Proofpoint

Open-Set Learning Literature Review

29th October 2020

Duke

Logistics

- Sent out Mid-Year presentation invites for Wednesday 11th November 11am-12pm.
- Our team will begin building our model this week and aim to have some results by the middle of next week.
- Should we have a Proofpoint meeting next week to discuss results?
- Thursday 12th November Meeting?
- November 2nd: 360 Action Plan.
- November 11th: Mid-Year Presentation.
- November 13th: Presentation to Duke MIDS.
- November 20th: Mid-Year Paper Submission.

Duke

Annotations

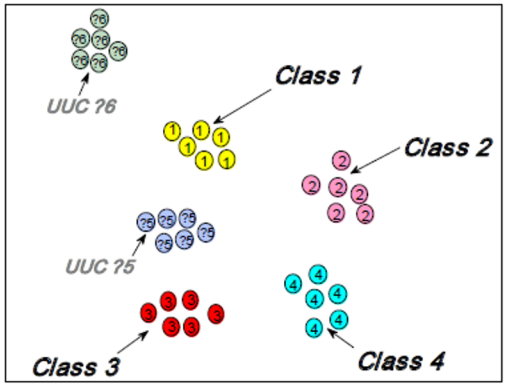
- Approx. 2700 images annotated + approx. 120 no logo images.
- Wrote a script to clean up the mismatches between annotations.
- Set up script to obtain urls for forms for 270 US government departments - https://www.usa.gov/forms.
- Have to treat each website individually to obtain pdf links.
- Pdf links can only screenshotted with headless = False.
- Majority of the forms do not have department logos on them (see appendix).
- Will work on scraping this content and others (social media, emails), over the winter / early next semester.

Duke

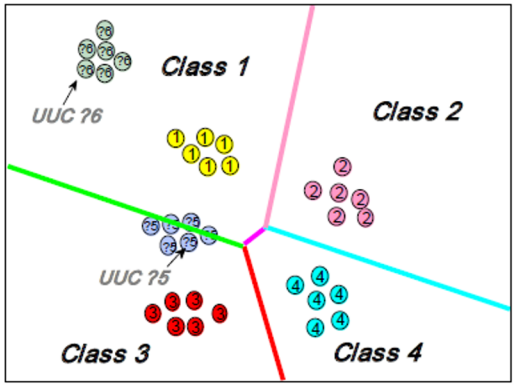
Open Set Learning Overview

- In many tasks, train and test data drawn from same sample/feature space
- More realistic to have unknown or unseen cases
  - Driverless cars, medical diagnoses
  - Logo detection
- Groups of classes
  - Known knowns (positive classes)
  - Known unknowns (negative classes)
  - Unknown knowns (no samples in training but have relevant semantic info)
  - Unknown unknowns (zero information during training)

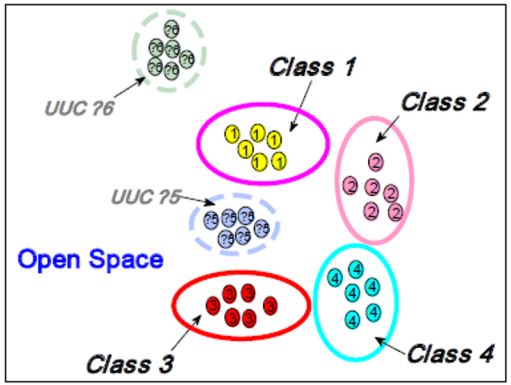
Traditional vs Open Set Classification



(a) Distribution of the original data set.



(b) Traditional recognition/classification problem.



(c) Open set recognition/classification problem.

Duke

Terminology

- Openness

$$O^* = 1 - \sqrt{\frac{2 \times |C_{\text{TR}}|}{|C_{\text{TR}}| + |C_{\text{TE}}|}}.$$

- Openness of 0 indicates closed set
- Focus is on minimizing the “Open Set Risk”

$$\arg\min_{f \in \mathcal{H}} \{R_{\mathcal{O}}(f) + \lambda_r R_{\varepsilon}(f(V))\}$$

Duke

Techniques - Discriminative Models

- Traditional ML
  - SVM → modifying hyperplanes
  - Sparse representation → model tails of matched and sum non-matched error distributions
  - Distance → find ratios of distance between sample and most similar classes
  - Margin Distribution (EVM) → uses marginal distributions to predict probability of sample associated with particular class
- Deep Neural Network
    - OpenMax → represent each class with mean activation vector (MAV), calculate sample distance from MAV and create new distribution
      - New distribution creates pseudo-activation for UUC
      - Probabilities of KKC and UUC recalculated using softmax
      - Works well against "fooling" images but struggles against adversarial ones

Duke

Techniques - Generative Models

- Instance-based
  - Generative OpenMax → creates probability estimation over generated UUCs
  - Counterfactual Image Generation → generate samples close to KKC but are not
    - Generate fake UUC data
- Non instance-based
  - CD-OSR
    - Uses hierarchical Dirichlet process (HDP) framework
    - Continuous co-clustering to create subclasses of KKC and UUC

