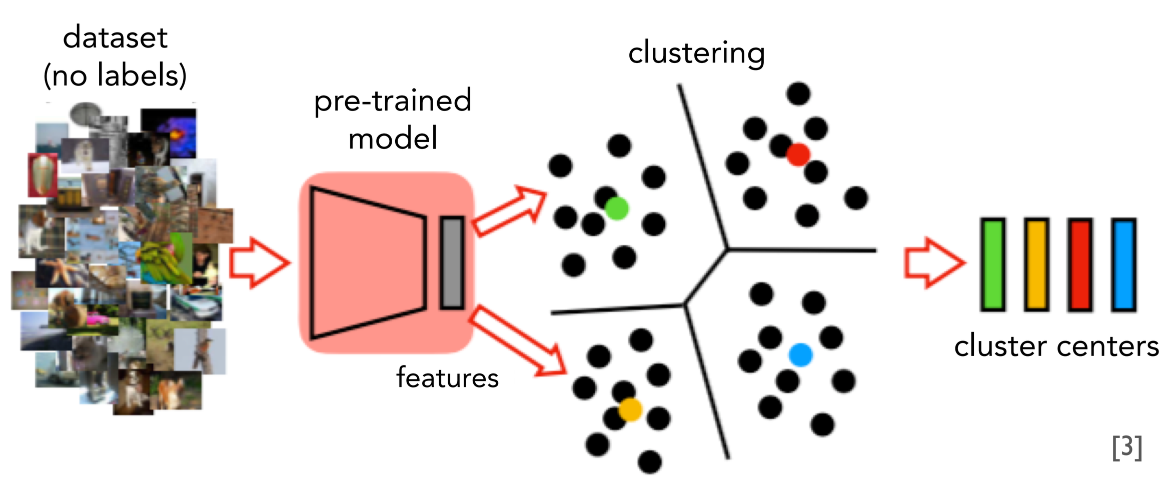
Combining Self-supervised Tasks with Deep Clustering of Visual Features

Cynthia Lai and T. Nathan Mundhenk

ABSTRACT

Iterative clustering of features and using those clusters for classification has proven to be an effective mechanism to improve self-supervised results. Combining this method with a meaningful self-supervised pre-text task may prove to provide fruitful results. With an improved pretraining method incorporating self-supervised tricks, our model has the potential to beat the current state-of-the-art.

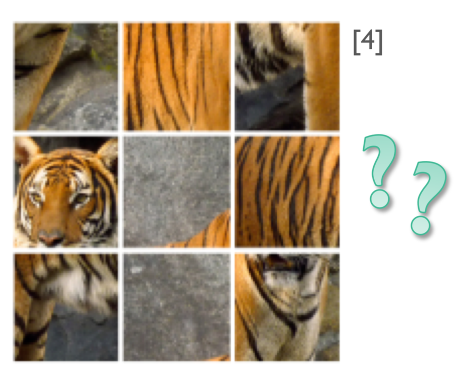


Clustering features outputted from a CNN using k-means

MOTIVATION

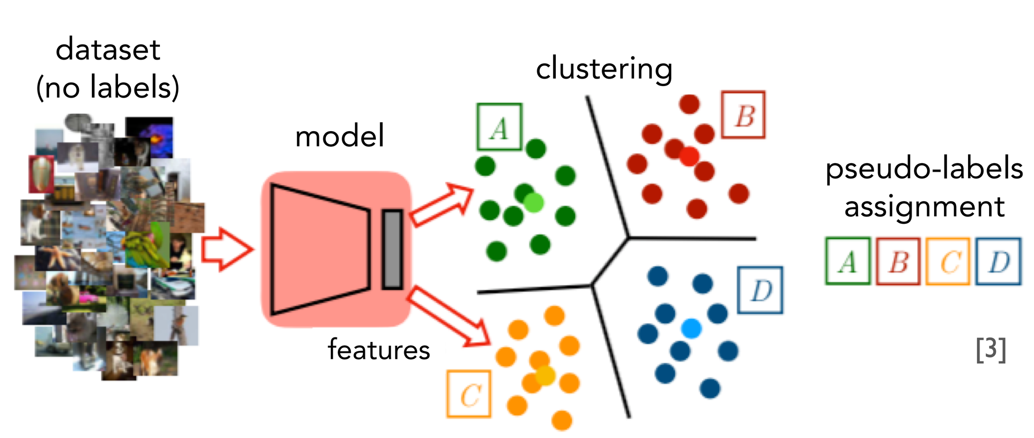
Self-supervision tries to bypass the problem of costly, manual-labeled data. The idea is to take attributes from the data and use those as pseudo-labels.

Ex: Colorization takes the greyscale version of an image and attempts to add color back.



The pre-text task used in this project is the **Jigsaw context problem**. By dividing an image into 9 pieces and shuffling the order, a model has to rely on context between patches to figure out the original orientation.

