OPEN-SET RECOGNITION:

A GOOD CLOSED-SET CLASSIFIER IS ALL YOU NEED

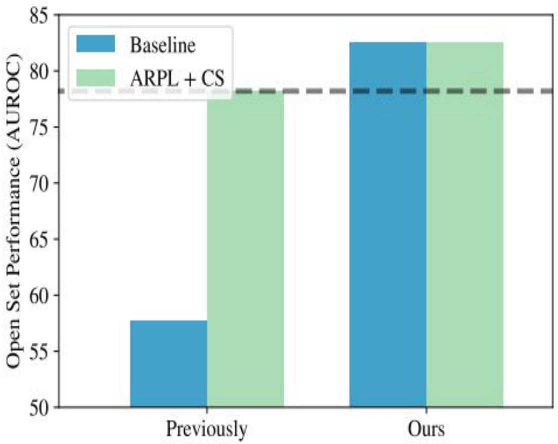
Suryam Arnav Kalra
(19CS30050)

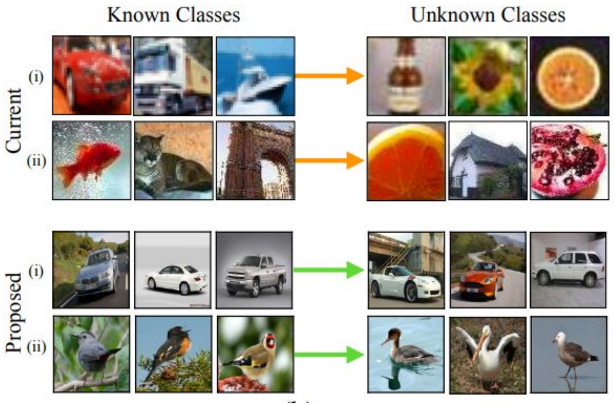
Kunal Singh
(19CS30025)

Overview

- Introduction: What is OSR?
- CORRELATION BETWEEN CLOSED-SET AND OPEN-SET PERFORMANCE.
- A GOOD CLOSED-SET CLASSIFIER IS ALL YOU NEED
- RETHINKING THE OSR BENCHMARKS
- Conclusion

Introduction: What is OSR?



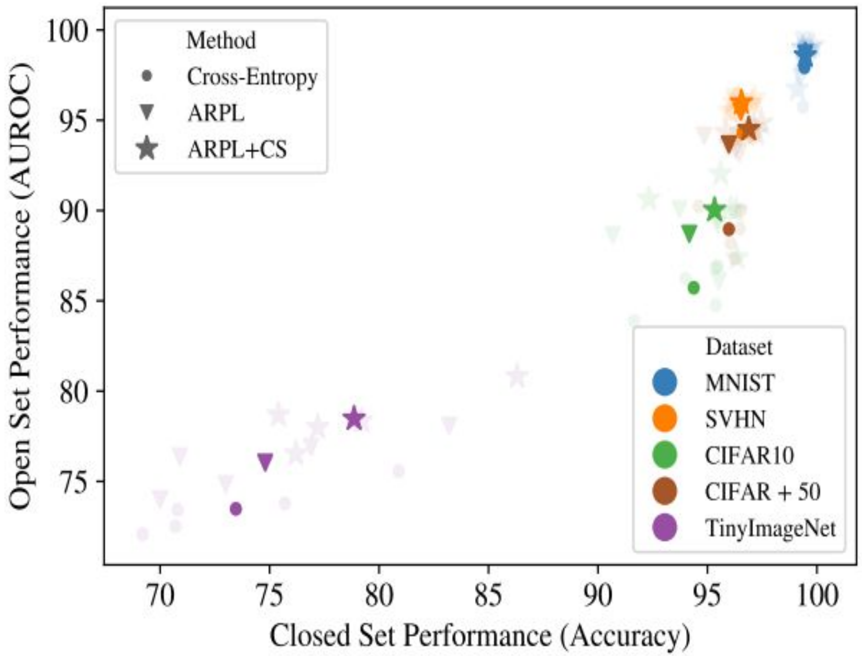


- After the success of modern deep learning systems on closed-set visual recognition tasks, a natural next challenge is open-set recognition (OSR).
- In closed set recognition, a model is tasked with recognizing a set of categories.
- In more realistic setting a model must also indicate if an image comes from a class it has not yet encountered.