Duke

Results

TABLE 4
Comparison Among The Representative OSR Methods Using Non-depth Features

Dataset / Method		1-vs-Set	W-OSVM	W-SVM	P_I-SVM	SROSR	OSNN	EVM	CD-OSR
LETTER	O^*=0%	81.51±3.94	95.64±0.37	95.64±0.25	96.92±0.36	84.21±2.49	83.12±17.41	96.59±0.50	**96.94±1.36**
	O^*=15.48%	55.43±3.18	83.83±2.85	91.24±1.48	90.89±1.80	74.36±5.10	73.20±15.21	89.81±0.40	**91.51±1.58**
	O^*=25.46%	42.08±2.63	73.37±1.67	85.72±0.85	84.16±1.01	66.50±8.22	64.97±13.75	82.81±2.42	**86.21±1.46**
PENDIGITS	O^*=0%	97.17±0.58	94.84±1.46	98.82±0.26	**99.21±0.29**	97.43±0.93	98.55±0.71	98.42±0.73	99.16±0.25
	O^*=8.71%	78.43±1.93	87.22±1.71	93.05±1.85	92.38±2.68	96.33±1.59	95.55±1.30	96.97±1.37	**98.75±0.65**
	O^*=18.35%	61.29±2.52	78.55±4.91	88.39±3.14	87.60±4.78	93.53±3.26	90.11±4.15	92.88±2.79	**98.43±0.73**
COIL20	O^*=0%	89.59±1.81	93.94±1.87	86.83±1.82	89.30±1.45	97.12±0.60	79.61±7.41	97.68±0.88	**97.71±0.94**
	O^*=10.56%	70.21±1.67	90.82±2.31	85.64±2.47	87.68±2.02	96.68±0.32	73.01±6.18	95.69±1.46	**97.32±1.50**
	O^*=18.35%	57.72±1.50	87.97±5.40	84.54±3.79	86.22±3.34	**96.45±0.66**	66.18±4.49	93.62±3.33	95.12±2.14
YALEB	O^*=0%	87.99±2.42	82.60±3.54	86.01±2.42	**93.47±2.74**	88.09±3.41	81.81±8.40	68.94±6.47	89.75±1.15
	O^*=23.30%	49.36±1.96	63.43±5.33	84.56±2.19	**88.96±1.16**	83.99±4.19	72.90±9.41	54.40±5.77	88.00±2.19
	O^*=35.45%	34.37±1.44	55.40±5.26	83.44±2.02	**86.63±0.60**	81.38±5.26	67.24±7.29	46.64±5.40	85.56±1.07

The results report the averaged micro-F-measure (%) over 5 random class partitions. Best and the second best performing methods are highlighted in bold and underline, respectively. O^* calculated from Eq. (3) denotes the openness of the corresponding dataset.

Duke

Results

TABLE 5
Comparison Among The Representative OSR Methods Using Depth Features

Dataset / Method		SoftMax	OpenMax	CROSR	C2AE	G-OpenMax	OSRCI
MNIST	O^*=13.40%	97.8	98.1	99.8	<u>98.9</u>	98.4	98.8
SVHN	O^*=13.40%	88.6	89.4	95.5	<u>92.2</u>	89.6	91.0
CIFAR10	O^*=13.40%	67.7	69.5	—	89.5	67.5	<u>69.9</u>
CIFAR+10	O^*=24.41%	81.6	81.7	—	95.5	82.7	<u>83.8</u>
CIFAR+50	O^*=61.51%	80.5	79.6	—	93.7	81.9	<u>82.7</u>
TinyImageNet	O^*=57.36%	57.7	57.6	<u>67.0</u>	74.8	58.0	58.6

The results report the averaged area under the ROC curve (%) over 5 random class partitions [87]. Best and the second best performing methods are highlighted in bold and underline, respectively. O^* calculated from Eq. (3) denotes the openness of the corresponding dataset. Following [85], we here only copy the AUROC values of these methods as some of the results do not provide standard deviations.

Duke

Review

- Challenges/Limitations
  - Many models use thresholding which still depends on KKCs
    - Very difficult to model UUCs
  - Hybrid discriminative and generative models
  - Most OSR models focus on rejecting unknown samples over classifying
  - Take samples individually as opposed to collectively
- Future Work
  - Open World Recognition
    - Label UUCs and add them to classifier/training
  - Improved generalization
    - Use more semantic information of KKCs in addition to feature-level info

Duke

Towards Open Set Deep Networks - 2015

- Present a modification to SoftMax called OpenMax, designed for open set recognition.
- Modifies the Softmax layer by redistributing the activation vector (final layer of neural network).

Duke

OpenMax Computation

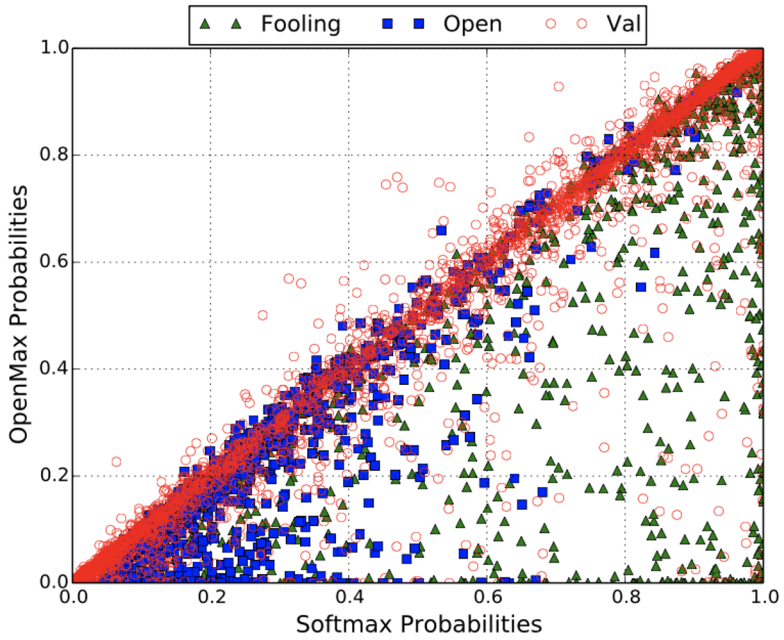
- Network first trained with SoftMax to minimize Cross Entropy Loss.
- Per class mean activation vectors (MAV) for each class calculated using activation vectors from this network.
- Each correctly classified training instance's distance from its MAV calculated (Euclidean-Cosine Distance or other metrics).
- Parameters of separate Weibull distribution are estimated on these distances for each class.
- Choice of Weibull Distribution inspired by Extreme Value Theory.

