Learn transferable features for semantic image segmentation in the presence of label noise

First Author
Institution1
Institution1 address
firstauthor@i1.org

Second Author
Institution2
First line of institution2 address
secondauthor@i2.org

July 20, 2017

Abstract

The ABSTRACT HERE

1 Introduction

The lack of "gold standard" annotations becomes the bottleneck for semantic segmentation.
The state-of-art Convolutional Neural Networks (CNNs) based semantic image segmentation models usually rely on the present pre-trained convolutional filters. [10] A typical method to pre-train the convolutional filters is to train a classification model with the large-scale ILSVRC dataset [17] which contains 1000 categories and 1.2 million images. However, this method constrains the semantic image segmentation models to have the same CNN architecture as the image classification models. The CNN design for semantic image segmentation does not necessarily follow the design of image classification architectures. The segmentation models need both global and local information to generate fine segmentations, whereas the classification models care less about local information that gives information about the object localization. For instance, the presence of the max-pooling layers enable the following convolutional filters to have larger receptive fields but, at the same time, reduce the resolution of the features. Additional upsampling layers, can recover the shape of the output segmentation but cannot fully recover the information thrown away. This first-pooling-and-then-upsampling pipeline can result in coarse segmentation output [2] with non-shape boundaries and blob-like shapes. [1] The coarse output can be refined with Conditional Random Field (CRF) inference[20, 2].
This paragraph should explain the difficulties for collecting perfect segmentation annotations on a large scale Alternatively, one can also pre-train CNNs by semantic segmentation tasks directly. Given the "data-hungry" nature of CNNs, one major challenge to train "good representation" with semantic segmentation tasks is the lack of "gold standard" segmentation on a large scale, [2] The largest segmentation dataset, COCO2014[9], contains annotations for only 164,000 training images, smaller by a factor of 10 than the ILSVRC2012 dataset with 1,281,167 labeled training images. Millions of images are available through Flikr, ImageNet and many other sources on the Internet but only a few of them [5, 13, 9] have been well-annotated for semantic segmentation tasks. Most state-of-art neural networks for semantic segmentation [10, 20] assume the existence of perfect segmentations with all instance annotated (exhaustive) and no misannotated instance (precise). However, it is natural for human beings to make mistakes due to the lack of expertise, the inherent ambiguity of tasks or unconscious bias. Enormous efforts are required to correct the mistakes made, including double-checking the annotations over and over again

[1] M: Also for this claim, I think we need a reference or something like that. Or other proof indeed... In addition, I wonder whether we can explain why this may be the case? J: Yes, I can refer to a paragraph in the intro of CRFasRNN and enrich the discussion a bit.

[2] M: For the rest a bit vague... what do we really mean by suffer? Maybe this becomes clear later in the intro...? J: Refrased.

and ensembling opinions from multiple annotators. Otherwise, it may lead to annotations that contain misannotated instances, unannotated instances and misclassification of instances. This makes collecting such perfect segmentations manually on a large scale become both expensive and time-consuming. [9] [3] In domains like medical imaging, collecting such annotations can be even more expensive because of the high compensation and workload for the medical experts. In addition, there are some freely available labels which may or may not be accurate. One exemple for this is the use of digital maps such as OpenStreetMap to annotate aerial images for segmentation, but the such dataset constructed from maps suffer from omission noise and registration problems.[12] If the "gold standard" annotation assumption can be relaxed somehow, collecting segmentation annotations would be easier to scale by avoiding tedious inspection and correction. But that requires methods that can still learn transferable features in the presence of annotation errors. [4]

Briefly formulate the problem.

Semantic Segmentation tasks can be considered as per-pixel classification problem. Each of the pixels is assigned a label of either 0, indicating a background pixel, or $k \in 1, \ldots, K$ denoting a foreground pixel corresponding to an instance from one of the K categories. The aforementioned errors can be interpreted by the pixel label flipping: *misannotation* is label flipped from 0 to k, *inexaustive annotation* flipped from k to 0, and *misclassification* flipped from i to j, where $i, j, k \in 1, \ldots, K$. The misannotated instances are assumed to be visually distinguishable against the background and have natural semantic meaning. All the label flips have one instance as the minimum fliping unit.

