENTATION WITH SELF-CORRECTING NETWORKS

Mostafa S. Ibrahim, Arash Vahdat, Mani Ranjbar, and William G Macready,

CVPR 2020

SEMANTIC OBJECT SEGMENTATION

Classifying each Pixel

Label noise is unavoidable: object boundaries, small resolution, …

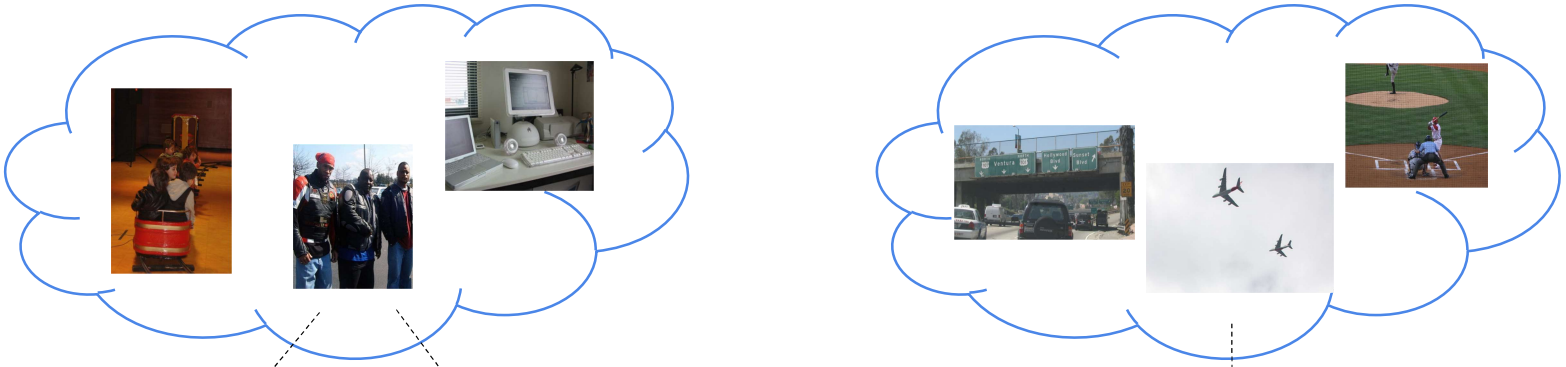
Labeling is expensive: ~8x slower than drawing bounding boxes and ~80x slower than image labeling [27]

SEMI-SUPERVISED SEMANTIC IMAGE SEGMENTATION

Problem Setup

Small Fully labeled Set + Large Weakly labeled Set

Images:



Annotations:

Bounding Box

Segmentation

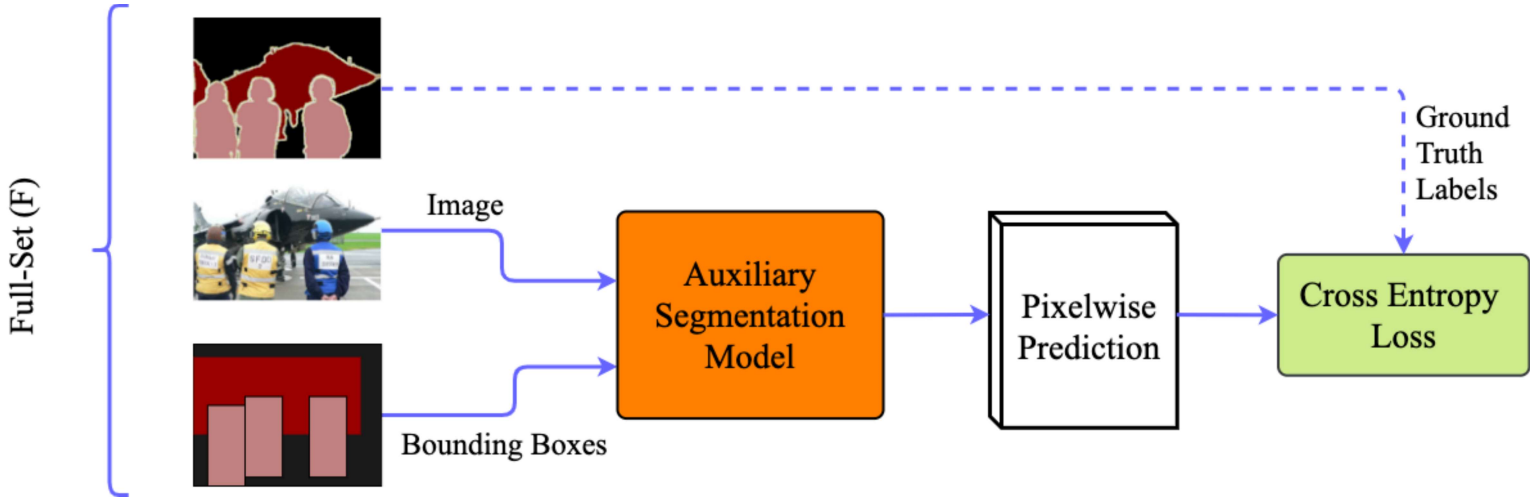
Bounding Box

AUXILIARY DISTRIBUTION $p_{aux}(t|y,x)$

Predict Segmentation Mask from Image and Bounding Box

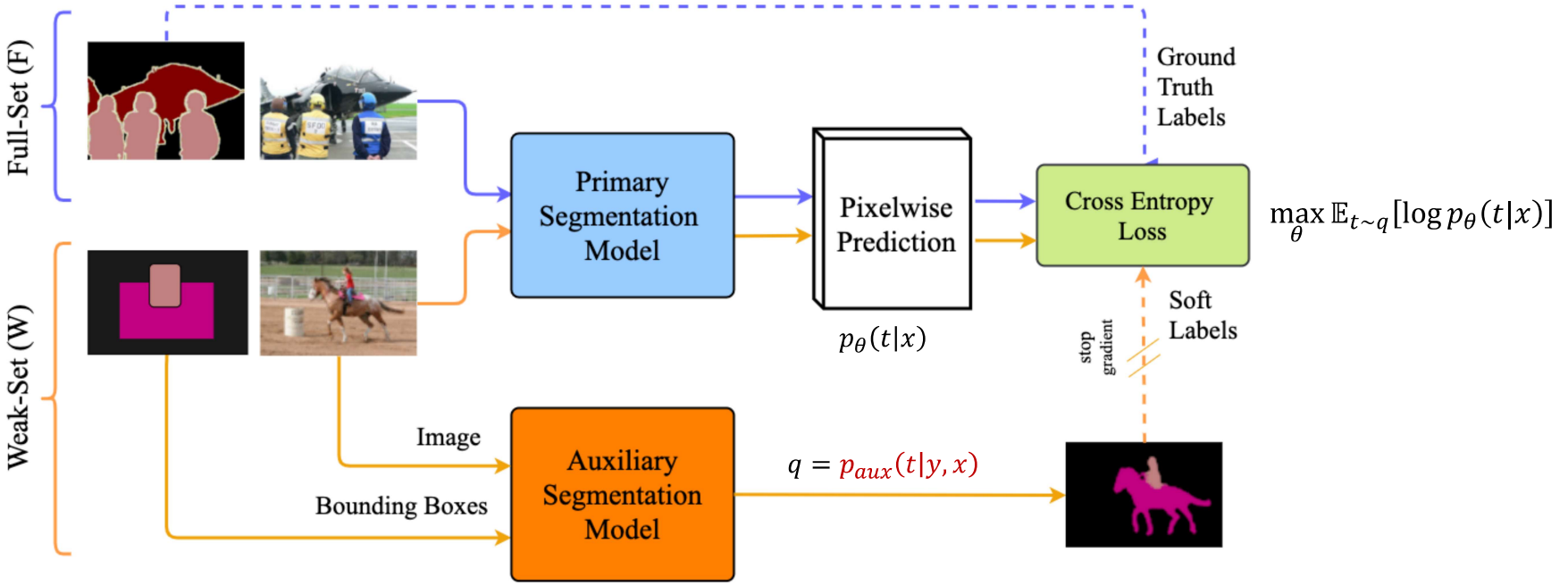
Recall that our method require an auxiliary model that infers true labels given noisy labels and the input image.