CORRELATION BETWEEN CLOSED-SET AND OPEN-SET PERFORMANCE

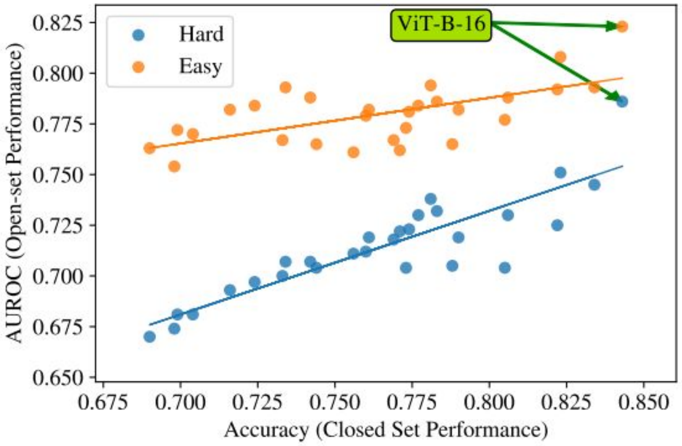
Trained 3 methods on standard openset benchmark datasets

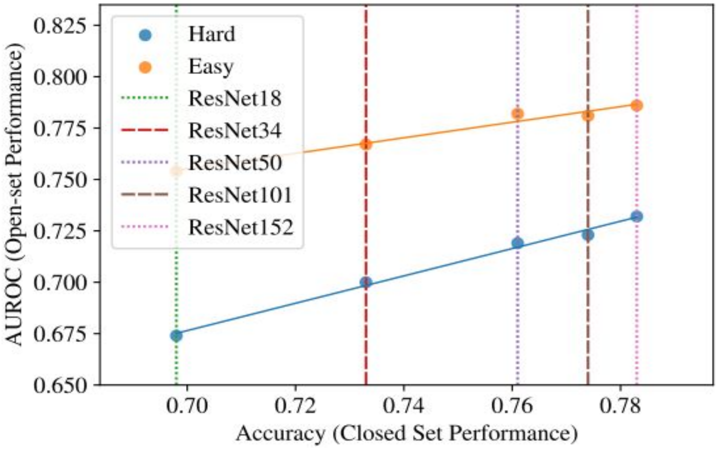
1. cross entropy baseline
2. ARPL
3. ARPL + CS