Instance Misannotation and Instance Misclassification

One hypothesis made in this work is that the *misannotated and misclassified* instances can still provide information for training multi-scaled features that are *transferable*[19] to the other datasets and categories. Supposing a dog toy is wrongly annotated as a dog, given that the "dog" is one of the target categories while the "toy" is not, the error would introduce bias into the last layer but not necessarily into the first convolutional layers. The high-level features were found more dependent on a particular category, i.e. more *specific*, than the low-level features, whereas the low-level features were found less category-dependent, i.e. more *general*. [19] We wanted to explore whether the "generality" of low-level features contributes to the annotation error robustness when we transfer the learned features to a new dataset with new categories. That leads us to the following research question:

1. How do misannotation and misclassification of instances influence the "transferability" of the learned features respectively?

We experimented, in Section 3, how transferable the learned features are in two special cases for the two types of annotation error:

instance misannotation all instances from the non-target categories are misannotated as instance from the target category

instance misclassification the classes for all the instances are completely randomly assigned

[5] The *transferability* of the features can be evaluated by how much they can boost the performance of training a new dataset. [19] *TODO One sentence summarizes the results.*

Inexaustive annotating

The exhaustive annotations can introduce bias to both the decoding layer and the encoding layers because they negatively contribute to the activations in all the layers. [6] The inexhaustive annotations need to be properly handled given the prior knowledge modeling the missing pattern of the annotations. Given that we believe any annotated instance provide information, all the foreground pixels that correspond to the annotated instances become reliable and the background pixels may contain both the true background pixels and object pixels unannotated. That satisfies a Positive and Unlabeled learning setup where the training dataset contains only the positive examples and unlabeled examples that are the mixed of the positive samples and negative samples.

[3] M: Do we actually have a reference for that? Or other proof? J: Yep. Microsoft's paper about the Microsoft COCO dataset actually mentioned how many working hours the crowd workers spent, etc.

[4] J: Why would this differ from training deep neural networks in the precence of annotation errors?

[5] M: So, to what extent is this the actual research question that you would like to answer? J: I think this section answers your question now.

[6] J: This argument need evidence too, or experiment/discussions in details in Section 4

Table of contents
Related works are summarized in the next section. In Section 3 we judge the possibility of learning convolutional representation with misannotations by learning to predict the pixel objectness. Section 4 explored the methods to compensate the inexaustive annotations in a Positive and Unlabeled Learning setup. Features learned by predicting the pixel objectness with inexaustive annotations were then validated with experiments described in Section 5.

2 Related work

Semantic Image Segmentation with Deep Neural Nets J. Long et.al.[10] defined a skip architecture to combine semantic information from a deep, coarse layer with appearance information from a shallow, fine layer to produce accurate and detailed segmentations and transfered the learned representations from the contemporray classification networks into fully convolutional networks. L. Chen et.al.[2] removed the last few max pooling layers of the CNNs and upsampled the corresponding filters to avoid the reduced feature resolution by the pooling layers. An additional fully connected Conditional Random Field (CRF) was added to refine the coarse last layer output for better localization performance. S. Zheng et.al.[20] integrate the CRFs-based probabilistic graphical modeling with CNNs in an end-to-end framework.

Transfer Learning The first-layer features in the modern CNNs are often observed converging to either Gabor filters or color blobs, regardless the exact learning objectives and the training dataset. This pheonomen is called the *generality* of the first-layer features. By contrast, the last-layer features depend significantly on the learning objective and dataset and they are called *specific*.[19] Yosinski et.al.[19] studied the transition from general to specific for the features in the intermediate layers by measuring how much the tranformed pre-trained features boost the fine-tuning performance on a new dataset, i.e. the "transferability" of the features. They found features from several bottom layers, not only the first layer, were "transferable" to a new dataset, meaning that they were not specific for a particular category but were shared across categories. This discovery led to our hypothesis that the general features can be more robust to annotation errors, either object misannotation or instance misclassidication, than the specific features. We used the same measure as Yosinki et.al. did to quantitatively measure the "transferability" of the features and reported the robustness to instance misannotation and instance misclassification for the learned representation.