We can use the fully-labeled set to train such model:



TRAIN THE PRIMARY SEGMENTATION MODEL

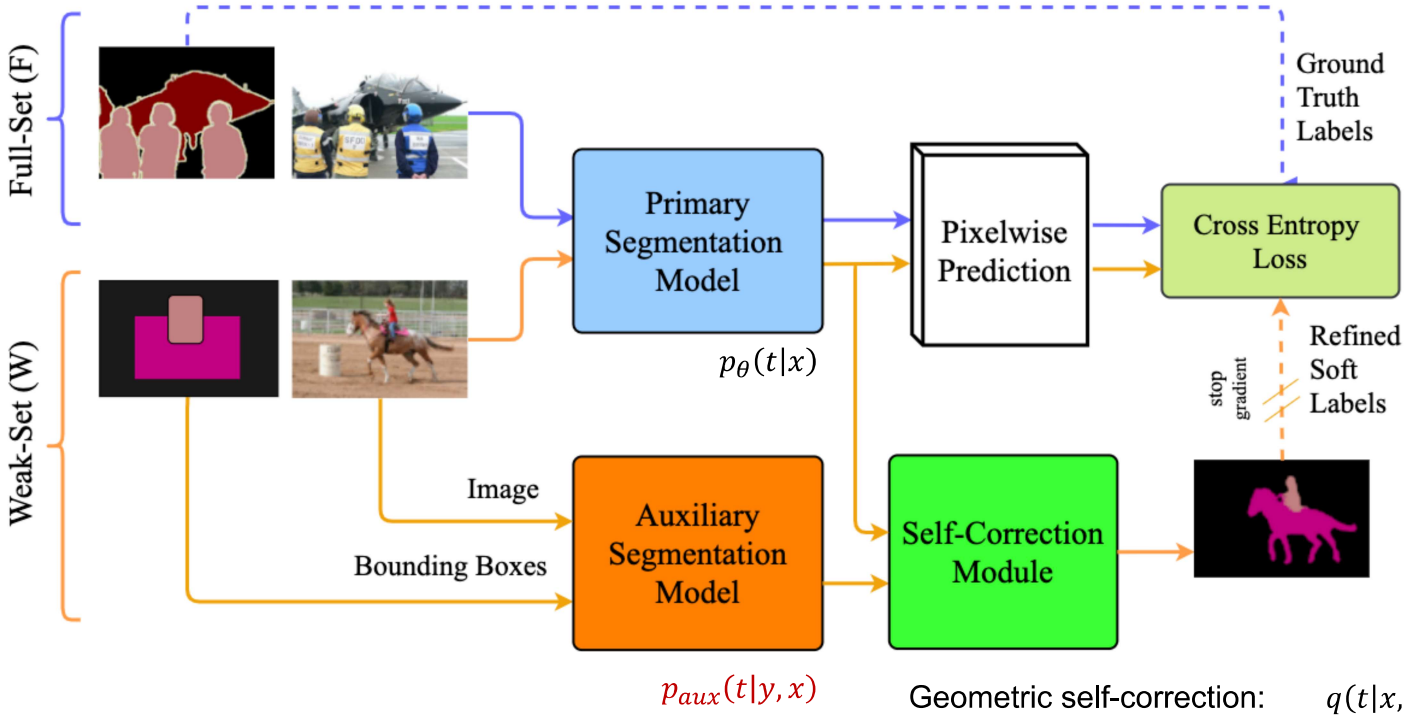
A Semi-Supervised Learning Approach



$$q(t|x,y) \propto \left(p_\theta(t|x)\, p_{aux}^{\alpha}(t|x,y)\right)^{\frac{1}{\alpha+1}} \text{ with } \alpha = \infty$$ No self-correction!

TRAIN THE PRIMARY SEGMENTATION MODEL

With Self-Correction Module



Geometric self-correction: $q(t|x,y) \propto \left(p_\theta(t|x)\, p_{aux}^{\alpha}(t|x,y)\right)^{\frac{1}{\alpha+1}}$

Convolutional self-correction: $q(t|x,y) = f\left(p_\theta(t|x), p_{aux}(t|x,y)\right)$

TAKE-HOME MESSAGE

Semi-Supervised Semantic Segmentation as Training with Noisy Labels

- Collecting annotation for semantic object segmentation is very costly
- Label noise is inevitable for this task
- We can consider bounding box annotations as noisy label in this problem
- The semantic object segmentation problem can benefit a lot from robust training models

A ROBUST LEARNING APPROACH TO DOMAIN ADAPTIVE OBJECT DETECTION

Mehran Khodabandeh, Arash Vahdat, Mani Ranjbar, and William G. Macready,

ICCV 2019

DOMAIN ADAPTATION IN OBJECT DETECTION

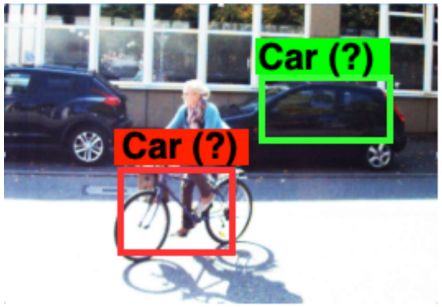
As Training with Noisy Labels

Domain shift is an unavoidable problem in computer vision: variations in viewpoint, background, object appearance, scene type and illumination.

Can we formulate domain adaptation in object detection as training with noisy labels?

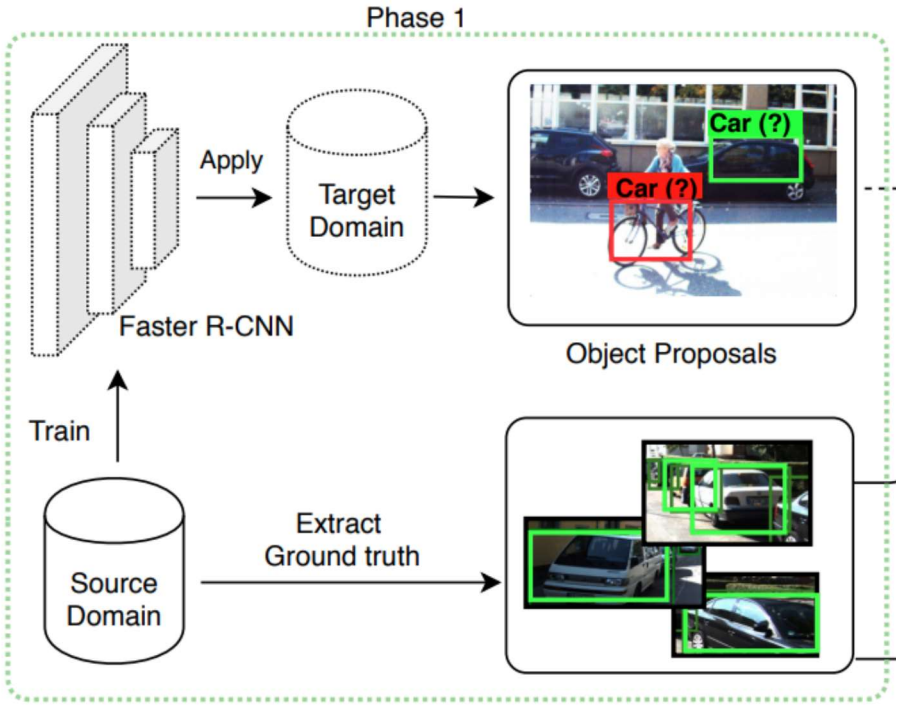
- We need a robust training framework for object detection!

