Data, Assemble: Towards Efficient Medical Image Analysis

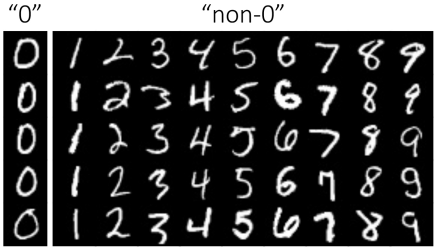
Zongwei Zhou, PhD

Johns Hopkins University

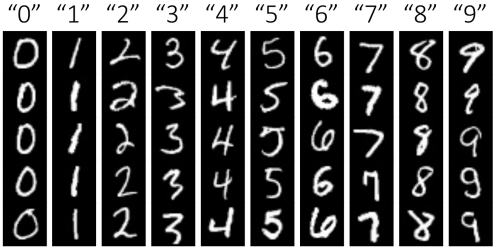
October 1, 2021

Example I

- Task: To classify images of "0"
- MNIST[1]: A dataset that provides images and annotations of "0–9"
- MNIST-*zero*: Derived from MNIST, wherein only the images of "0" are labeled as positives and the remainder are negatives (sufficient for the task)
- The numbers of images are the same in MNIST and MNIST-*zero*
- The only difference is that the makeup of negatives ("1–9") is unknown in MNIST-*zero*, yet it is known in MNIST



MNIST-*zero*

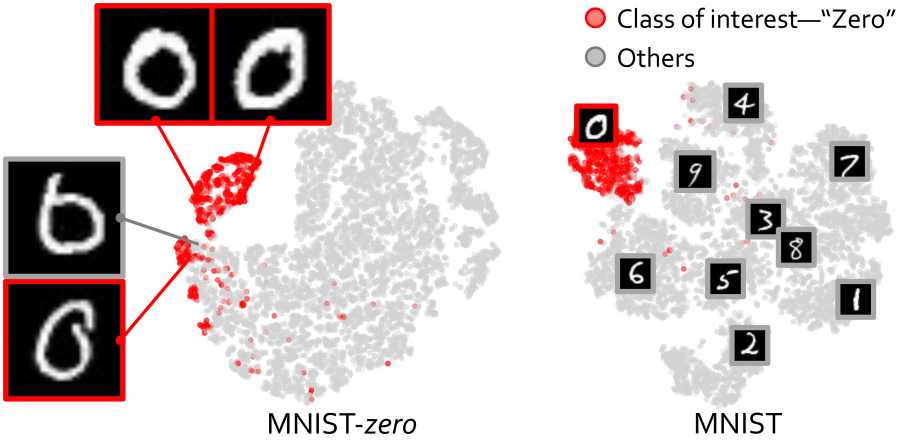


MNIST

[1] Yann LeCun et al. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

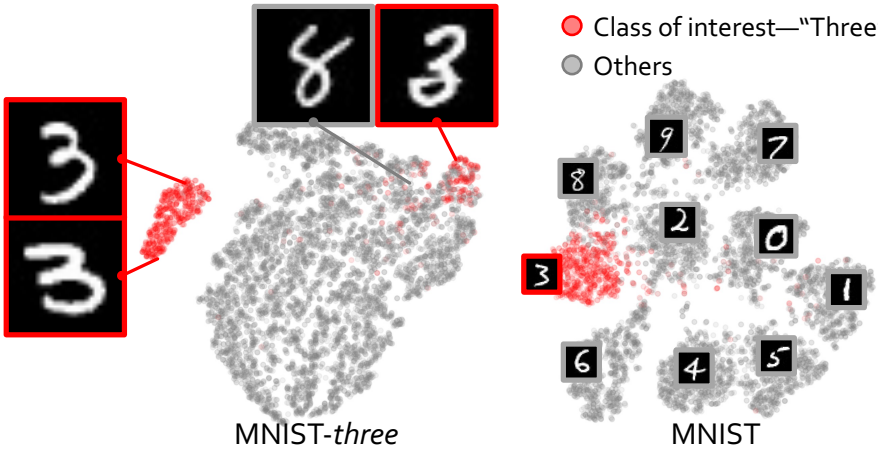
Example I

- MNIST outperforms MNIST-*zero* for classifying images of "0"
- Fine-grained labels (e.g., 1–9) in negative examples ("non-0") positively affect the classification of "0".
- The lack of fine-grained labels causes confusion between zero-like "6" and "0"