Pearson Product-Moment correlation $\rho = 0.95$

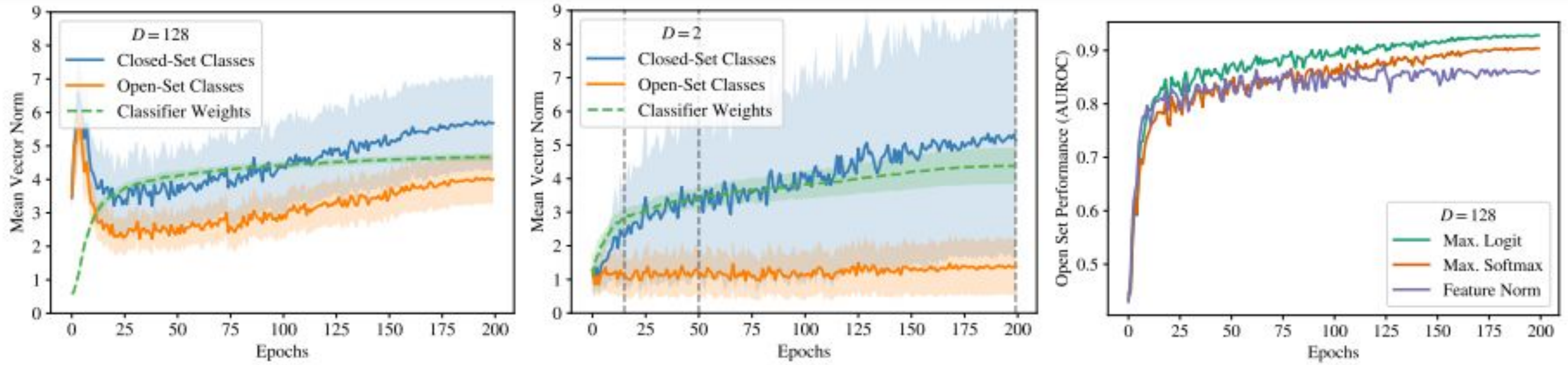
LARGE-SCALE EXPERIMENTS AND ARCHITECTURE ABLATION





- Open-set results on a range of architectures on the ImageNet dataset. ‘Easy’ and ‘Hard’ OSR splits are constructed from the ImageNet-21K-P dataset.
- Here two open-set splits for ImageNet are not randomly sampled, but rather designed to be ‘Easy’ and ‘Hard’ based on the semantic similarity of the open-set categories to the training classes.
- **Results** - A positive correlation between closed and open-set performance can be seen, linear relationship is found to be weaker. In general, no particular model family is found to be remarkably better for OSR than others except the ViT model.

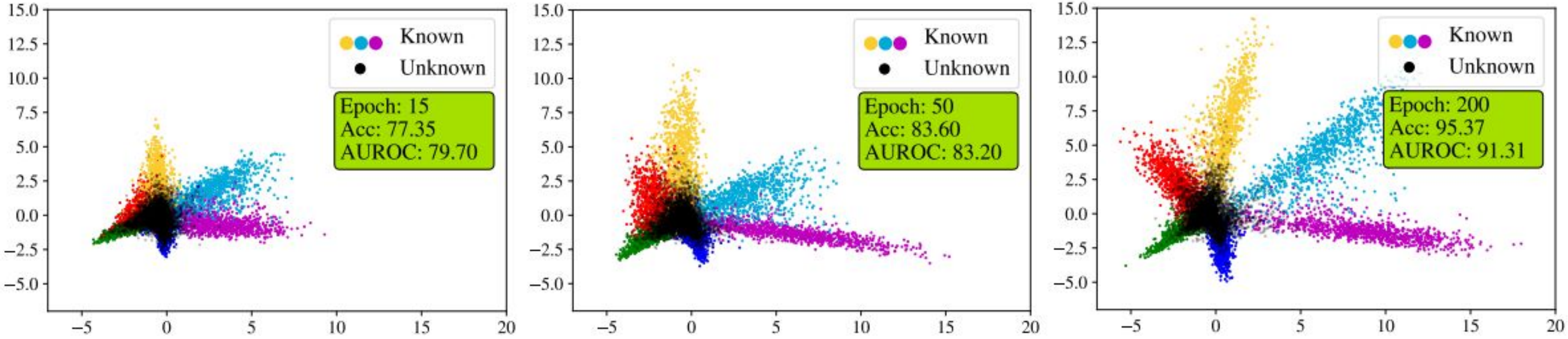
Cross Entropy loss and mean vector plots



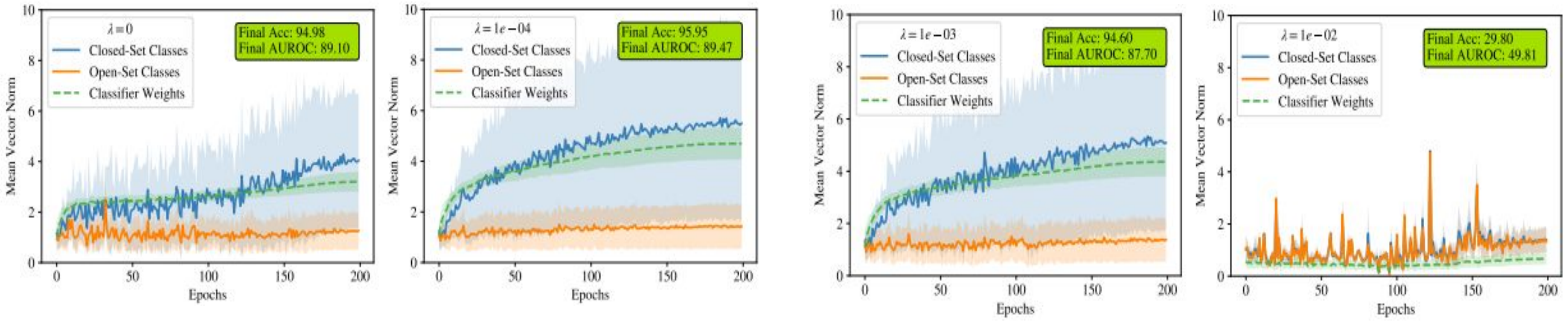
The cross-entropy loss for a single sample in the batch is given by the equation below:

$$\mathcal{L}_i(\theta, \mathbf{W}) = -\hat{y}_{i,c} + \log\left(\sum_{j=1}^{C} \exp(\hat{y}_{i,j})\right) = -\mathbf{w}_c \cdot \Phi_\theta(\mathbf{x}_i) + \log\left(\sum_{j=1}^{C} \exp(\mathbf{w}_j \cdot \Phi_\theta(\mathbf{x}_i))\right)$$

Correlation between open-set and closed-set performance via plots



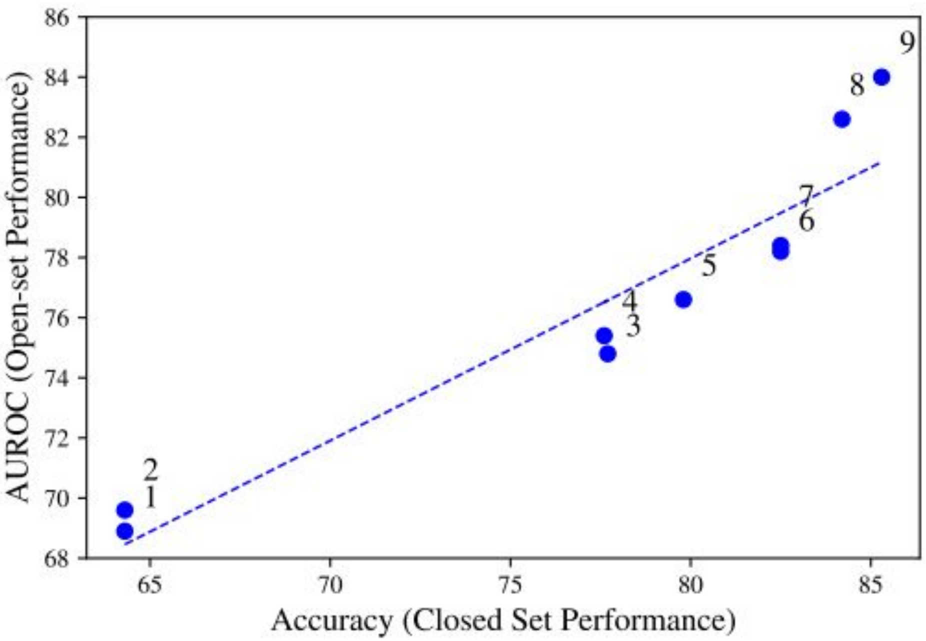
Feature projections for images from seen and unseen classes at different epochs



Ablation for different weight decay values (λ)

A GOOD CLOSED-SET CLASSIFIER IS ALL YOU NEED

- They Leverage the correlation between open and closed-set classifier to improve the performance of the baseline OSR method.
- Specifically improving the closed set accuracy made the model competitive/stronger than state-of-the-art open-set models.



Gains in open-set performance as closed-set performance increases on TinyImageNet

Breakdown of methods used to improve the closed-set classification accuracy of the baseline cross-entropy method.