However, self-supervision has not been able to fully replace supervised learning in terms of performance. **Deep clustering** of features also creates pseudo-labels and can be used in conjunction with the jigsaw puzzle to improve results.



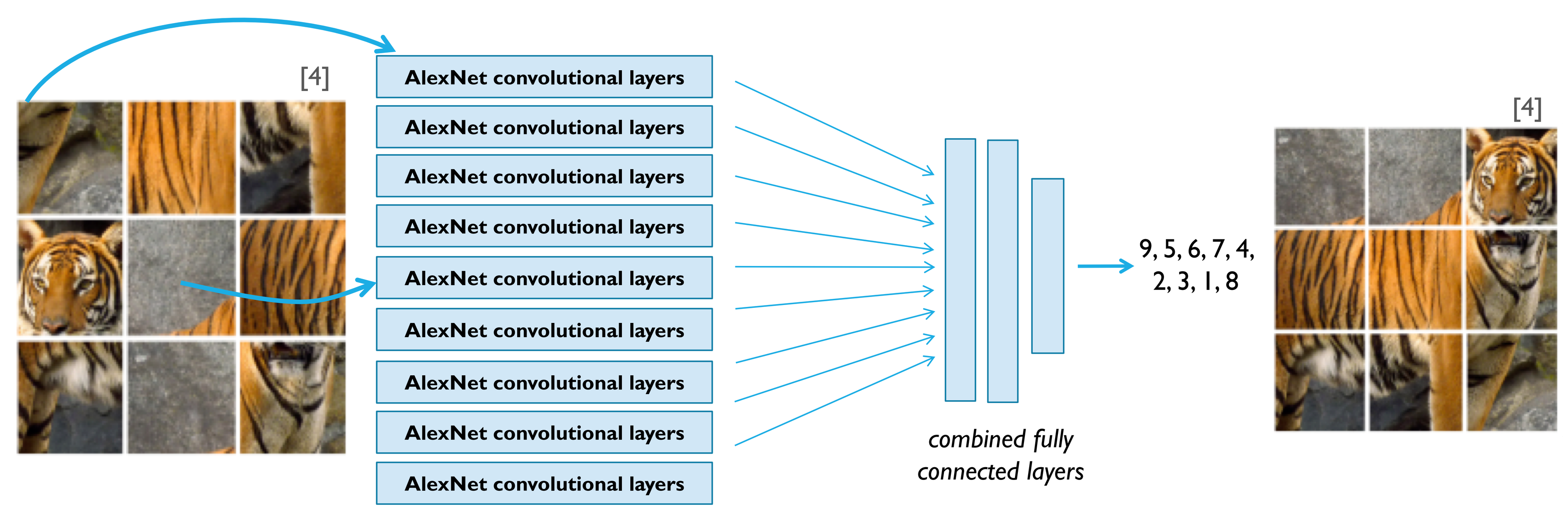
Using cluster centers as pseudo-labels for classification

METHOD

1. Pretraining: Jigsaw puzzle pre-text task

We divide an image into a 3x3 grid and shuffle the pieces.
Each individual patch has its own array of convolutional layers for a 9-Siamese AlexNet.

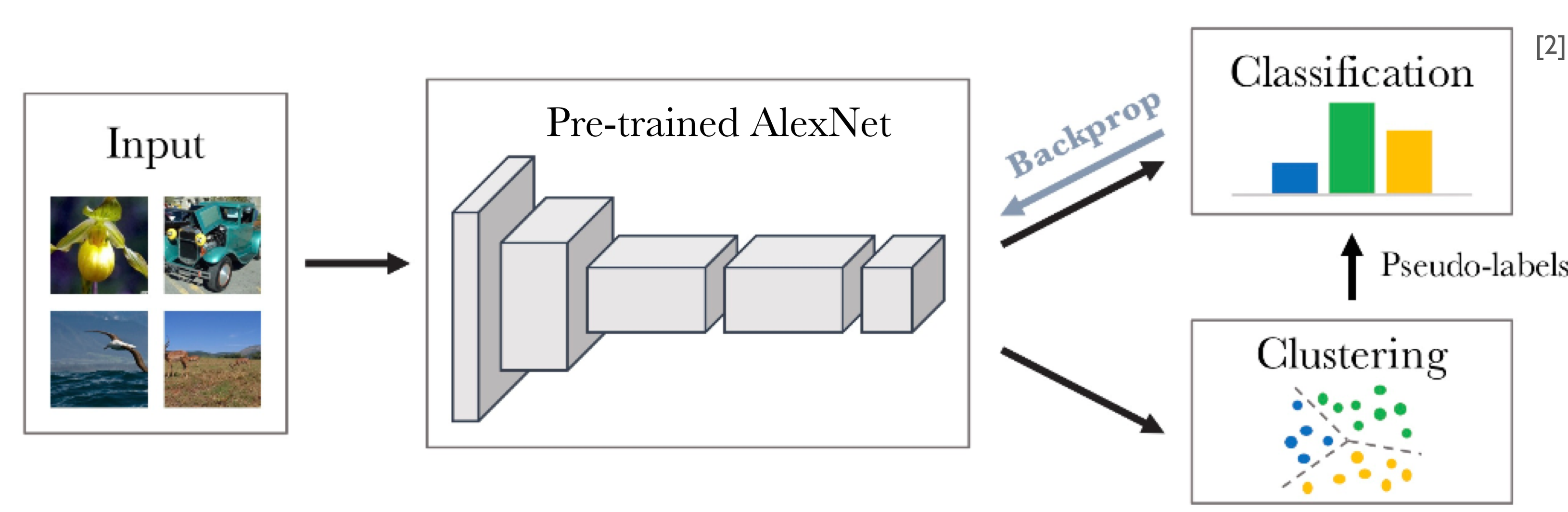
Goal: The network must figure out which permutation the patches are arranged in.