Unsupervised and semi-supervised pre-training Apart from supervised pre-training, one can also obtain the pre-trained features with unsupervised or semi-supervised pre-training. Vincent et al.[18] trained multilevel representations robust to corrupted inputs with stacked denoising auto-encoders. Masci et al.[11] presented a stacked convolutional auto-encoder unsupervised pre-training for hierarchical feature extraction. Hinton et al.[7] proposed a greedy learning algorithm to train *deep belief nets* one layer at a time in an unsupervised generative training setup. Lee et al.[8] presented a hierarchical generative model, *convolutional deep belief network*, to learn hierachical representation. A couple of works[4, 3, 1] highlighted the advantage of unsupervised pre-training compared to the random initialization, connecting unsupervised pre-training to either a norm of regularization or better disentangling the sample variations. However, xavier initialization[6] shorten the gap between the unsupervised pre-trained representations and random initialization. Unsupervised deep representation learning is in general not comparable to supervised representation learning given large scale dataset is available. [7] Supervised pre-training even in the presence of annotation errors is expected to outperform unsupervised pre-training because more information is provided. [8] *TODO Semi-supervised representation learning*

Deep Learning with Noisy Labels *Robustness analysis* Deep Learning is Robust to Massive Label Noise [16]

Entropy regularization Training deep neural networks on noisy labels with bootstrapping [15] Regularizing Neural Networks by Penalizing Confident Output Distributions [14]

Positive and Unlabeled Learning

[7] J: Citation needed.
[8] J: Evidence/Experiment needed!

Initial Representation	mean IU (aerospace, bicycle, bird, boat, bottle)	mean IU (bus, car, cat, chair, cow)	mean IU (dining table, dog, horse, motorbike, person)	mean IU (potted plant, sheep, sofa, train, TV)
ImageNetModel	0.42 ± 0.01	0.51 ± 0.01	0.49 ± 0.01	0.47 ± 0.01
SingleCategory	0.42 ± 0.01	0.51 ± 0.01	0.49 ± 0.01	0.47 ± 0.01
PixelObjectness	0.30 ± 0.02	0.35 ± 0.01	0.29 ± 0.02	0.35 ± 0.03
CompleteCategory	0.29 ± 0.01	0.36 ± 0.01	0.29 ± 0.01	0.37 ± 0.01
RandomCategory	0.29 ± 0.01	0.33 ± 0.03	0.26 ± 0.01	0.28 ± 0.01
FromScratch	0.29 ± 0.01	0.29 ± 0.03	0.27 ± 0.01	0.30 ± 0.02

Table 1: Performances of FCN with Alexnet trained to segment 5 categories from the PASCAL VOC2011 dataset with different representation initializations. *Complete-Category* is the model pre-trained to segment the other 15 categories from the PASCAL VOC2011 dataset; The *PixelObjectness* model was pre-trained to distinguish the instance against the background; The *RandomCategory* model was pre-trained to segment instances with randomly assigned categories from 1 to 15.

3 Pre-train features by learning "objectness"

4 Positive and Unlabeled Learning

One sentence summary of Positive and Unlabeled Learning

Formulation *This part should explain the Positive and Unlabeled Learning setup with mathematical representation when necessary.*

Weighted Logistic Regression *This part should discuss the linear model for observing positive conditioning on true positive and its relationship to changing the class weight.*

Exponential Loss for unlabeled examples *This part should explain why the exponential loss could perform better than the cross-entropy loss, potentially with a figure of 2D Gaussians.*

This paragraph should explain why fade-in was introduced to avoid all-positive inital prediction

This part should explain the influence of the imbalanced problem and how to overcome.

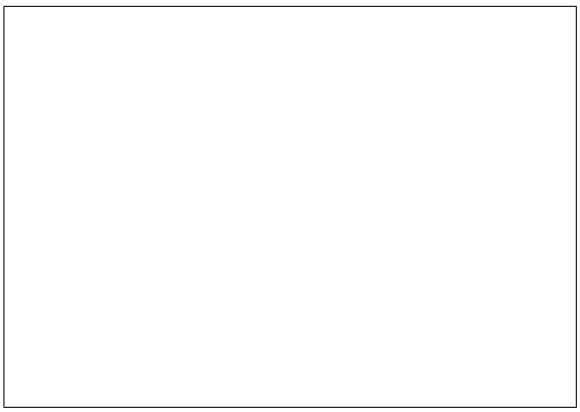

Figure 1: Visualization of first-layer features from different pre-trained models.