Source Domain



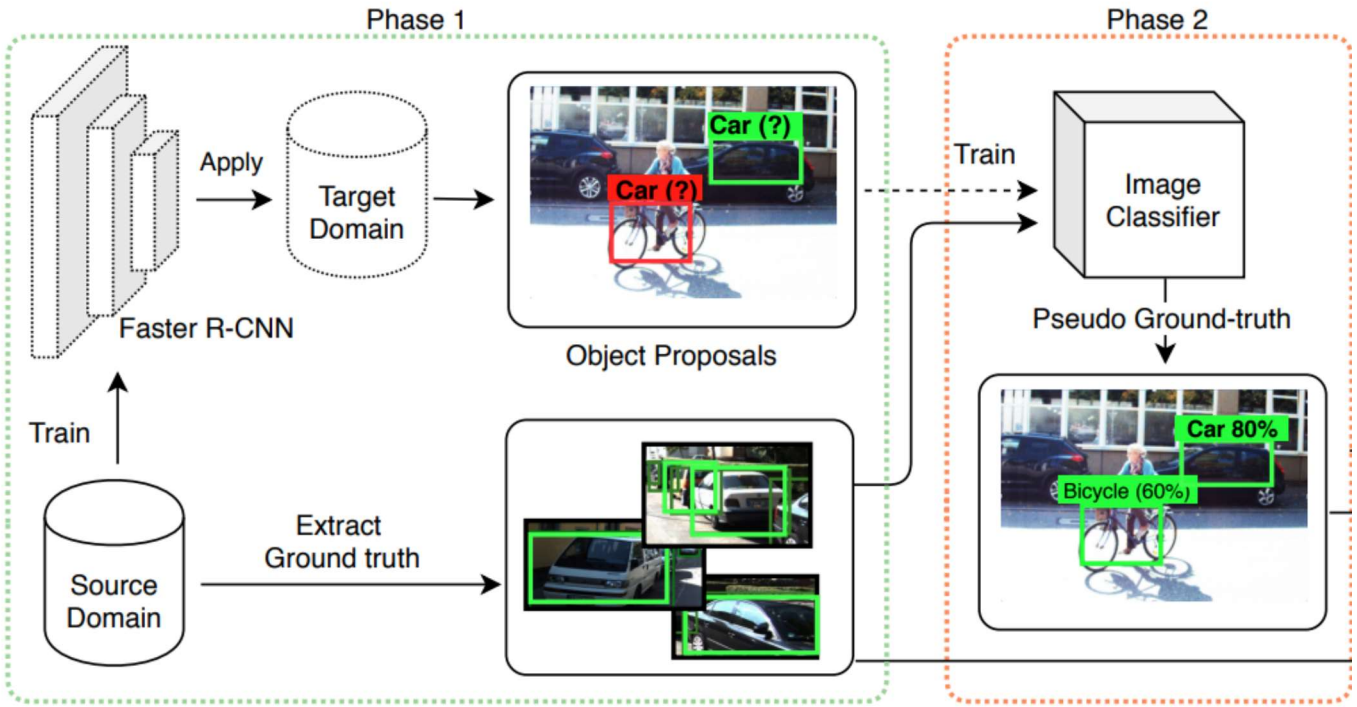
Target Domain

DOMAIN ADAPTIVE OBJECT DETECTION



Extract object proposals
in the target domain

DOMAIN ADAPTIVE OBJECT DETECTION



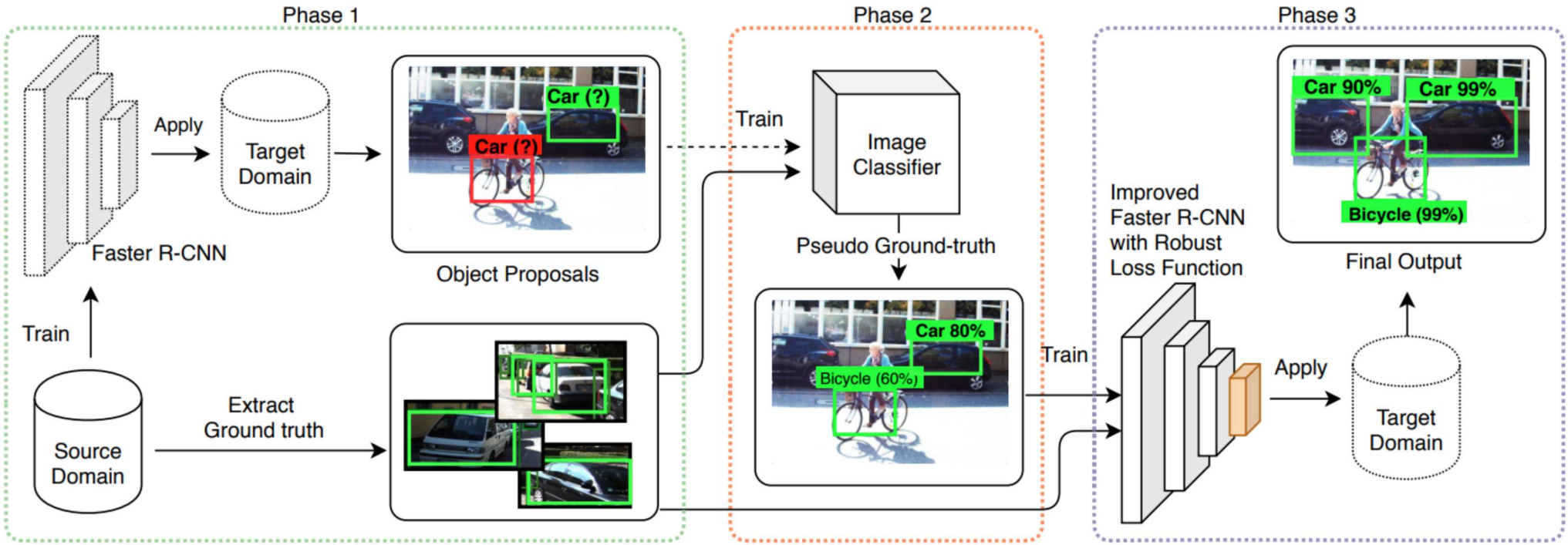
Extract object proposals in the target domain

Train image classifier: $p_{img}(t_C|x,b)$

We can mix other clues:

- motion,
- stronger classifiers
- geometry

DOMAIN ADAPTIVE OBJECT DETECTION



Extract object proposals in the target domain

Train image classifier: $p_{img}(t_C|x,b)$

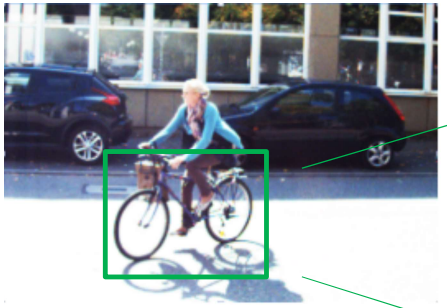
We can mix other clues:
- motion,
- stronger classifiers
- geometry

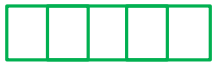
Train a noise-robust object detection model

NOISE ROBUST OBJECT DETECTION

Robust Faster R-CNN

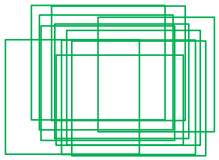
Faster R-CNN





$p_{cls}(t_c|x,b)$

Categorical Distribution



$p_{loc}(t_l|x,b)$

Normal Distribution

Faster R-CNN as two models:

1. Region Classification model
2. Object localization model

Classification Error Correction

$$p_{aux}(t_C|x,b) \coloneqq p_{img}(t_C|x,b)$$

$$q(t_c|x,b) \propto \left(p_{cls}(t_c|x,b)\; p_{img}^{\alpha}(t_C|x,b)\right)^{\frac{1}{\alpha+1}}$$

Guide the classification model with the auxiliary image classifier from phase 2

Bounding Box Refinement

$$p_{aux}(t_l|x,b) \coloneqq \mathcal{N}(b,\sigma I)$$

$$q(t_l|x,b) \propto \left(p_{loc}(t_l|x,b)\; p_{aux}^{\alpha}(t_l|x,b)\right)^{\frac{1}{\alpha+1}}$$