Epochs	Scheduler	Aug.	Logit Eval	Warmup	Label Smoothing	Ensemble	Closed Set (Accuracy)	Open Set (AUROC)	Combined (OSCR)
100	Step	RandCrop	✗	✗	✗	✗	64.3	68.9	51.4
100	Step	RandCrop	✓	✗	✗	✗	64.3	69.6	50.7
200	Cosine (0)	RandCrop	✓	✗	✗	✗	77.7	74.8	64.3
200	Cosine (0)	CutOut	✓	✗	✗	✗	77.6	75.4	64.7
200	Cosine (0)	RandAug	✓	✗	✗	✗	79.8	76.6	67.3
600	Cosine (2)	RandAug	✓	✗	✗	✗	82.5	78.2	70.3
600	Cosine (2)	RandAug	✓	✓	✗	✗	82.5	78.4	70.3
600	Cosine (2)	RandAug	✓	✓	✓	✗	84.2	82.6	74.3
600	Cosine (2)	RandAug	✓	✓	✓	✓	85.3	84.0	76.1

- Experiments were conducted with a VGG32 backbone over five ‘known/unknown’ splits of the TinyImageNet dataset.
- We can simply see that leveraging standard training strategies for cross-entropy models leads to a significant boost in open-set performance.

Details of Cross Entropy + , OSRCI + , (ARPL + CS) +

Cross Entropy +

- Train the VGG32 model with a batch size of 128 for 600 epochs.
- Initial learning rate of 0.1 for all datasets except TinyImageNet, for which it is 0.01
- RandAugment is used for all experiments.
- Optimal value for label smoothing found to be 0 for all dataset except TinyImageNet, where it is at 0.9.

(ARPL + CS) +

- Same experimental procedure as (ARPL + CS) is used tuning the RandAugment and label smoothing hyperparameters.
- Batch size of 64 and learning rate of 0.001 lead to better performance on TinyImageNet.
- Took significantly longer to train.

OSRCI +

- It involves multiple stages of training.
- First training a GAN to synthesize images similar to the training data, before using generated images as 'open-set' examples for training.
- Reducing all learning rates by a factor 10 compared to Cross-Entropy+ significantly improved performance.

Comparing their improved baseline with other deep learning based OSR methods on the standard benchmark datasets.

Method	Backbone	MNIST	SVHN	CIFAR10	CIFAR + 10	CIFAR + 50	TinyImageNet
Cross-Entropy (Neal et al., 2018)	VGG32	97.8	88.6	67.7	81.6	80.5	57.7
OpenMax (Bendale & Boult, 2016)	VGG32	98.1	89.4	69.5	81.7	79.6	57.6
G-OpenMax (Ge et al., 2017)	VGG32	98.4	89.6	67.5	82.7	81.9	58.0
OSRCI (Neal et al., 2018)	VGG32	98.8	91.0	69.9	83.8	82.7	58.6
CROSR (Yoshihashi et al., 2019)	DHRNet	99.1	89.9	-	-	-	58.9
C2AE (Oza & Patel, 2019)	VGG32	98.9	92.2	89.5	95.5	93.7	74.8
GFROSR (Perera et al., 2020)	VGG32 / WRN-28-10	-	93.5 / 95.5	80.7 / 83.1	92.8 / 91.5	92.6 / 91.3	60.8 / 64.7
CGDL (Sun et al., 2021)	CPGM-AAE	99.5	**96.8**	**95.3**	96.5	96.1	77.0
OpenHybrid (Zhang et al., 2020)	VGG32	99.5	94.7	95.0	96.2	95.5	79.3
RPL (Chen et al., 2020a)	VGG32 / WRN-40-4	99.3 / 99.6	95.1 / **96.8**	86.1 / 90.1	85.6 / 97.6	85.0 / **96.8**	70.2 / 80.9
PROSER (Zhou et al., 2021)	WRN-28-10	-	94.3	89.1	96.0	85.3	69.3
ARPL (Chen et al., 2021)	VGG32	99.6	96.3	90.1	96.5	94.3	76.2
ARPL + CS (Chen et al., 2021)	VGG32	**99.7**	96.7	91.0	97.1	95.1	78.2
Cross-Entropy+	VGG32	98.4 (**+0.6**)	95.9 (**+7.3**)	91.0 (**+23.3**)	95.4 (**+13.8**)	93.9 (**+13.4**)	**82.6** (**+24.9**)
OSRCI+	VGG32	98.5 (-0.3)	89.9 (-1.1)	87.2 (**+17.3**)	91.1 (**+7.3**)	90.3 (**+7.6**)	62.6 (**+4.0**)
(ARPL + CS)+	VGG32	99.0 (-0.7)	**96.7** (+0.0)	93.9 (**+2.9**)	**97.7** (**+0.6**)	**96.0** (**+0.9**)	**82.6** (**+4.4**)

- We can see that their improved baseline significantly outperforms reported baseline figures and largely closes the gap to state-of-the-art OSR models. It also achieves a new state-of-the-art on the most challenging TinyImageNet benchmark.