Annotation	Loss	acc.	prec.	rec.	F_1
Complete	CrossEntropyU.	0.87 ± 0.01	0.88 ± 0.01	0.82 ± 0.01	0.85 ± 0.01
50%(P+N)	CrossEntropyU.	0.83 ± 0.01	0.84 ± 0.01	0.78 ± 0.01	0.80 ± 0.01
50%P+U	CrossEntropyU.	0.64 ± 0.04	0.93 ± 0.08	0.34 ± 0.02	0.44 ± 0.06
50%P+U	WeightedU.	0.78 ± 0.01	0.75 ± 0.01	0.75 ± 0.01	0.76 ± 0.01
50%P+U	ExponentialU.	0.82 ± 0.01	0.86 ± 0.01	0.73 ± 0.01	0.78 ± 0.01
50%P+U	BootstrapHard	0.74	0.81	0.60	0.67
50%P+U	DropoutReg.				

Table 2: Image classification with positive examples partially annotated. The complete dataset contains images from CIFAR10 as the **positive** (P) set and images from CIFAR110 as the **negative** (N) set. The unannotated positive examples from P set construct the **unlabeled** (U) set together with the N set.

5 Results

6 Conclution

References

[1] Yoshua Bengio. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, pages 17–36, 2012.

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915*, 2016.

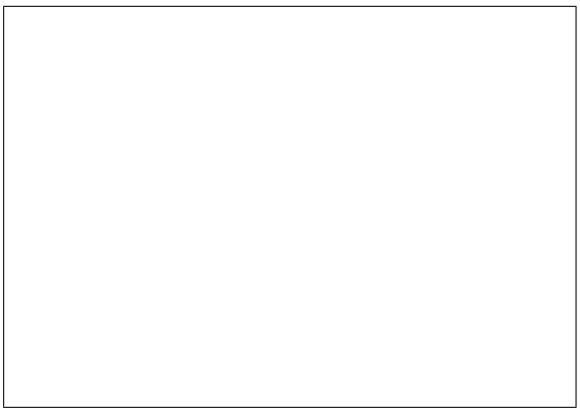

Figure 2: Varying the number of categories while pre-training the representation and the pre-trained weights were fine-tuned to segment 5 categories from the PASCAL VOC2011 dataset.

Annotation	Loss	pixel acc.	mean acc.	mean IU	f.w. IU
Complete	CrossEnt.U				
50%(P+N)	CrossEnt.U				
50%P+U	CrossEnt.U				
50%P+U	WeightedU				
50%P+U	ExponentialU				
50%P+U	BootstrapHard				
50%P+U	DropoutReg.				

Table 3: Image semantic segmentation with images contain single instance only from the PASCAL VOC2011 segmentation dataset. The complete **positive** (P) set denotes the foreground instances and the **negative** (N) set consists of the background. The unannotated instances from P set construct the **unlabeled** (U) set together with the N set.

[3] Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, 11(Feb):625–660, 2010.

[4] Dumitru Erhan, Pierre-Antoine Manzagol, Yoshua Bengio, Samy Bengio, and Pascal Vincent. The difficulty of training deep architectures and the effect of unsupervised pre-training. In *Artificial Intelligence and Statistics*, pages 153–160, 2009.

[5] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015.

[6] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256, 2010.

[7] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[8] Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th annual international conference on machine learning*, pages 609–616. ACM, 2009.

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[10] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[11] Jonathan Masci, Ueli Meier, Dan Cireşan, and Jürgen Schmidhuber. Stacked convolutional auto-encoders for hierarchical feature extraction. *Artificial Neural Networks and Machine Learning–ICANN 2011*, pages 52–59, 2011.

[12] Volodymyr Mnih and Geoffrey E Hinton. Learning to label aerial images from noisy data. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 567–574, 2012.

[13] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2014.

[14] Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017.

[15] Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*, 2014.