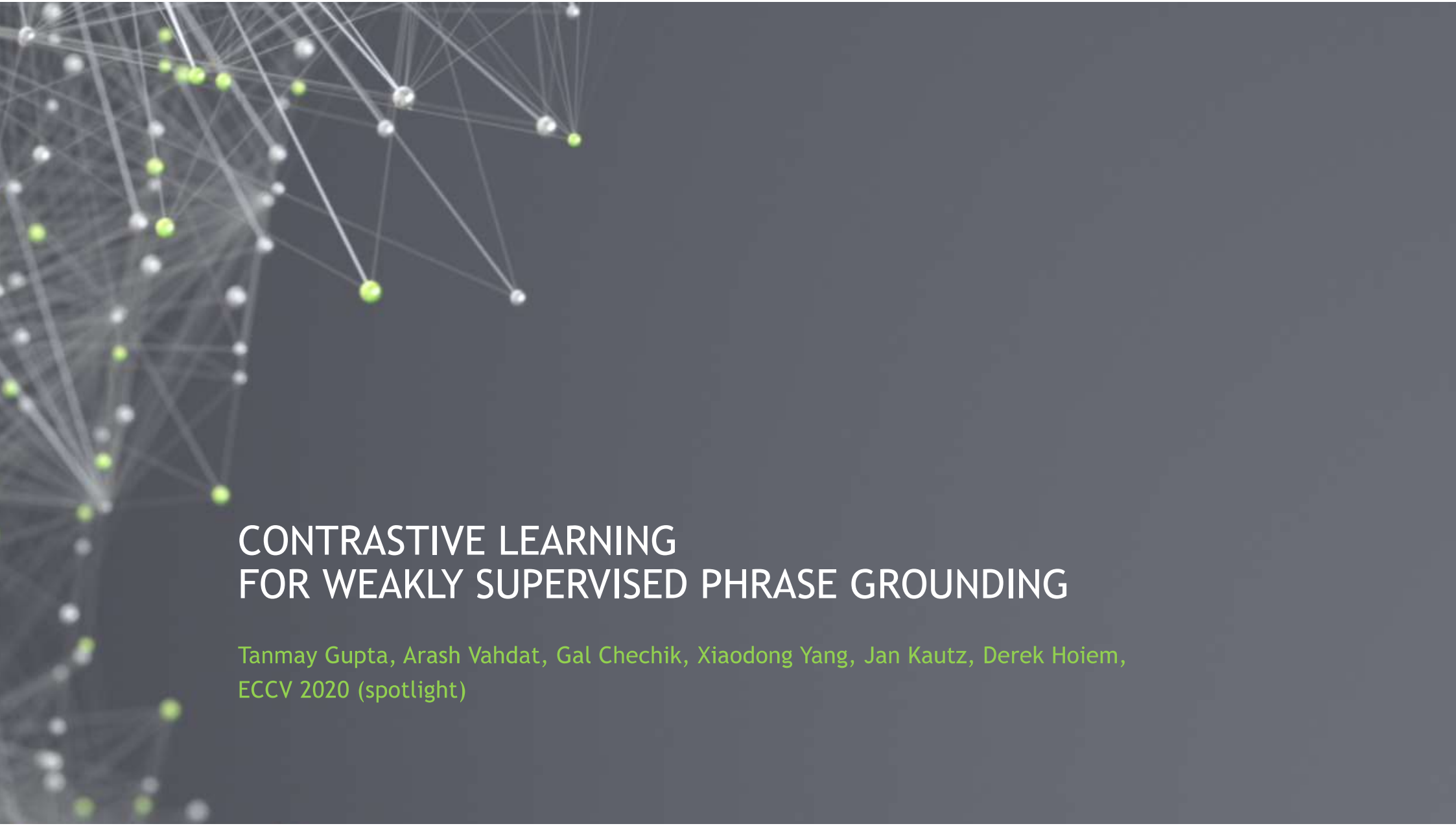
We show that the geometric mean of two Normal distributions is also Normal

Allow small refinement around the original bounding boxes

TAKE-HOME MESSAGE

Domain Adaptive Object Detection as Training with Noisy Labels

- Semi-supervised domain adaptation can be modeled as training with noisy labels
  - Recent techniques such as self-training with noisy students explored this for image classification [28]
- In object detection, auxiliary distribution can be constructed using image classifiers
- We can use the same self-correction principal for both categorical classifiers as well as Normally distributed object localization models



CONTRASTIVE LEARNING FOR WEAKLY SUPERVISED PHRASE GROUNDING

Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, Derek Hoiem,

ECCV 2020 (spotlight)

TEXT AS NOISY LABELS

Limitless Labels in a Labelless World

Text and Images are abundant on the Web

Many early works in weakly supervised learning focused on “webly supervised” vision models [19]

(1) Recently, there has been tremendous progress in unsupervised representation learning for natural language:

- BERT [20], GPT-3 [21], …

(2) In computer vision, contrastive learning is showing a promising performance for representation learning:

- Deep InfoMax [22], InfoNCE [23], MoCo [24], SimCLR [25], …

Can we use contrastive learning for training a deep model from image captions?

WEAKLY SUPERVISED PHRASE GROUNDING

The Problem of Associating Image Regions to Caption Words

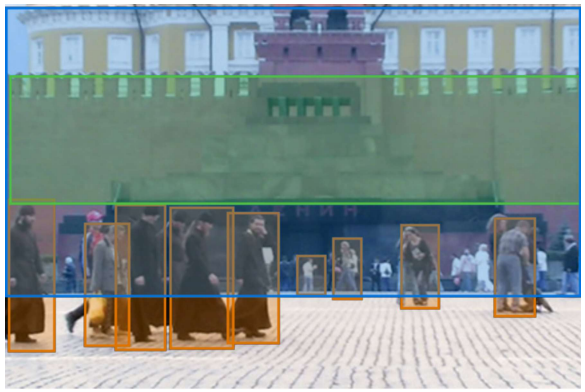
1. A town square of cobblestones with many people walking through it .
2. A city square with a large red wall and people walking about .
3. Average , everyday people walk by a facility .

1. The little girl is happily making her way past the yellow tables and benches .
2. Somebody is reaching out their arm to a little girl that is about to walk by .
3. The smiling girl runs down the aisle between the crowd-filled tables .

1. The skateboarder is leaping in the air in front of the telegraph pole .
2. Skateboarder in jeans and t-shirt performing jump .
3. A skateboarder in the air in front of wires .

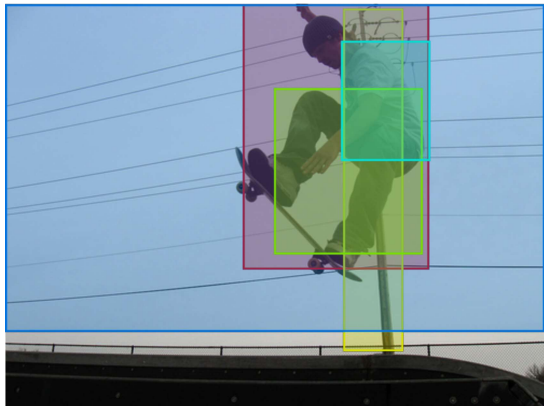
WEAKLY SUPERVISED PHRASE GROUNDING

The Problem of Associating Image Regions to Caption Words



1. A town square of cobblestones with many people walking through it .
2. A city square with a large red wall and people walking about .
3. Average , everyday people walk by a facility .

1. The little girl is happily making her way past the yellow tables and benches .
2. Somebody is reaching out their arm to a little girl that is about to walk by .
3. The smiling girl runs down the aisle between the crowd-filled tables .



1. The skateboarder is leaping in the air in front of the telegraph pole .
2. Skateboarder in jeans and t-shirt performing jump .
3. A skateboarder in the air in front of wires .

WEAKLY SUPERVISED PHRASE GROUNDING

The Problem of Associating Image Regions to Caption Words

1. A town square of cobblestones with many people walking through it .
2. A city square with a large red wall and people walking about .
3. Average , everyday people walk by a facility .

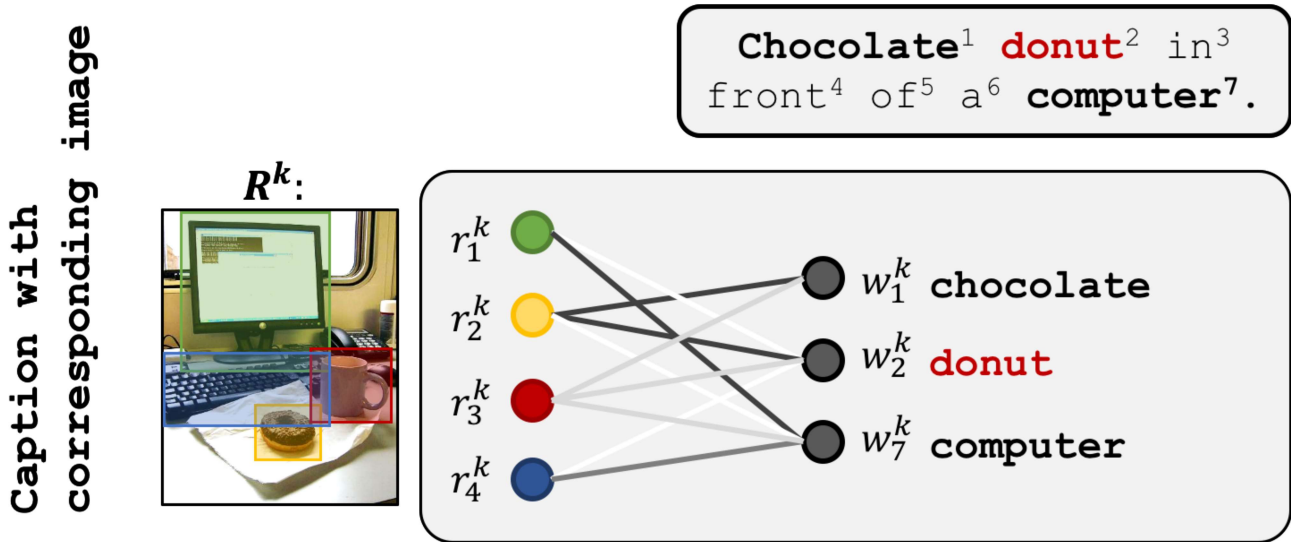
1. The little girl is happily making her way past the yellow tables and benches .
2. Somebody is reaching out their arm to a little girl that is about to walk by .
3. The smiling girl runs down the aisle between the crowd-filled tables .