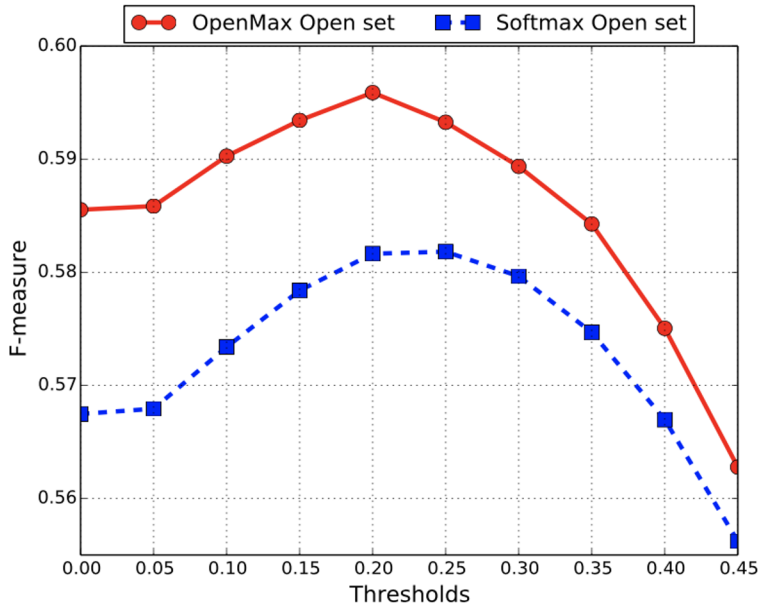
OpenMax Computation - Cont.

- Activation vector values redistributed using Weibull Distribution probabilities.
- Redistributed value summed to represent unknown class activation value.
- Class probabilities (including the unknown class) calculated using softmax on redistributed activation vector.

Duke

Performance Results





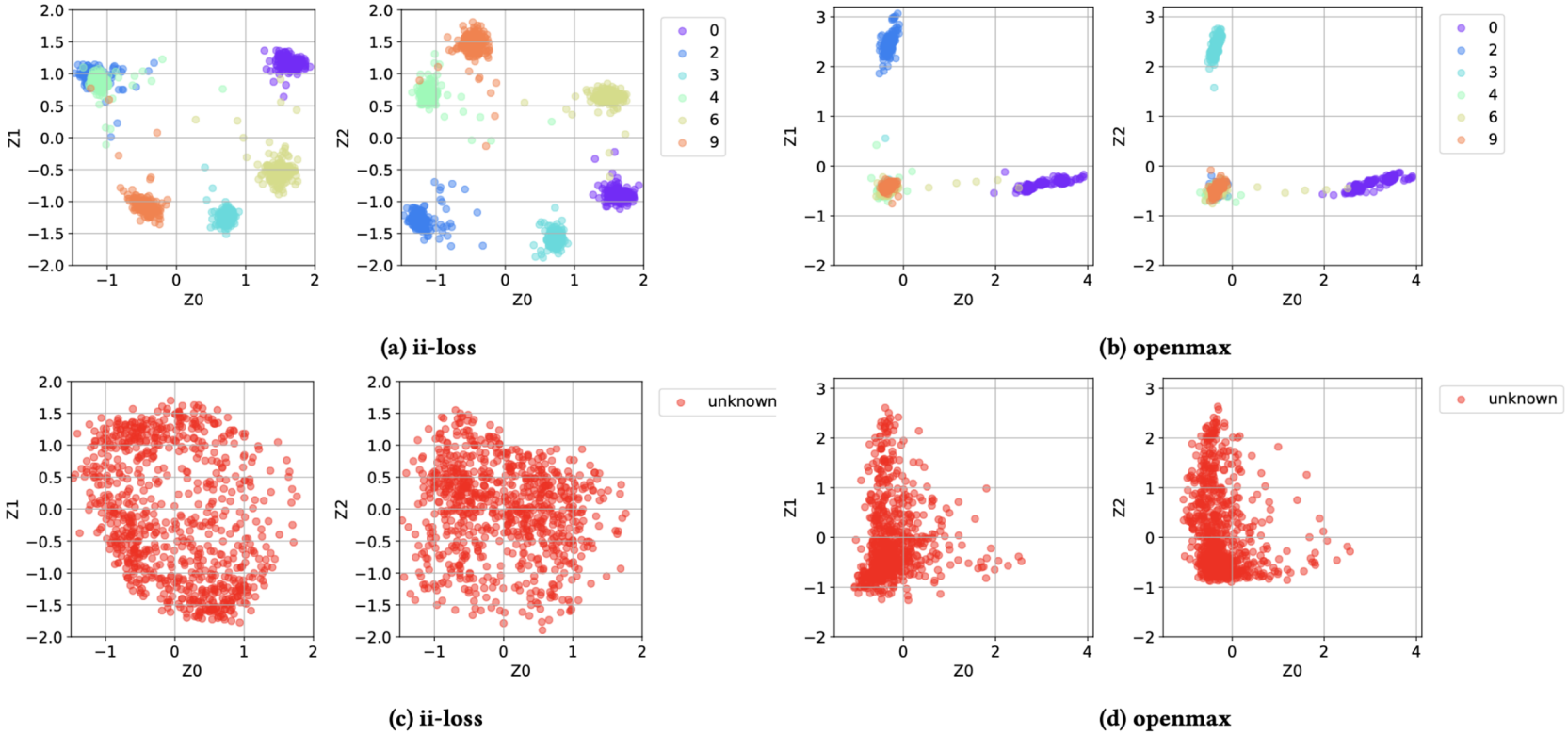
Duke

Learning a Neural-network-based Representation for Open Set Recognition - 2018

- Present a distance based Open Set Recognition approach as an improvement to the OpenMax methodology.
- Drawbacks of OpenMax that this paper attempts to resolve:
  - Does not use a loss function that directly incentivizes projecting class instances around the mean class activation vector.
  - The distance function used by openmax is not necessarily the right distance function for final activation vector space.
- Present a method that increases the occupation of open space between known classes by the unknown class instances.