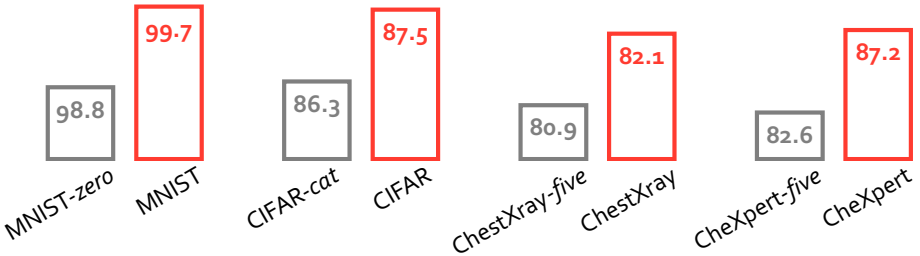
Example I

- Similarly, MNIST outperforms MNIST-*three* for classifying images of "3"
- Fine-grained labels (e.g., 1–2, 4–9) in negative examples ("non-3") positively affect the classification of "3".
- The lack of fine-grained labels causes confusion between three-like "8" and "3"



More Demonstrations for Example I (Classification)

- The numbers of images are the same in
  - CIFAR-*cat* vs. CIFAR
  - ChestXray-*five* vs. ChestXray
  - CheXpert-*five* vs. CheXpert
- The performance was evaluated on the class of interest (e.g., zero, cat, five diseases)
- Conclusion: A dataset that is labeled with many classes can foster more powerful *classification* models than one that is only labeled the class of interest

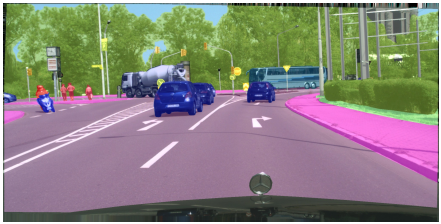


More Demonstrations for Example I (Segmentation)

- Task: To segment "Bus" from images
- Cityscapes[2]: A dataset that provides images and pixel-wise annotations of 19 classes.
- Cityscapes-*five*: Derived from Cityscapes, wherein only five classes (Bus, Road, Sidewalk, Building, and Wall) are labeled (sufficient for the task)
- The numbers of images are the same in Cityscapes and Cityscapes-*five*
- The only difference is that the makeup of background is unknown in Cityscapes-*five*, yet it is known in Cityscapes

	bus	road	sidewalk	building	wall	
terrain	sky	traffic light	traffic sign	car	fence	pole
motorcycle	bicycle	person	rider	vegetation	truck	train

Cityscapes-*five*

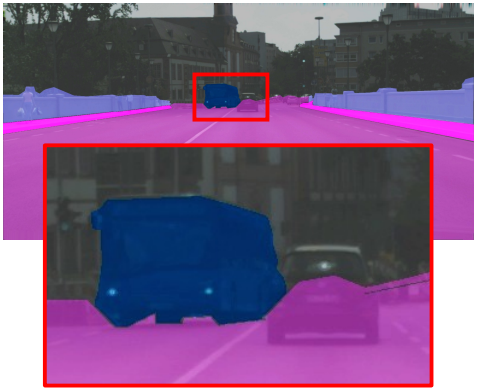


Cityscapes

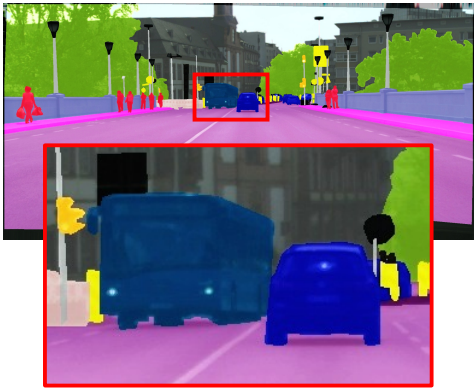
[2] Marius Cordts et al. "The cityscapes dataset for semantic urban scene understanding". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 3213–3223.

More Demonstrations for Example I (Segmentation)

- Cityscapes outperforms Cityscapes-*five* for segmenting bus
- Fine-grained labels in negative examples ("background") positively affect the segmentation of "Bus".
- The lack of fine-grained labels causes confusion between "car" and "bus"
- Conclusion: A dataset that is labeled with many classes can foster more powerful *segmentation* models than one that is only labeled the class of interest