1. The skateboarder is leaping in the air in front of the telegraph pole .
2. Skateboarder in jeans and t-shirt performing jump .
3. A skateboarder in the air in front of wires .

PHRASE GROUNDING WITH ATTENTION

Pretrained models

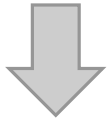
- Pretrained **object detector** extracts region proposals and their features
- Pretrained **language model** extracts contextualized word representations
- Regions assigned to words through a **learnable attention mechanism**



MUTUAL INFORMATION

Estimating mutual information between a set of image regions and caption words requires aligning regions and words.

Maximize InfoNCE [23] bound with respect to parameters of region-word attention



Attention mechanism learns grounding

Contrastive Training

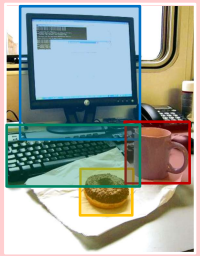
$\mathcal{L}_{img}(\theta)$

Caption with corresponding image

Same caption; different image

Chocolate[1] donut[2] in[3] front[4] of[5] a[6] computer[7].

R^k:



r_1^k r_2^k r_3^k r_4^k

$\phi_\theta(R^k, w_j^k)$

w_1^k chocolate

w_2^k donut

w_7^k computer

$R^{k'}$:

$r_1^{k'}$ $r_2^{k'}$ $r_3^{k'}$ $r_4^{k'}$ $r_5^{k'}$

$\phi_\theta(R^{k'}, w_j^k)$

w_1^k chocolate

w_2^k donut

w_7^k computer

$$\min_\theta -\log\left(\frac{e^{\phi_\theta(R^+, w^+)}}{e^{\phi_\theta(R^+, w^+)} + \sum_{(R^-, w^-)\in N} e^{\phi_\theta(R^-, w^-)}}\right)$$

$\phi_\theta(R, w)$: Compatibility between set of region features R from an image and contextualized word representation w that uses the region-word attention.

$\min_\theta \mathcal{L}_{img}(\theta)$

Contrastive Training

$\mathcal{L}_{img}(\theta)$

$\mathcal{L}_{lang}(\theta)$

Caption with corresponding image

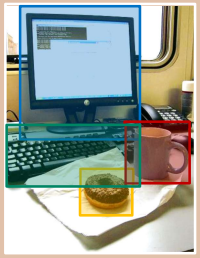
Same caption; different image

Same image; context-preserving negative caption

Chocolate[1] donut[2] in[3] front[4] of[5] a[6] computer[7].

Chocolate[1] cookie[2] in[3] front[4] of[5] a[6] computer[7].

R^k:



$\phi_\theta(R^k, w_j^k)$

r_1^k r_2^k r_3^k r_4^k

w_1^k chocolate

w_2^k donut

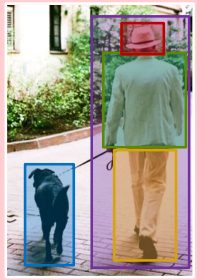
w_7^k computer

$\phi_\theta(R^k, w_j^{k^-})$

r_1^k r_2^k r_3^k r_4^k

$w_2^{k^-}$ cookie

$R^{k'}$:



$\phi_\theta(R^{k'}, w_j^k)$

$r_1^{k'}$ $r_2^{k'}$ $r_3^{k'}$ $r_4^{k'}$ $r_5^{k'}$

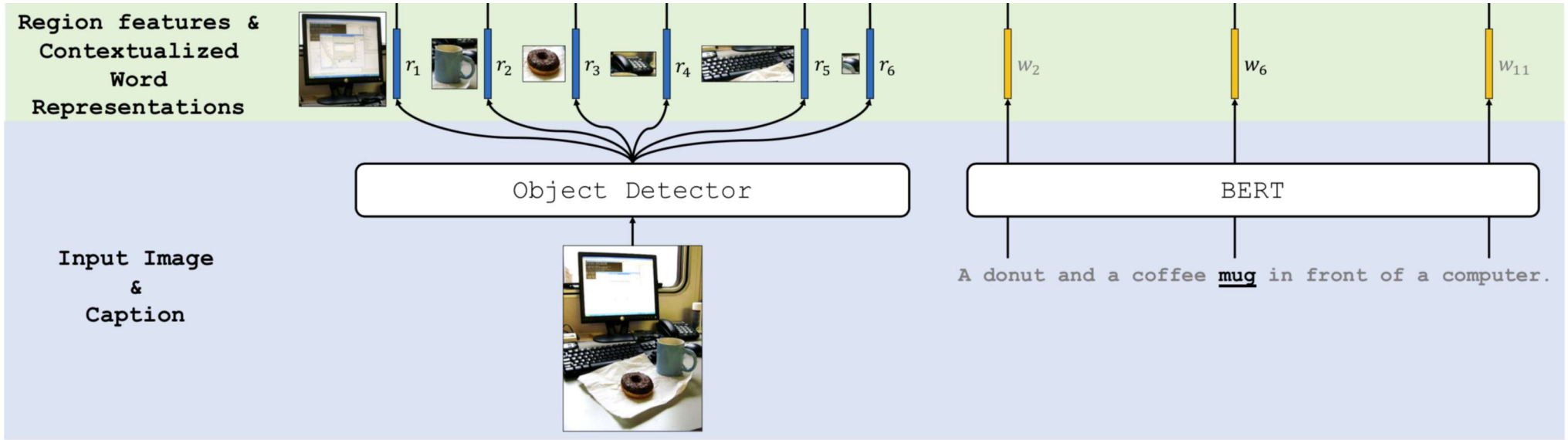
w_1^k chocolate

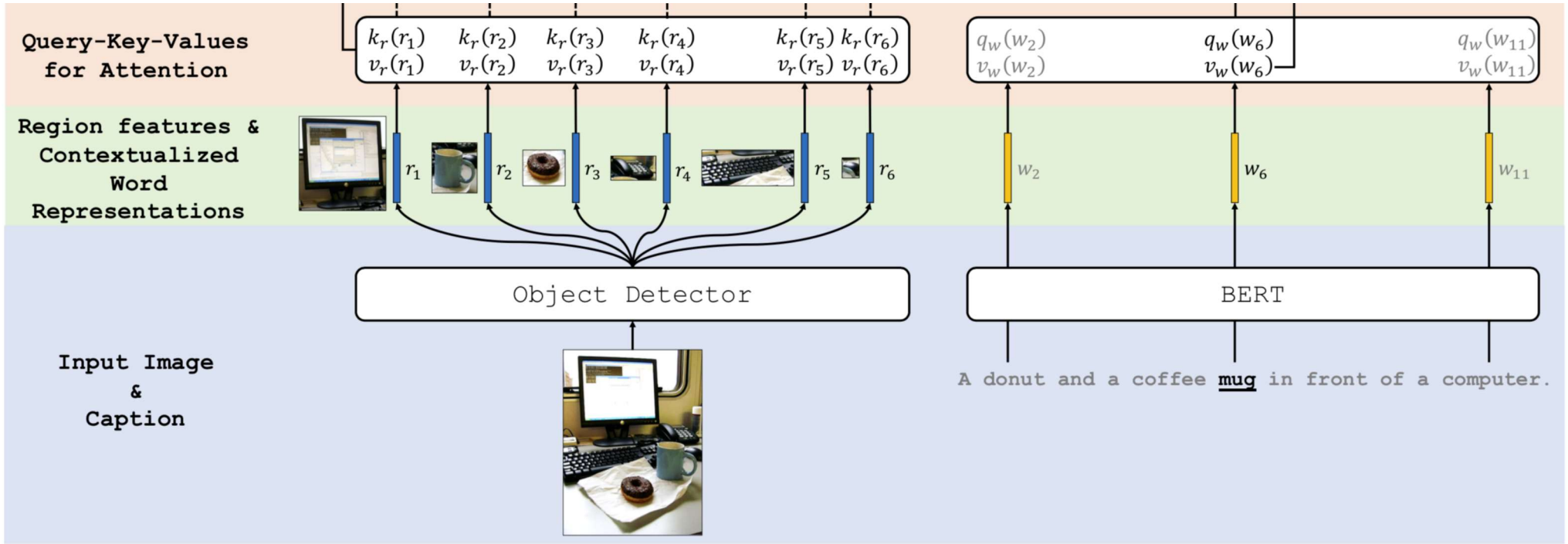
w_2^k donut

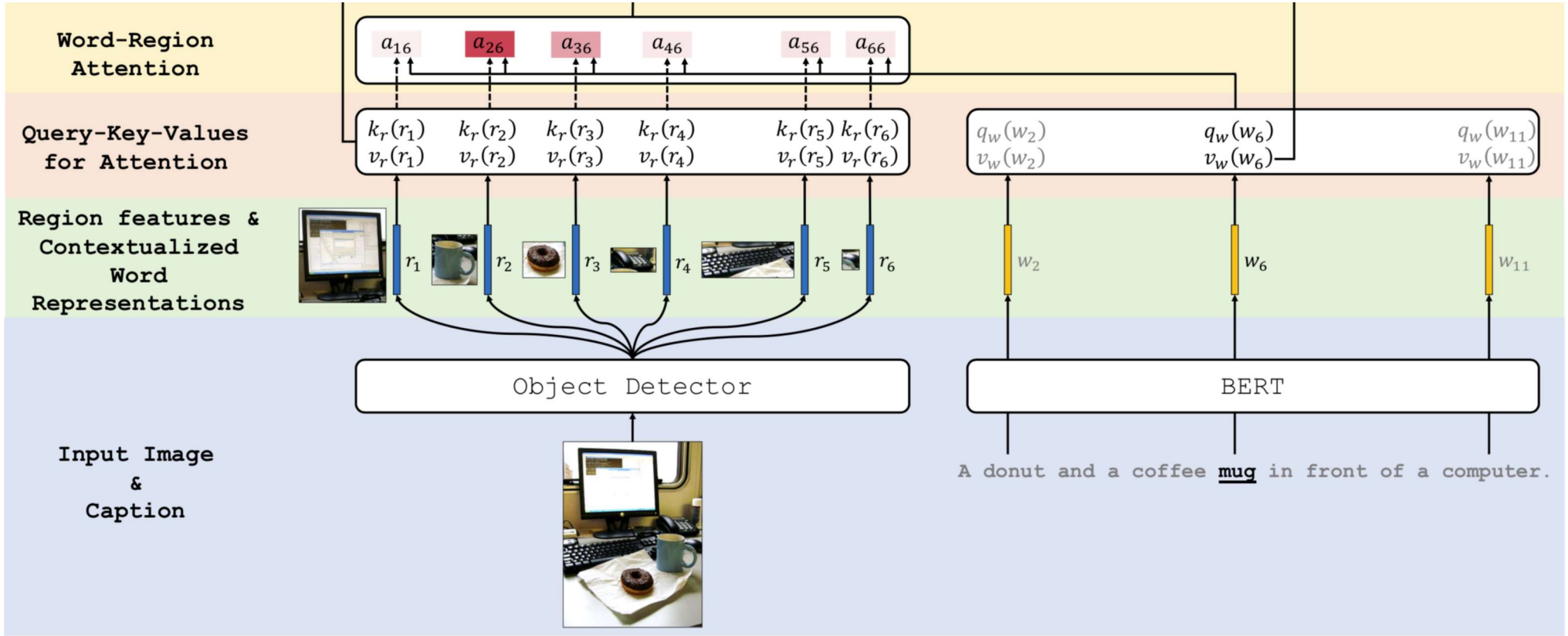
w_7^k computer

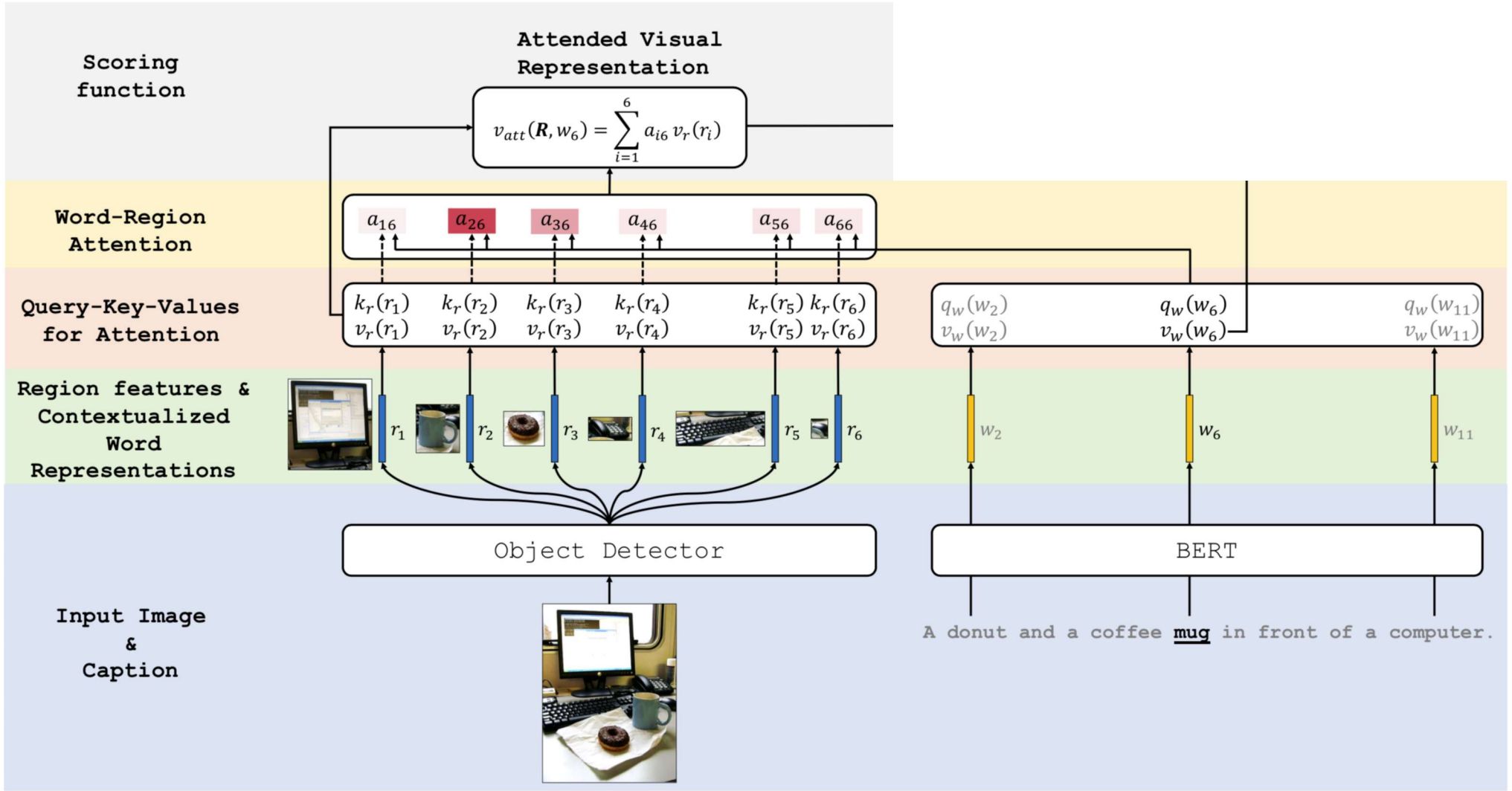
$\phi_\theta(R, w)$: Compatibility between set of region features R from an image and contextualized word representation w that uses the region-word attention.