Duke

Increased Open Space Occupation



Train Model Using Il-Loss

- **Il-Loss = Intra Spread - Inter Separation**
- Intra Spread: Average distance of instances from their class means.
- Inter Separation: distance between the closest two class means among all K known classes.
- Network trained to project members of a class close to the class mean.
- Choose Activation Vector size through parameter tuning.

$$intra\_spread = \frac{1}{N}\sum_{j=1}^{K}\sum_{i=1}^{|C_j|}\left\|\vec{\mu}_j - \vec{z}_i\right\|_2^2$$

$$inter\_sparation = \min_{\substack{1\le m\le K \\ m+1\le n\le K}}\left\|\vec{\mu}_m - \vec{\mu}_n\right\|_2^2$$

$$\vec{\mu}_j = \frac{1}{|C_j|}\sum_{i=1}^{|C_j|}\vec{z}_i$$

Duke

Identify New Classes with Outlier Score

- Outlier Score = distance between activation score of new instance and its closest class mean vector.
- Choose threshold through estimating what percentage of incoming data is outlier.

$$outlier\_score(\vec{x}) = \min_{1 \le j \le K} \left\| \vec{\mu}_j - \vec{z} \right\|_2^2$$

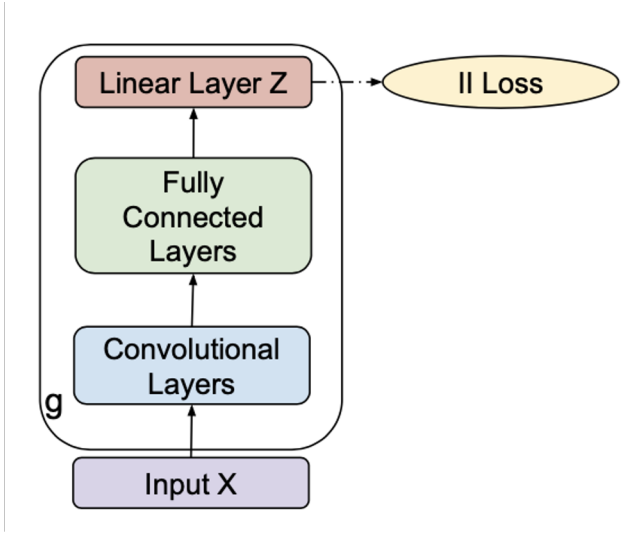
$$P(y = j \mid \vec{x}) = \frac{e^{-\left\| \vec{\mu}_j - \vec{z} \right\|_2^2}}{\sum_{m=1}^{K} e^{-\left\| \vec{\mu}_m - \vec{z} \right\|_2^2}}$$

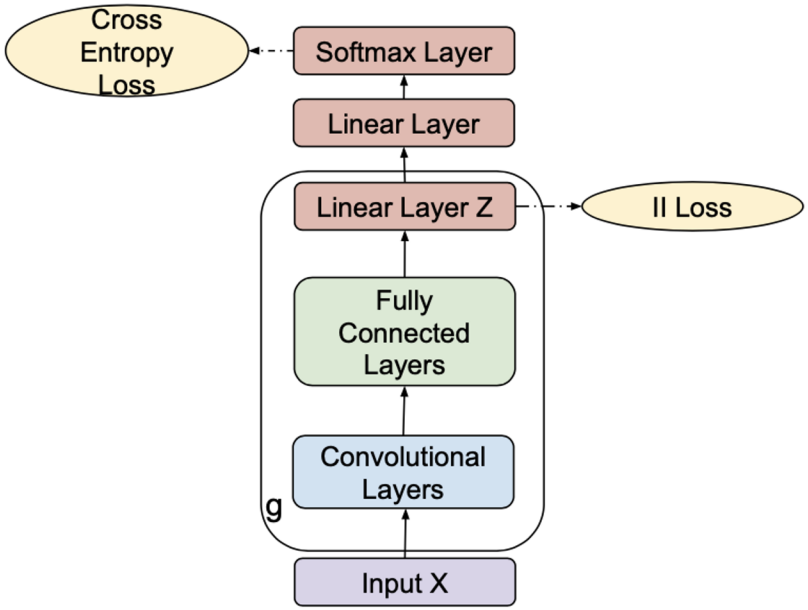
$$y = \begin{cases} K + 1, & \text{if } outlier\_score > threshold \\ \underset{1 \le j \le K}{\operatorname{argmax}} P(y = j \mid \vec{x}), & \text{otherwise} \end{cases}$$

Duke

Model Architecture

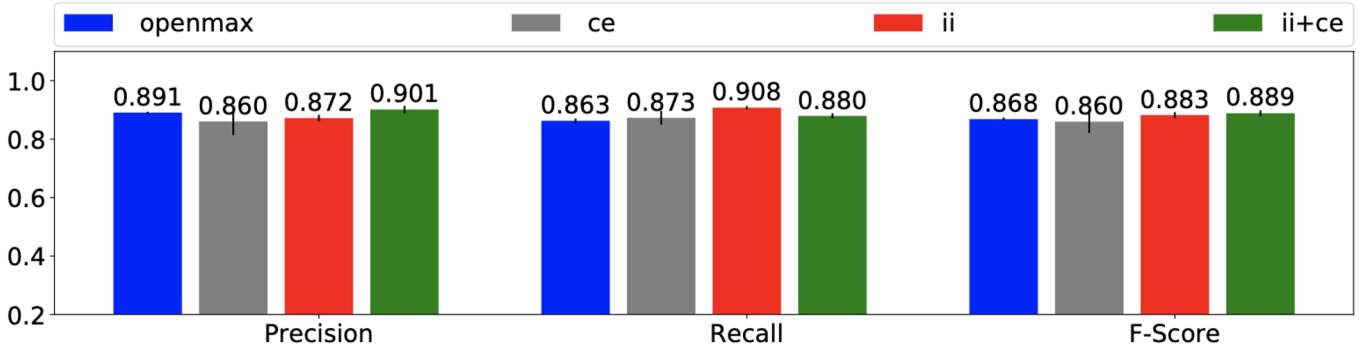
- Can train with just II-Loss or in tandem with Cross Entropy Loss.





