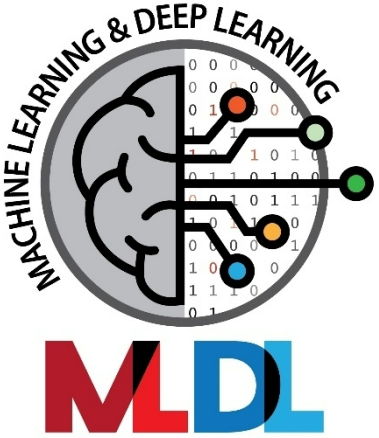
SAND2021-8213C



MLDL

Machine Learning and Deep Learning Conference 2021

Machine Learning with Synthetic Data: What Approaches are Most Effective?

- Jenny Galasso (6364)
- Rob Forrest (8716), Arvind Prasadan (8762)

Synthetic data for ML

- For many domain areas, labeled data is difficult or expensive to obtain and having enough data for Machine Learning isn't feasible.
- In these cases, it may be possible to synthetically generate massive amounts of data for training purposes.
- But how much synthetic data is needed and what approaches are most effective?
- Using the Lynceus Organics AgBio dataset and physics-based synthetically-generated data, this effort hopes to answer some of these questions.

The data

- We have dual view X-ray scans of luggage with eight classes of AgBio items that we are trying to classify (e.g. apple, orange, banana, etc)
  - Ten of these packings have been held out as the test set
- Arvind Prasadan has taken the remaining empty-bag packings along with individually-scanned items and has created synthetic images
  - Items are randomly located within packed bag boundaries
  - 36846 samples created per ‘set’
  - Six of these sets were created (221076 samples)
    - Each set would have the same packing and individual item, but in a different position
  - 80/20 Train/Validation Splits
- Convert 2-channel raw data to a 3-channel png file (3rd channel empty)

Questions/Experiments

- Can we use a single model trained on real data to evaluate various synthetic data techniques?
- Can we add multiple ‘sets’ of synthetic data to improve accuracies?
- Is using semi-supervised learning effective with synthetic data?
- Can we add a small amount of real data to the training set to improve accuracies?

Can we use a model trained on real data to evaluate various synthetic data techniques?

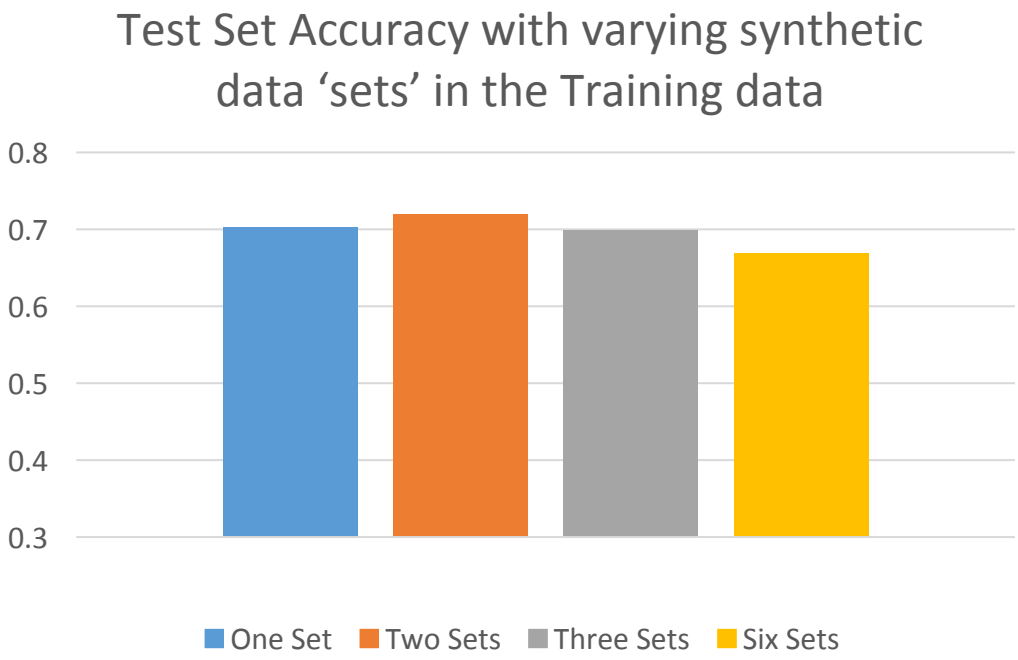
- Yes, we can quickly identify the best synthetic data techniques using a single model trained on real data!
  - The first row was from a single-trained model
  - The remaining rows required a trained model per Synthetic Data Set (i.e. time consuming!)
- There appears to be minimal effect from 'packing contamination'