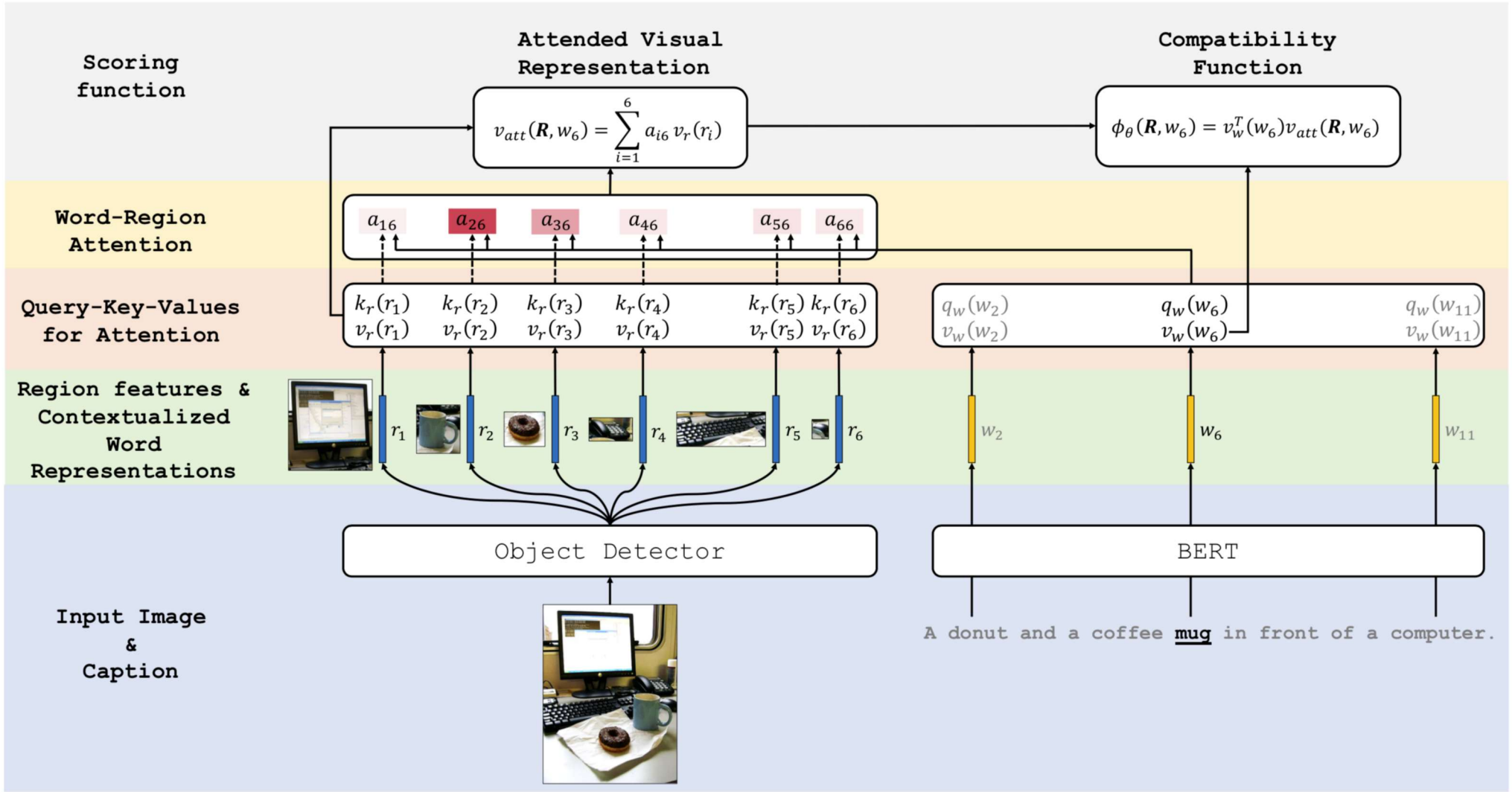
$$\min_\theta \mathcal{L}_{img}(\theta) + \mathcal{L}_{lang}(\theta)$$


Region features &
Contextualized
Word
Representations
r_1
r_2
r_3
r_4
r_5
r_6
w_2
w_6
w_{11}
Object Detector
BERT
Input Image
&
Caption
A donut and a coffee mug in front of a computer.


Query-Key-Values for Attention
Region features & Contextualized Word Representations
Input Image & Caption
$k_r(r_1)$ $v_r(r_1)$
$k_r(r_2)$ $v_r(r_2)$
$k_r(r_3)$ $v_r(r_3)$
$k_r(r_4)$ $v_r(r_4)$
$k_r(r_5)$ $v_r(r_5)$
$k_r(r_6)$ $v_r(r_6)$
r_1 r_2 r_3 r_4 r_5 r_6
Object Detector
$q_w(w_2)$ $v_w(w_2)$
$q_w(w_6)$ $v_w(w_6)$
$q_w(w_{11})$ $v_w(w_{11})$
w_2 w_6 w_{11}
BERT
A donut and a coffee mug in front of a computer.


Word-Region Attention
a_{16} a_{26} a_{36} a_{46} a_{56} a_{66}
Query-Key-Values for Attention
$k_r(r_1)$ $v_r(r_1)$
$k_r(r_2)$ $v_r(r_2)$
$k_r(r_3)$ $v_r(r_3)$
$k_r(r_4)$ $v_r(r_4)$
$k_r(r_5)$ $v_r(r_5)$
$k_r(r_6)$ $v_r(r_6)$
$q_w(w_2)$ $v_w(w_2)$
$q_w(w_6)$ $v_w(w_6)$
$q_w(w_{11})$ $v_w(w_{11})$
Region features & Contextualized Word Representations
r_1 r_2 r_3 r_4 r_5 r_6
w_2 w_6 w_{11}
Object Detector
BERT
Input Image & Caption
A donut and a coffee mug in front of a computer.


Scoring function
Attended Visual Representation
$v_{att}(\boldsymbol{R}, w_6) = \sum_{i=1}^{6} a_{i6}\, v_r(r_i)$
Word-Region Attention
a_{16} a_{26} a_{36} a_{46} a_{56} a_{66}
Query-Key-Values for Attention
$k_r(r_1)$ $v_r(r_1)$
$k_r(r_2)$ $v_r(r_2)$
$k_r(r_3)$ $v_r(r_3)$
$k_r(r_4)$ $v_r(r_4)$
$k_r(r_5)$ $v_r(r_5)$
$k_r(r_6)$ $v_r(r_6)$
$q_w(w_2)$ $v_w(w_2)$
$q_w(w_6)$ $v_w(w_6)$
$q_w(w_{11})$ $v_w(w_{11})$
Region features & Contextualized Word Representations
r_1 r_2 r_3 r_4 r_5 r_6
w_2 w_6 w_{11}
Object Detector
BERT
Input Image & Caption
A donut and a coffee mug in front of a computer.


Scoring function
Attended Visual Representation
Compatibility Function
$v_{att}(\boldsymbol{R}, w_6) = \sum_{i=1}^{6} a_{i6}\, v_r(r_i)$
$\phi_\theta(\boldsymbol{R}, w_6) = v_w^T(w_6) v_{att}(\boldsymbol{R}, w_6)$
Word-Region Attention
a_{16} a_{26} a_{36} a_{46} a_{56} a_{66}
Query-Key-Values for Attention
$k_r(r_1)$ $v_r(r_1)$
$k_r(r_2)$ $v_r(r_2)$
$k_r(r_3)$ $v_r(r_3)$
$k_r(r_4)$ $v_r(r_4)$
$k_r(r_5)$ $v_r(r_5)$
$k_r(r_6)$ $v_r(r_6)$
$q_w(w_2)$ $v_w(w_2)$
$q_w(w_6)$ $v_w(w_6)$
$q_w(w_{11})$ $v_w(w_{11})$
Region features & Contextualized Word Representations
r_1 r_2 r_3 r_4 r_5 r_6
w_2 w_6 w_{11}
Object Detector
BERT
Input Image & Caption
A donut and a coffee mug in front of a computer.