Duke

Open Set Learning Improvements



	FPR	ce	ii	ii+ce
MNIST	100%	0.9282 (±0.0179)	**0.9588** (±0.0140)	0.9475 (±0.0151)
	10%	0.0775 (±0.0044)	**0.0830** (±0.0045)	0.0801 (±0.0044)
MS Challenge	100%	0.9143 (±0.0433)	0.9387 (±0.0083)	**0.9407** (±0.0135)
	10%	0.0526 (±0.0091)	**0.0623** (±0.0030)	0.0596 (±0.0035)
Android Genom	100%	0.7755 (±0.1114)	0.8563 (±0.0941)	**0.9007** (±0.0426)
	10%	0.0066 (±0.0052)	0.0300 (±0.0193)	**0.0326** (±0.0182)

Duke

Large-Scale Long-Tailed Recognition in an Open World

Large-Scale Long-Tailed Recognition in an Open World

Ziwei Liu[1,2*] Zhongqi Miao[2*] Xiaohang Zhan[1] Jiayun Wang[2] Boqing Gong[2†] Stella X. Yu[2]
[1] The Chinese University of Hong Kong [2] UC Berkeley / ICSI
{zwliu,zx017}@ie.cuhk.edu.hk, {zhongqi.miao,peterwg,stellayu}@berkeley.edu, bgong@outlook.com

Abstract

Real world data often have a long-tailed and open-ended distribution. A practical recognition system must classify among majority and minority classes, generalize from a few known instances, and acknowledge novelty upon a never seen instance. We define Open Long-Tailed Recognition (OLTR) as learning from such naturally distributed data and optimizing the classification accuracy over a balanced test set which include head, tail, and open classes.

OLTR must handle imbalanced classification, few-shot learning, and open-set recognition in one integrated algorithm, whereas existing classification approaches focus only on one aspect and deliver poorly over the entire class spectrum. The key challenges are how to share visual knowledge between head and tail classes and how to reduce confusion between tail and open classes.

We develop an integrated OLTR algorithm that maps an image to a feature space such that visual concepts can easily relate to each other based on a learned metric that respects the closed-world classification while acknowledging the novelty of the open world. Our so-called dynamic meta-embedding combines a direct image feature and an associ-

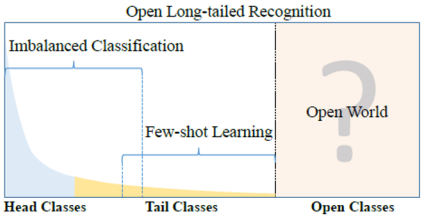
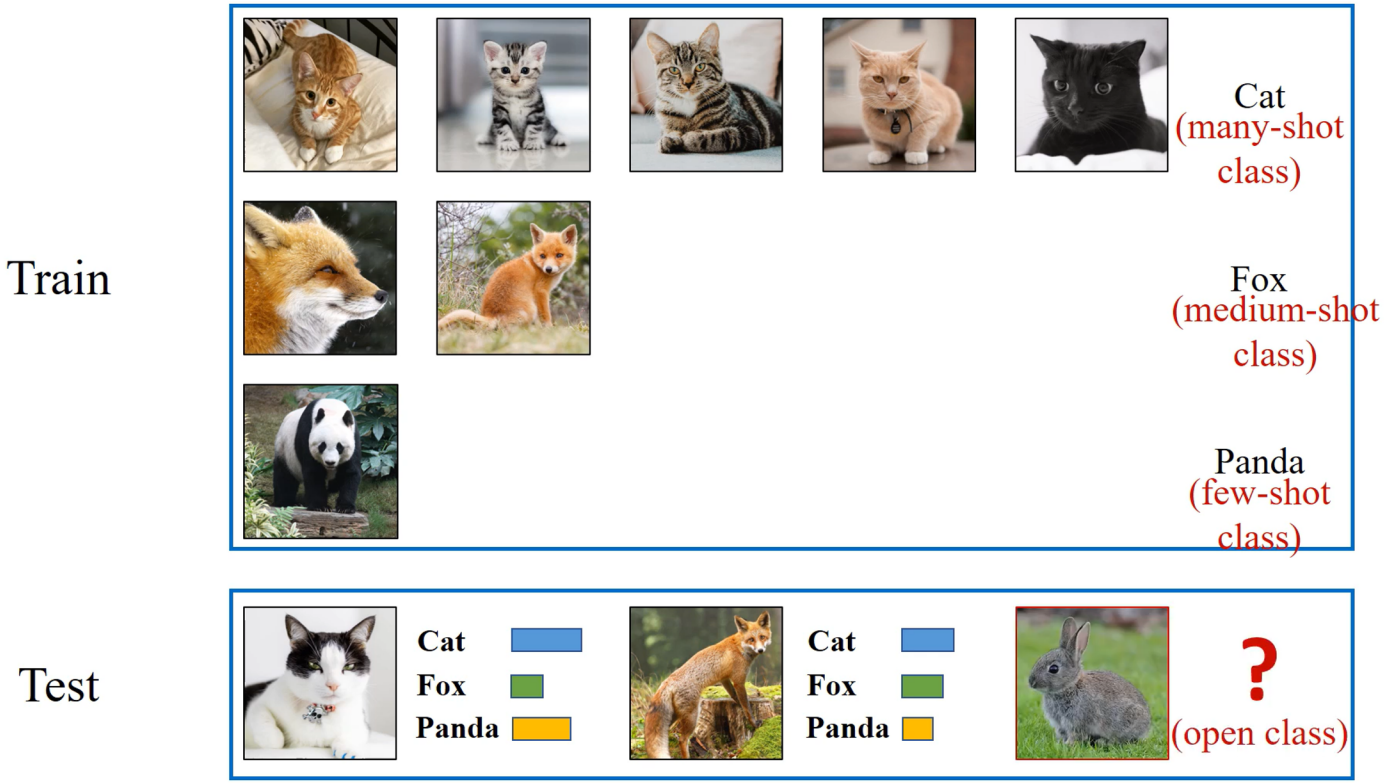


Figure 1: Our task of open long-tailed recognition must learn from long-tail distributed training data in an open world and deal with imbalanced classification, few-shot learning, and open-set recognition over the entire spectrum.

While the natural data distribution contains head, tail, and open classes (Fig. 1), existing classification approaches focus mostly on the head [8, 30], the tail [55, 27], often in a closed setting [59, 34]. Traditional deep learning models are good at capturing the big data of head classes [26, 20]; more recently, few-shot learning methods have been developed for the small data of tail classes. [52, 18].