RETHINKING THE OSR BENCHMARKS

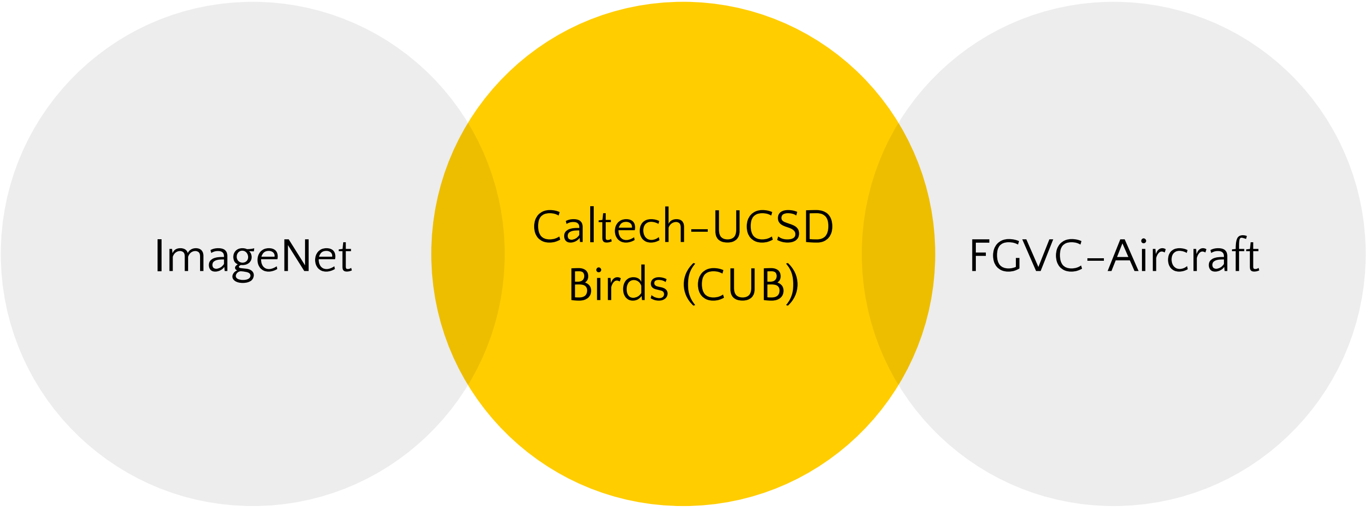
Drawbacks in current OSR Benchmarks

- They all involve small scale datasets
- They lack a clear definition of what constitutes a 'semantic class'.