constructing hard negative captions with BERT

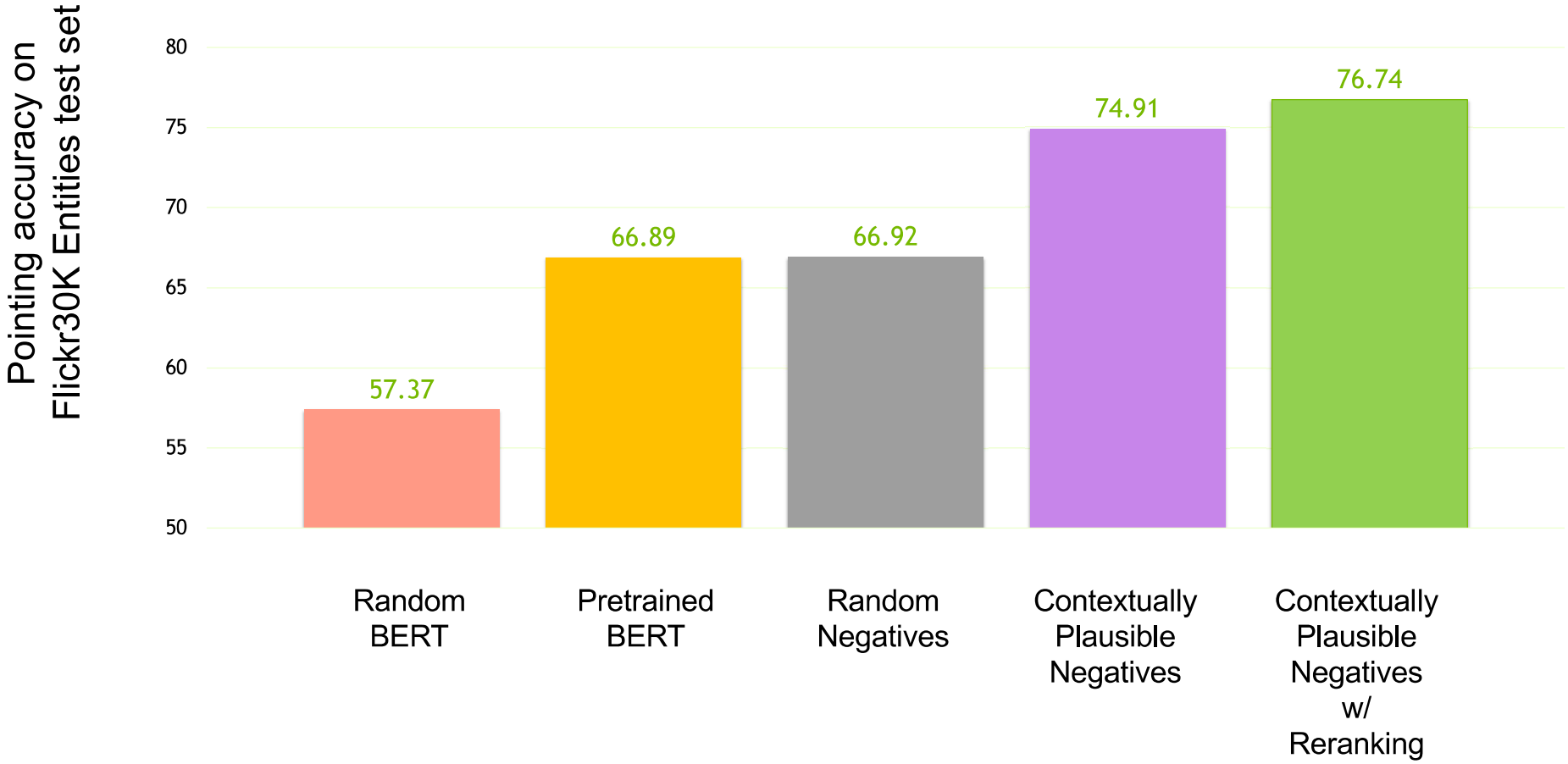
Use a pretrained language model to construct effective negative captions for contrastive learning through word substitutions.

A man in a blue jumpsuit stands next to a red van pulling a trailer.

bike
sedan
horse
cart

} Contextually plausible negative word substitutes

Gains from Language Modeling



TAKE-HOME MESSAGE

Contrastive Learning for Vision-Text Applications

- Recent advances in natural language processing have brought us some of the most advanced language models with rich representation of words/sentences
- We showed that contrastive learning can discover object location when maximizing a lower bound on mutual information
- We showed how language models like BERT can help us extract plausible negative word substitutes

For code and other project material visit
http://tanmaygupta.info/info-ground/

CONCLUSIONS

Final Thoughts

Label Noise is present in real-world problems!

Self-correction mechanism to handle label noise

Small progress in this area despite its importance

Noisy Labels

Contrastive learning is an effective framework for learning from text data in a weakly supervised setting

Recent language models contain rich representation of natural language

More work in this space!

Contrastive Learning

Big gap between theory and practice currently!

The existing theory only considers multiclass classification

Theoretical Understanding

THANKS

@ArashVahdat
avahdat@nvidia.com

REFERENCES

[1] Natarajan, N., Dhillon, I. S., Ravikumar, P. K., and Tewari, A. Learning with noisy labels. In NeurIPS, 2013

[2] Patrini, G., Rozza, A., Menon, A., Nock, R., and Qu, L. Making neural networks robust to label noise: a loss correction approach. In CVPR, 2017.

[3] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In CVPR, 2016.

[4] Vahdat, A. Toward robustness against label noise in training deep discriminative neural networks. In NeurIPS, 2017.

[5] Veit, A., Alldrin, N., Chechik, G., Krasin, I., Gupta, A., and Belongie, S. Learning from noisy large-scale datasets with minimal supervision. In CVPR, 2017.

[6] Li, Y., Yang, J., Song, Y., Cao, L., Luo, J., and Li, J. Learning from noisy labels with distillation. In ICCV, 2017b.

[7] Ghosh, A., Kumar, H., and Sastry, P. Robust loss functions under label noise for deep neural networks. In AAAI, 2017.

[8] Zhang, Z. and Sabuncu, M. R. Generalized cross entropy loss for training deep neural networks with noisy labels. In NeurIPS, 2018.

[9] Ma X, Huang H, Wang Y, Erfani SR, Bailey J. Normalized Loss Functions for Deep Learning with Noisy Labels. In ICML 2020

REFERENCES

[10] Ren M, Zeng W, Yang B, Urtasun R. Learning to Reweight Examples for Robust Deep Learning. In ICML 2018

[11] Jiang, L., Zhou, Z., Leung, T., Li, L.-J., and Fei-Fei, L. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In ICML, 2018.

[12] Malach, E. and Shalev-Shwartz, S. Decoupling" when to update" from" how to update". In NeurIPS, 2017

[13] Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., and Sugiyama, M. Co-teaching: robust training deep neural networks with extremely noisy labels. In NeurIPS, 2018.

[14] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, The journal of machine learning research, vol. 15, no. 1, pp. 1929–1958, 2014.

[15] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, ,mixup: Beyond Empirical Risk Minimization, In ICLR, 2018.

[16] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In ACM transactions on graphics (TOG). ACM, 2004.

[17] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In IEEE International Conference on Computer Vision (ICCV), 2015.

REFERENCES

[18] George Papandreou, Liang-Chieh Chen, Kevin P. Murphy, and Alan L. Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In ICCV, 2015.

[19] Chen X, Gupta A. Webly supervised learning of convolutional networks. In ICCV 2015

[20] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

[21] Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S. Language models are few-shot learners. 2020.

[22] Hjelm, R.D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., Bengio, Y., Learning deep representations by mutual information estimation and maximization. In ICLR, 2019.

[23] Oord, A.v.d., Li, Y., Vinyals, O., Representation learning with contrastive predictive coding. arXiv 2018.

[24] He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., Momentum contrast for unsupervised visual representation learning. In CVPR, 2020.

[25] Chen, T., Kornblith, S., Norouzi, M., Hinton, G., A simple framework for contrastive learning of visual representations. In ICML, 2020.

REFERENCES

[26] Algan G, Ulusoy I. Image classification with deep learning in the presence of noisy labels: A survey. arXiv preprint arXiv:1912.05170. 2019 Dec 11.

[27] Bearman A, Russakovsky O, Ferrari V, Fei-Fei L. What's the point: Semantic segmentation with point supervision. In ECCV, 2016

[28] Xie Q, Luong MT, Hovy E, Le QV. Self-training with noisy student improves imagenet classification. In CVPR, 2020