2. Iterative deep clustering and classification

Every epoch, images are passed through the AlexNet convolutional layers and the features are clustered using k-means.

After a round of clustering, cluster labels are used as classification labels.

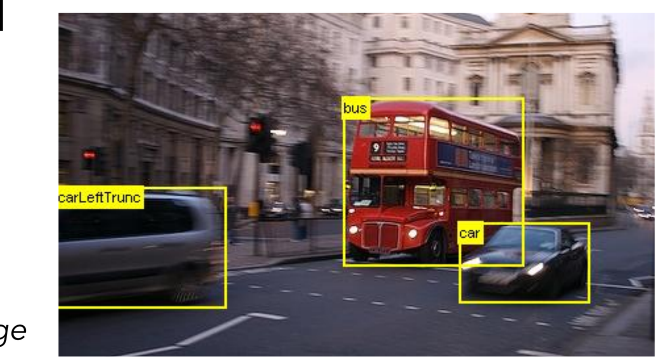


CONCLUSION

These are the current best performances for supervised and unsupervised/self-supervised classification on the PASCAL Visual Object Classes dataset:

	Classification	Detection	Segmentation
ImageNet labels (supervised)	79.9%	56.8%	48.0%
DeepCluster [2] (current best unsupervised)	73.7%	55.4%	45.1%

By incorporating a cleverly constructed self-supervised task as pretraining, *improvements will come!*

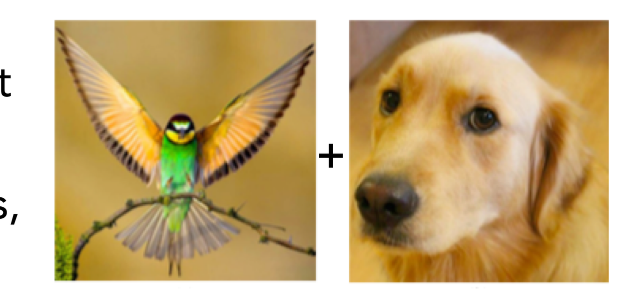


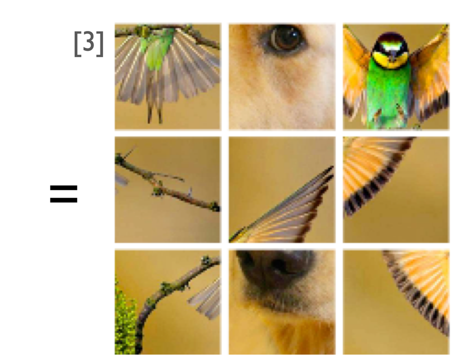
Detection task example to the right, taken from the PASCAL VOC 2007 challenge page

FUTURE WORK

Future possibilities include:

- ✓ **Simultaneously train** the self-supervised pre-text task with the clustering and classification.
- ✓ Try **other clustering algorithms** besides k-means, like expectation maximization.
- ✓ Add in **miscellaneous jigsaw pieces** to make the model figure out which pieces *do not belong* and learn to *organize the remaining pieces.*





REFERENCES

[1] Mundhenk, T. Nathan, Daniel Ho, and Barry Y. Chen. "Improvements to context based self-supervised learning." *arXiv preprint arXiv:1711.06379* (2017).

[2] Caron, Mathilde, et al. "Deep Clustering for Unsupervised Learning of Visual Features." arXiv preprint arXiv:1807.05520 (2018).

[3] Noroozi, Mehdi, et al. "Boosting Self-Supervised Learning via Knowledge Transfer." arXiv preprint arXiv:1805.00385 (2018).

[4] Noroozi, Mehdi, and Paolo Favaro. "Unsupervised learning of visual representations by solving jigsaw puzzles." European Conference on Computer Vision. Springer, Cham, 2016.

LLNL-POST-755944