Accuracies when evaluating several synthetic data sets

	Synth DS 0	Synth DS 1	Synth DS 2	Synth DS 3	Synth DS 4	Synth DS 5
Train on Organics, Test on Synth DS (packing contamination)	0.511	0.504	0.589	0.589	**0.709**	0.706
Train on Synth DS, Test on all Organics (packing contamination)	0.349	0.325	0.511	0.212	**0.623**	0.378
Train on Synth DS, Test on Held-Out Organics Test Set	0.34	0.327	0.502	0.227	**0.631**	0.393

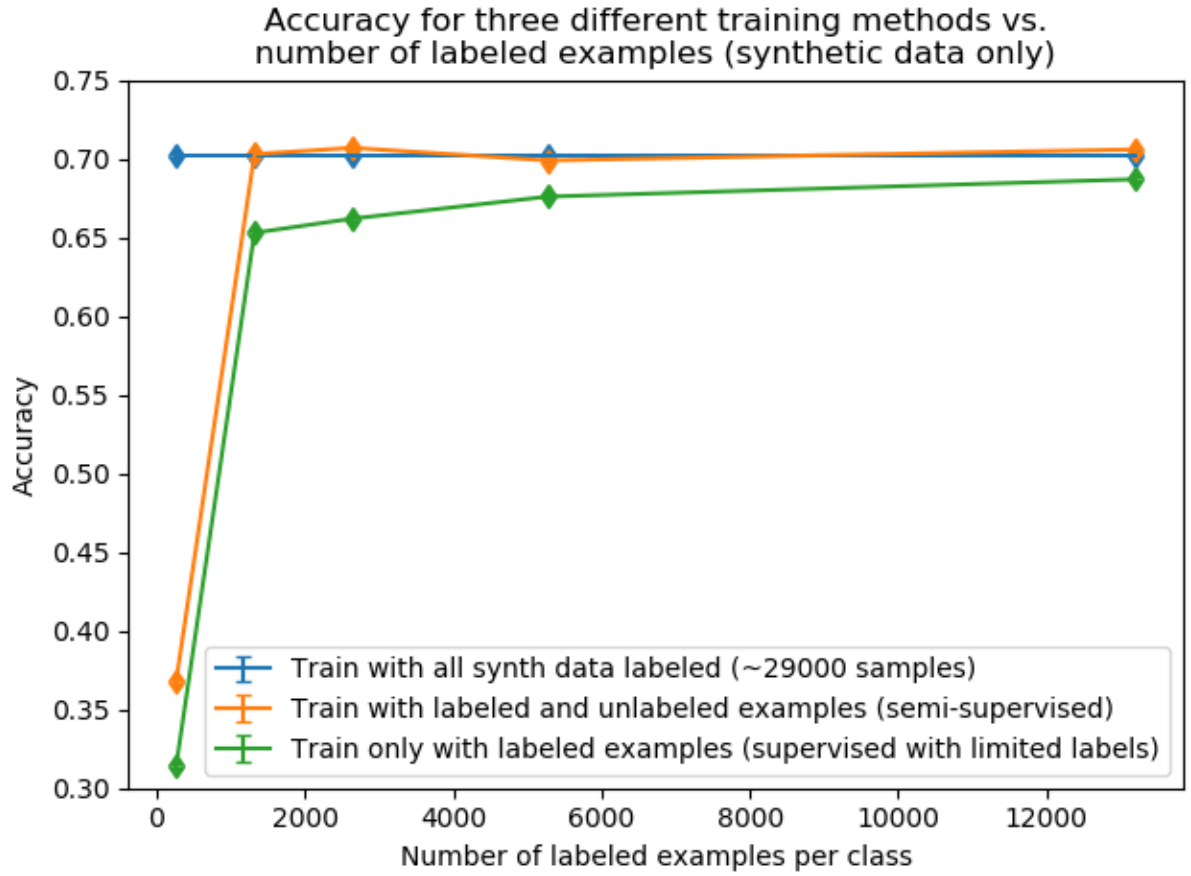
Can we add multiple ‘sets’ of synthetic data to the training set to improve accuracies?