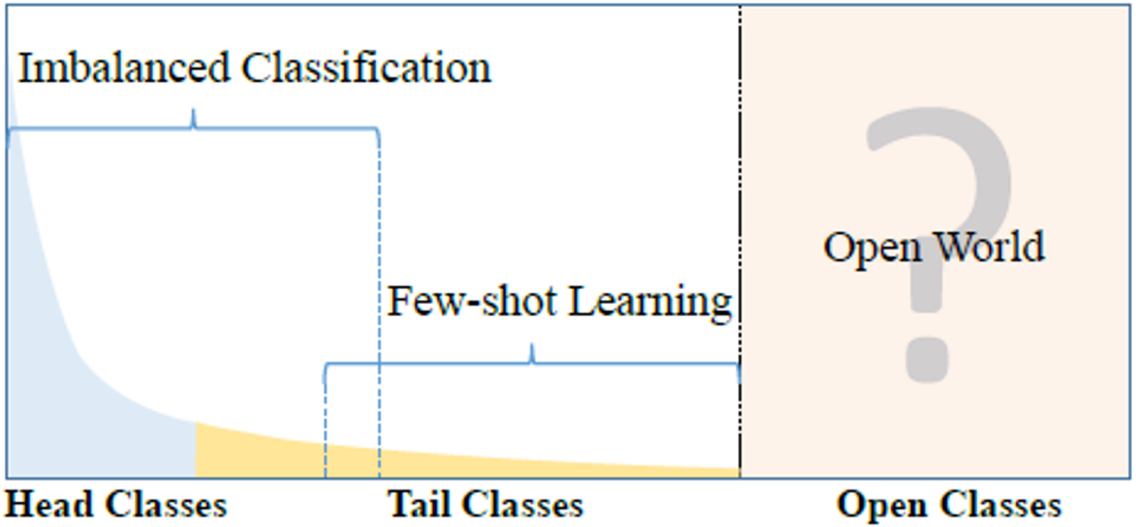
Duke

Introduction



Duke

Two Main Challenges



Open-Set Sensitivity

Tail Recognition Robustness (Remember the training)

Duke

Two-Pronged Solution

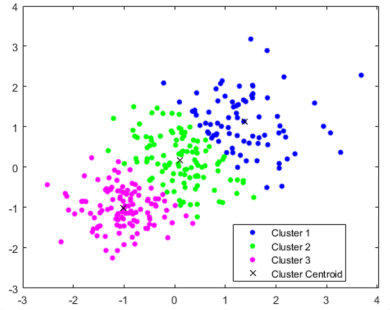
- Dynamic Meta Embedding
  - Direct feature lacks sufficient supervision
  - Direct feature + Induced Feature
  - Visual Memory

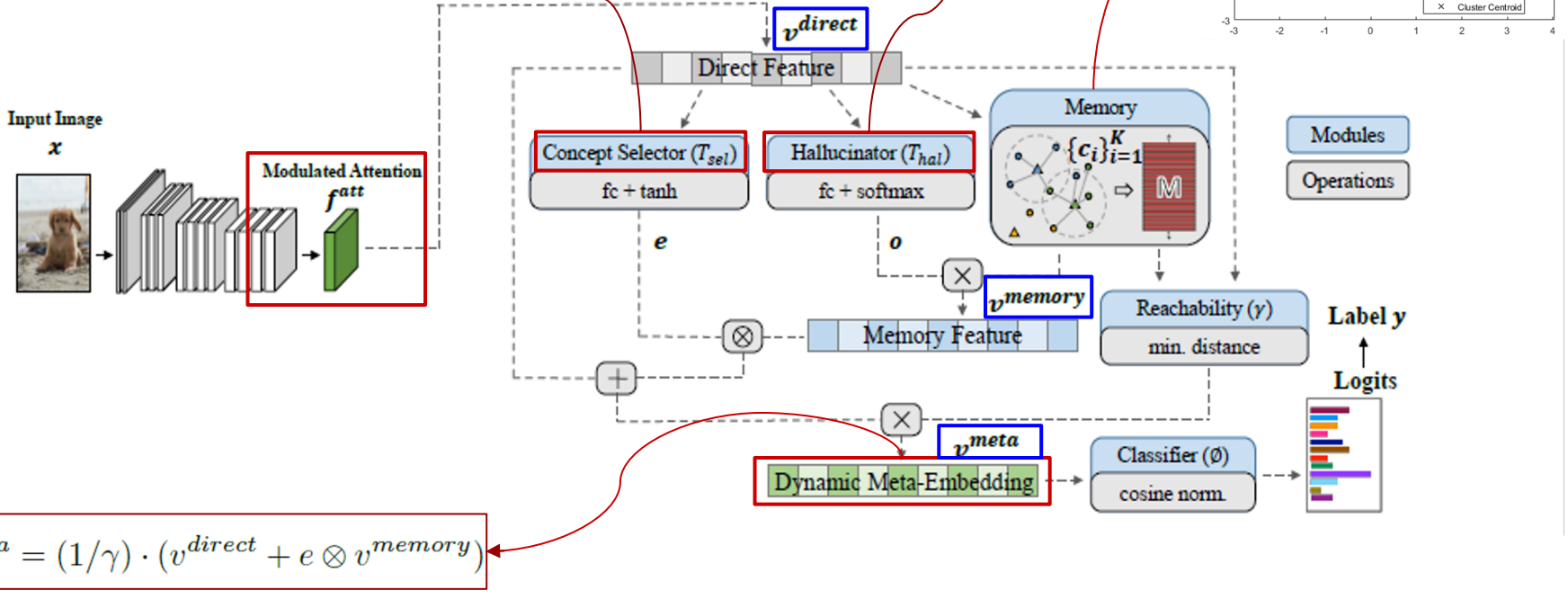
- Modulated Attention
  - Encourages head and tail classes to use different features
  - Enhances discrimination

Duke

How much of direct feature to infuse?