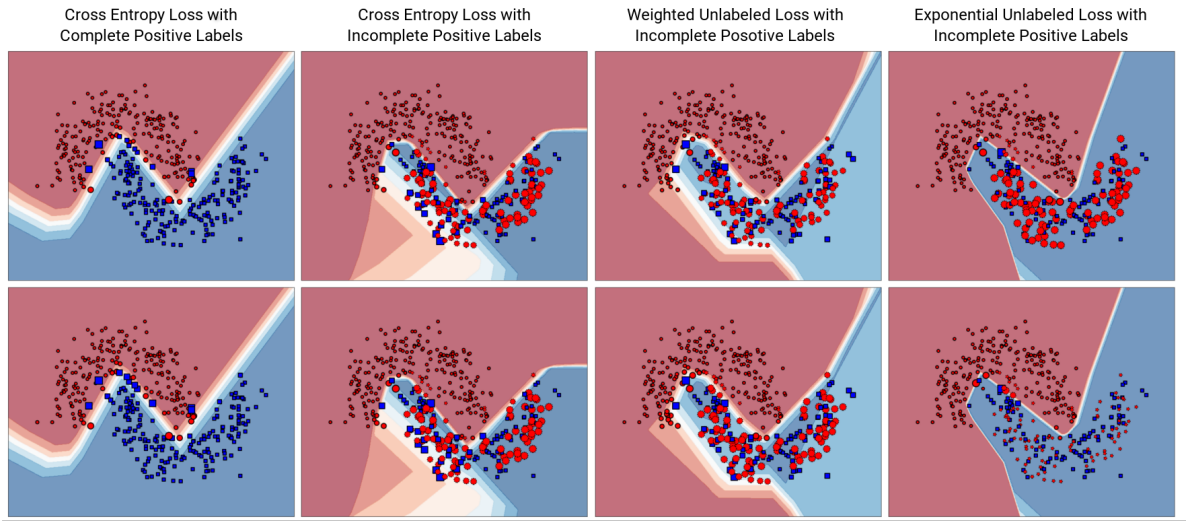

Figure 3: 2D moons dataset with non-linear separable decision boudary. Four hundreds samples per class were drawn randomly from two interleaving half circles with noises added with a minor standard deviation. A **red circle** indicates an example labelled as positive whilst a **blue square** indicates the example has a negative label. The **leftmost** figures have complete positive labels, meaning the positive and negative labels are all correct, whereas, in **the other figures** only half of the positives were correctly labelled and the rest were mixed with the negative samples. The **background colors** represent the probability for the area to be positive given by the classifier trained with the given samples and labels: **red** for high probability areas, **blue** for low probability areas and **white** for the class transition areas, i.e.decision boundaries. The **size of the markers** in the top row denotes the per-class normalized training losses and the **size of the markers** in the bottom row the per-class normalized derivatives w.r.t the output of the last layer for the trained Multilayer Perceptron (MLP) with the different losses.

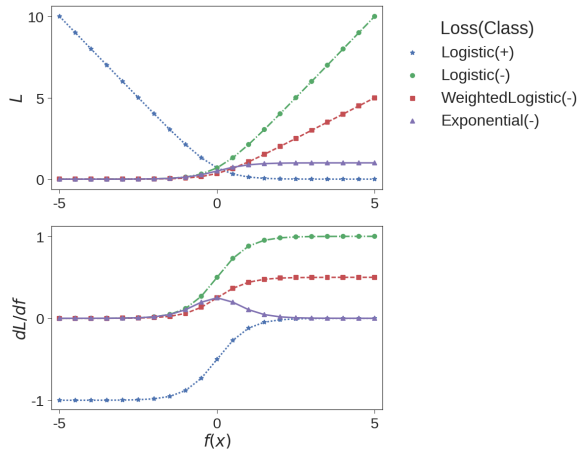

Figure 4: The Logistic Loss, Weighted Logistic Loss, Exponential Loss and their dirivatives with respect to the model output.

[16] David Rolnick, Andreas Veit, Serge Belongie, and Nir Shavit. Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694*, 2017.

[17] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[18] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(Dec):3371–3408, 2010.

[19] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014.

[20] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1529–1537, 2015.

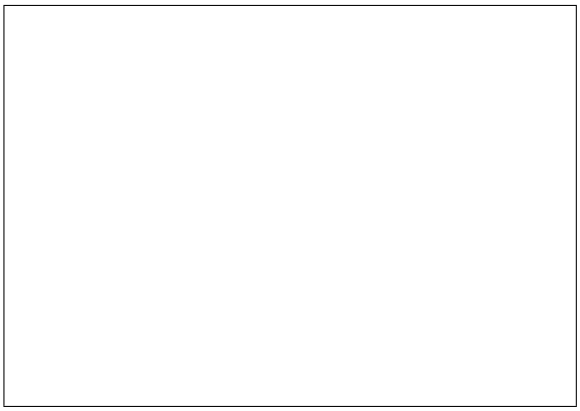

Figure 5: Varying percentage of annotated positives 10%, 20%, 50%, 80% and 100% with images from CIFAR10 as the positives and images from CIFAR110 as the negatives.