- No, adding multiple sets of synthetic data only makes it take longer to train
  - Each set includes ~36K samples
- Each set repeats the same background packing and varies in only the placement of the item
- We are looking into creating synthetic data with more variation to overcome this



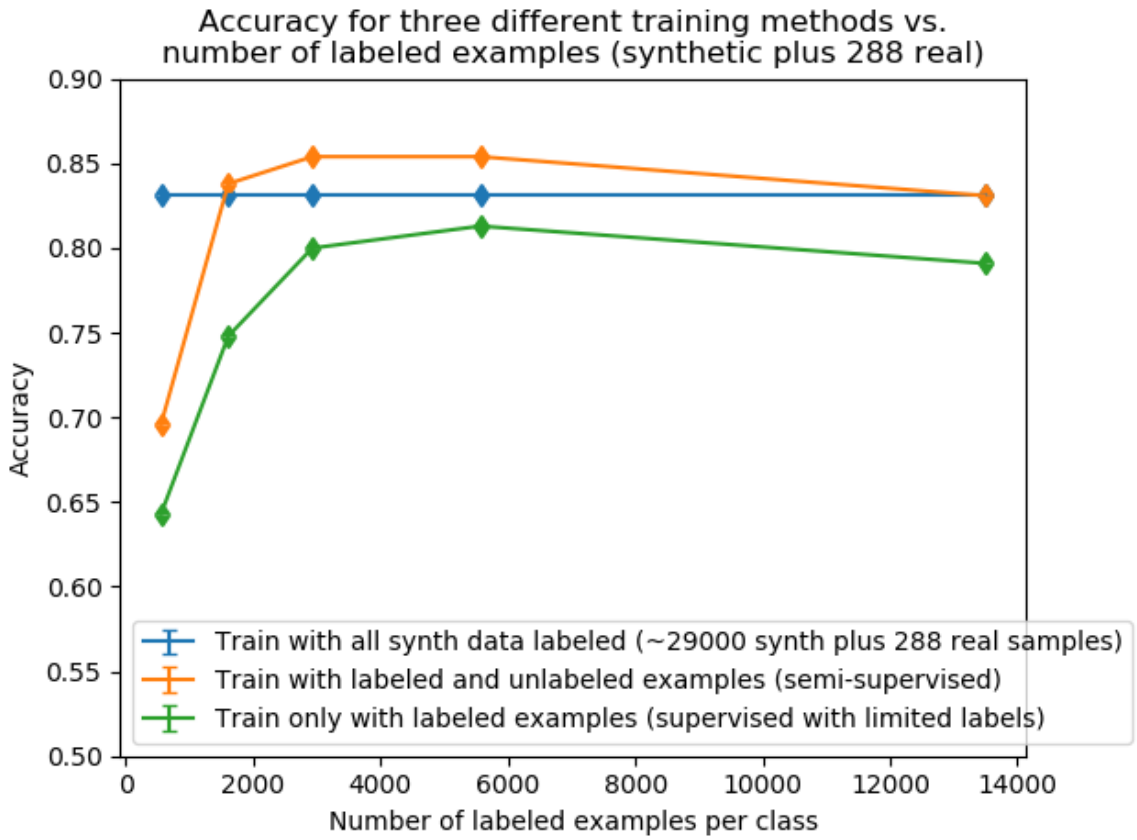
Is using semi-supervised learning effective with synthetic data alone?

- No, there is no improvement over fully-supervised



Can we add a small amount of real data to the training set to improve accuracies?

- Yes, adding only 288 labeled samples, and keeping everything else the same as the previous experiment, dramatically improves accuracies AND makes semi-supervised outperform fully-supervised!
  - Fully-supervised synthetic only: 0.702
  - Fully-supervised synthetic plus 288 real samples: 0.831



Conclusions

- Using a single model trained on real data to evaluate various synthetic data techniques is fast and effective
- More synthetic data isn't better if there isn't a lot of variation in the additional data
- Adding only a small amount of labeled real data samples dramatically improves performance
- Semi-supervised trained models outperform fully-supervised models if a small amount of real data is included in the labeled training set