Getting GAN vibes, take a subject and place it in a different surrounding (Takes direct feature and noise as input)

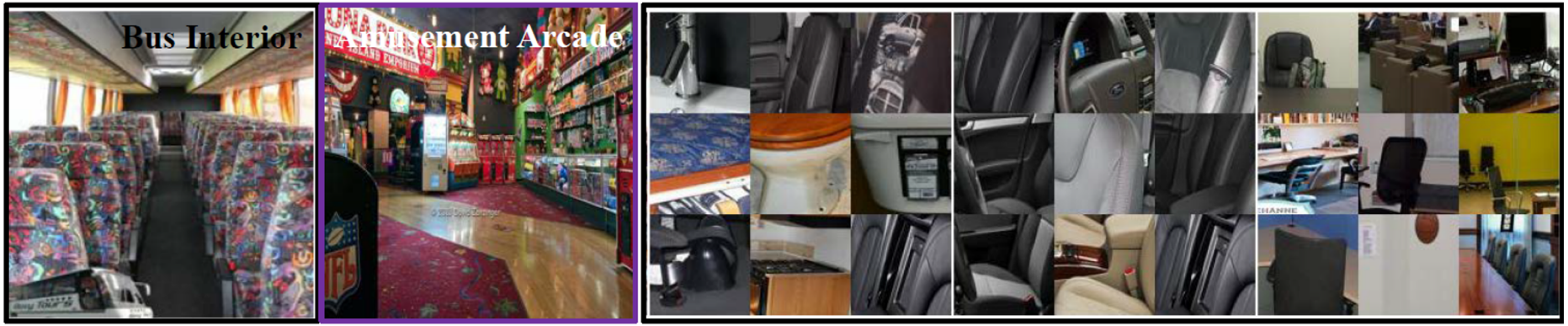


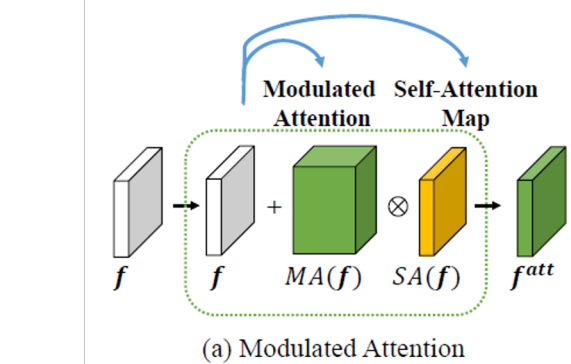


$$v^{meta} = (1/\gamma) \cdot (v^{direct} + e \otimes v^{memory})$$

Duke

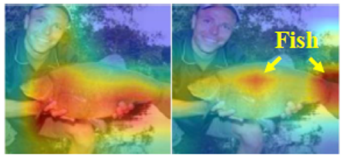
Input Image | Plain Model Prediction | Top-3 Transferred Neurons from Memory Feature





(a) Modulated Attention

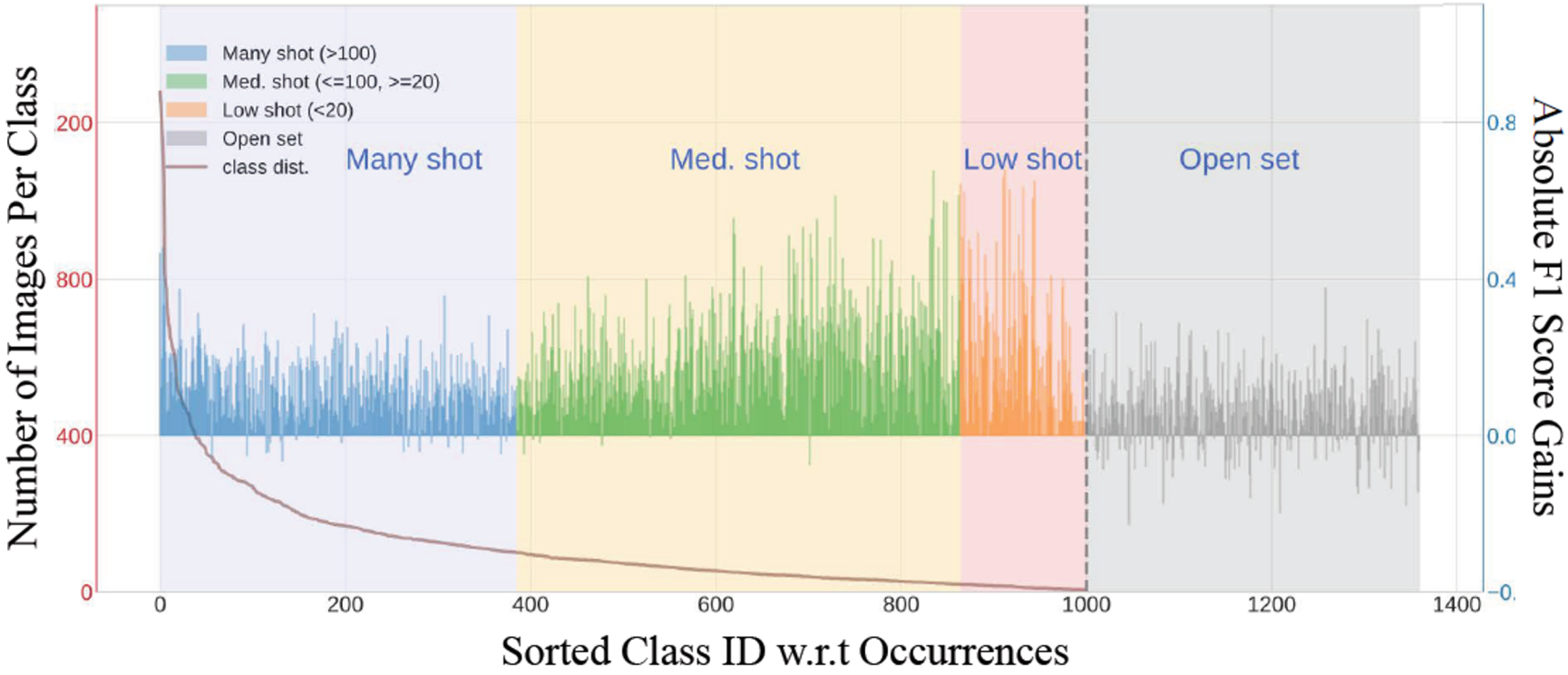
(b.1) Input Image (b.2) Feature Map of Plain ResNet Model



(b.3) Feature Map of Our Model (b.4) Modulated Attention

Duke

Result


Many shot (>100)
Med. shot (<=100, >=20)
Low shot (<20)
Open set
class dist.
Many shot
Med. shot
Low shot
Open set
Number of Images Per Class
Absolute F1 Score Gains
Sorted Class ID w.r.t Occurrences
0
200
400
800
0
200
400
600
800
1000
1200
1400
0.8
0.4
0.0

Model Building Updates

- Have around 2700 annotated screenshots (1200 organised and cleaned)
  - Have a script ready to organise and clean them
  - Typos in labels will arise
- We have 3.3 annotations per image (based on the cleaned ones)
- Need to convert Pascal VOC to COCO format for Detectron2

Duke

End-of-Semester Presentation

- Type of audience?