Addressing the above issues

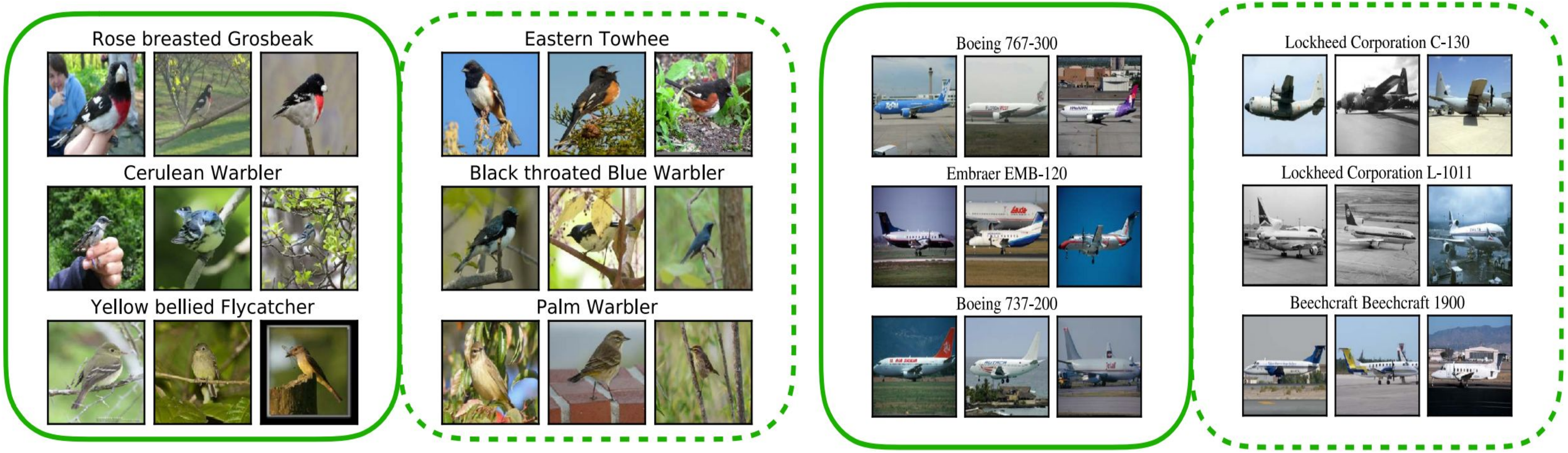
- Use fine-grained datasets which have clear definitions of a semantic class.
- The evaluation settings should aim to explicitly capture the notion of semantic novelty.
- Benchmark Cross-Entropy and ARPL on the new benchmark suite to motivate future research.

PROPOSED BENCHMARK DATASETS



- For ImageNet they introduced a large - scale evaluation benchmark for OSR with open splits based on **semantic distances** to the training splits.
- Designated the ImageNet-1k classes for closed sets.
- Chose open set classes from disjoint sets of ImageNet-21k

Sample classes from closed and open-set splits for the CUB dataset and FGVC-Aircraft dataset



- These are Fine-grained datasets for open-set recognition
- All classes are a variants of a single category and hence define a single **axis of semantic variation**.

Statistics of the proposed OSR benchmarks.

Dataset	Known	Easy	Medium	Hard
CUB	100 (2884)	32 (915)	34 (1004)	34 (991)
FGVC-Aircraft	50 (1668)	20 (667)	17 (565)	13 (433)
ImageNet	1000 (50000)	1000 (50000)	-	1000 (50000)

OSR results on the proposed benchmark datasets.

Method	CUB			FGVC-Aircraft			ImageNet		
	Acc.	AUROC	OSCR	Acc.	AUROC	OSCR	Acc.	AUROC	OSCR
Cross-Entropy+	**86.2**	**88.3 / 82.3 / 76.3**	**79.8 / 75.4 / 70.8**	**91.7**	**90.7 / 86.4 / 77.6**	**86.8 / 83.1 / 75.4**	**78.8**	78.2 / - / 72.6	**66.1 / - / 62.7**
ARPL+	85.9	83.5 / 78.9 / 72.1	76.0 / 72.4 / 66.8	91.5	87.0 / 85.0 / 68.1	83.3 / 81.6 / 66.1	78.1	**79.0 / - / 73.6**	65.9 / - / 62.6

Average Precision (AP) results on the proposed benchmark datasets

	CUB	FGVC-Aircraft	ImageNet
Cross-Entropy+	**67.1 / 58.2 / 47.2**	**69.2** / 58.2 / **39.6**	76.6 / - / 68.6
ARPL+	59.9 / 53.3 / 45.3	66.9 / **58.9** / 34.4	**78.2** / - / **71.2**

- Average precision (AP) for the binary ‘known/unknown’ decision are reported.
- It can be that the ‘Hard’ FGVC open-set splits (with a small number of classes) report substantially poorer AP than AUROC in absolute terms.

Conclusion

- A strong correlation between closed-set and open-set performance of models was observed.
- Closed-set performance was increased using a variety of optimization strategies, model architectures and experimenting on a range of datasets.
- Baseline cross-Entropy OSR model can achieve comparable performance to most state-of-the-art architectures.
- The benchmark datasets used for OSR were reconstructed to differentiate the task of detecting semantic novelty from low level distributional shifts.

Thanks!

Any questions ?

You can find us at

- suryamkalra35@gmail.com
- ksingh19136@gmail.com