- Duration?

- Potential topics
- Set up problem- user story - who is using it, their experience, why do we care
  - Talk about how we collected data (and the issues we encountered)
  - Literature review (object detection, small-object detection open-set)
  - Model results

Duke

Next Steps

- Convert from VOC to COCO & YOLO
- Develop initial models and obtain results by middle of next week.
- Utilize Faster-RCNN implementation by Detectron2 and YOLO implementation by Ultralytics.

Duke

References

1. Geng, Chuanxing, Sheng-jun Huang, and Songcan Chen. "Recent advances in open set recognition: A survey." *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).
2. Bendale, A., & Boult, T. E. (2016). Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1563-1572).
3. Hassen, M., & Chan, P. K. (2020). Learning a neural-network-based representation for open set recognition. In *Proceedings of the 2020 SIAM International Conference on Data Mining* (pp. 154-162). Society for Industrial and Applied Mathematics.

Duke

Appendix

Duke

Application to Replace Permanent Resident Card

Department of Homeland Security
U.S. Citizenship and Immigration Services

USCIS
Form I-90
OMB No. 1615-0082
Expires 07/31/2019

For USCIS Use Only	☐ Applicant Interviewed Date: _____	Receipt	Action Block
	Class of Admission		
	Remarks		

► **START HERE - Type or print in black ink.**

Part 1. Information About You

1. Alien Registration Number (A-Number)
   A-
2. USCIS Online Account Number (if any)
   ►

Your Full Name

NOTE: Your card will be issued in this name.

3.a. Family Name (Last Name)
3.b. Given Name (First Name)
3.c. Middle Name
4. Has your name legally changed since the issuance of your Permanent Resident Card?

Mailing Address

(USPS ZIP Code Lookup)

6.a. In Care Of Name
6.b. Street Number and Name
6.c. ☐ Apt. ☐ Ste. ☐ Flr.
6.d. City or Town
6.e. State
6.f. ZIP Code
6.g. Province
6.h. Postal Code
6.i. Country

CMS
CENTERS FOR MEDICARE & MEDICAID SERVICES

Medicare

Beneficiary Services: 1-800-MEDICARE (1-800-633-4227)
TTY/TDD: 1-877-486-2048

This form is used to advise Medicare of the person or persons you have chosen to have access to your personal health information.

Where to Return Your Completed Authorization Forms:
After you complete and sign the authorization form, return it to the address below:

Medicare CCO, Written Authorization Dept.
PO Box 1270
Lawrence, KS 66044

For New York Medicare Beneficiaries ONLY
The New York State Public Health Law protects information that reasonably could identify someone as having HIV symptoms or infection, and information regarding a person's contacts. Because of New York's laws protecting the privacy of information related to alcohol and drug abuse, mental health treatment, and HIV, there are special instructions for how you, as a New York resident, should complete this form.

- For question 2A, check the box for *Limited Information*, even if you want to authorize Medicare to release any and all of your personal health information.

- **Then proceed to question 2B.** You may also check any of the remaining boxes and include any additional limitations in the space provided. For example, you could write "payment information".

TRAVEL VOUCHER OR SUBVOUCHER
(Continuation Sheet)

PAGE 3 OF 3 PAGES

4. NAME *(Last, First, Middle Initial) (Print or type)*

15. ITINERARY | **3. FOR D.O. USE ONLY**

a. DATE 2012		b. PLACE *(Home, Office, Base, Activity, City and State; City and Country, etc.)*	c. MEANS/ MODE OF TRAVEL	d. REASON FOR STOP	e. LODGING COST	f. POC MILES
	DEP	Insert Address (Duty Station)	PA			
	ARR			MC		
	DEP	CITY, STATE ZIP Code (HOR)				
	ARR					
	DEP					
	ARR					
	DEP					
	ARR					
	DEP					
	ARR					
	DEP					
	ARR					
	DEP					
	ARR					
	DEP					
	ARR					
	DEP					
	ARR					
	DEP					
	ARR					
	DEP					
	ARR					
	DEP					
	ARR					
	DEP					
	ARR					
	DEP					
	ARR					
	DEP					
	ARR					

Fill in this information to identify your case:

Debtor 1 _____
First Name Middle Name Last Name

Debtor 2 (Spouse, if filing) _____
First Name Middle Name Last Name

United States Bankruptcy Court for the: _____ District of _____

Case number (If known) _____

☐ Check if this is an amended filing

Official Form 107

Statement of Financial Affairs for Individuals Filing for Bankruptcy

04/19

Be as complete and accurate as possible. If two married people are filing together, both are equally responsible for supplying correct information. If more space is needed, attach a separate sheet to this form. On the top of any additional pages, write your name and case number (if known). Answer every question.

Part 1: Give Details About Your Marital Status and Where You Lived Before

1. **What is your current marital status?**

   ☐ Married
   ☐ Not married

2. **During the last 3 years, have you lived anywhere other than where you live now?**

   ☐ No
   ☐ Yes. List all of the places you lived in the last 3 years. Do not include where you live now.

Debtor 1:	Dates Debtor 1 lived there	Debtor 2:	Dates Debtor 2 lived there
		☐ Same as Debtor 1	☐ Same as Debtor 1
_____ Number Street	From _____ To	_____ Number Street	From _____ To