Cityscapes-*five*



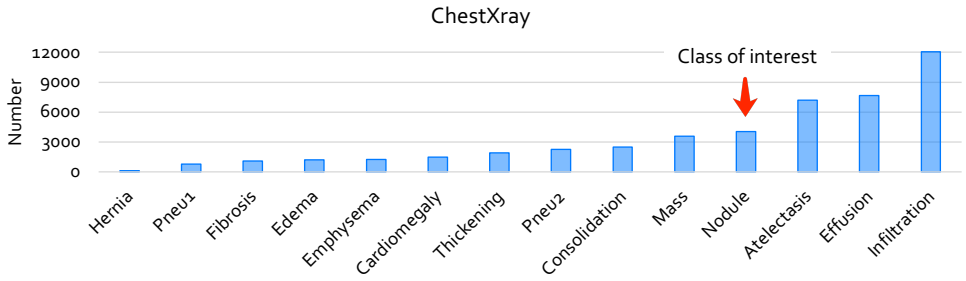
Cityscapes

Thoughts on Example I

- A dataset that is labeled with many classes can foster more powerful models than one that is only labeled the class of interest
  - Classification—true
  - Segmentation—true
  - Detection, localization, other problems
- Learning from classes in "negative examples" can better delimit the decision boundary of the class of interest
- Wanted: A dataset that is labeled with many classes.

Example II

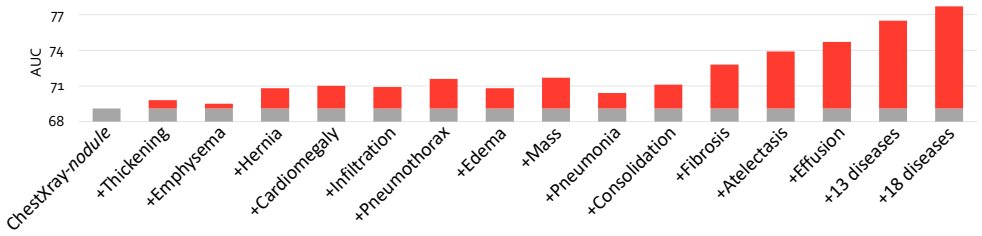
- Task: To classify images of "lung nodule"
- ChestXray-*nodule*: Derived from NIH ChestXray[3], wherein only the disease of "nodule" is annotated (sufficient for the task)
- "Lung nodule" accounts for 8.6%, 4,060/47,115 examples on NIH ChestXray
- ChestXray-*nodule* offers AUC score of lung nodule classification equals to 69.1%
- We then progressively add more annotation of other chest diseases



[3] Xiaosong Wang et al. "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2097–2106.

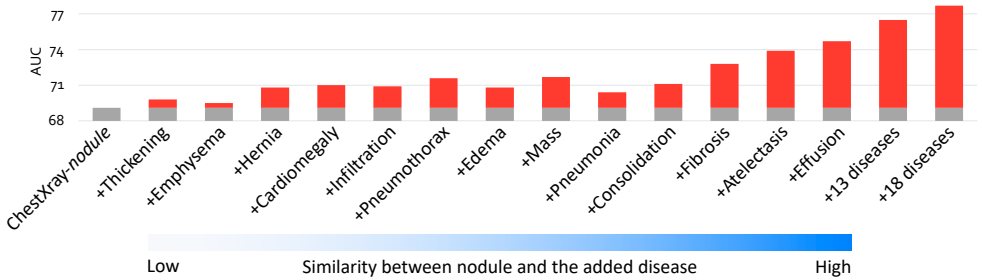
Example II

- The performance of "nodule" classification improves when adding more labels of other chest diseases (non-nodule)
- Learning from additional classes can better delimit the decision boundary of the class of interest (lung nodule)



Example II

- The performance gain (in red) is positively correlated to inter-class similarity ($r = 0.83$; $p = 4.93e\text{-}4$)



Thoughts on Example II

- Annotating "negative examples" facilitates the diagnosis of the interested disease
  - A unique annotation scheme for computer-aided diagnosis of rare diseases and emerging pandemics, where "positive examples" are hard to collect, yet "negative examples" are relatively easier to assemble.
- (I) Assembling existing labels of medically related diseases can improve the classification of the disease of interest
- (II) Assembling existing labels of spatially related organs/diseases can improve the segmentation of the organ/disease of interest
- Wanted: A dataset that is labeled with many classes

Needs for Large Datasets

- Examples I & II stress the need for a dataset that is labeled with many classes
- The creation of large-scale, multi-center, fully-labeled datasets will be fundamental to foster future research in deep learning applied to medical images[4]
- The acquisition and utilization of large-scale labeled datasets are not common amongst medical imaging communities.
- Two main reasons:
- (I) Creating large-scale datasets from scratch requires prohibitively high annotation costs, far exceeding the capability of an individual institute
  - The NLST study involved over 53,000 patients and cost over $250 million[5].
- (II) Sharing medical data involves several privacy concerns

[4] Gabriel Chartrand et al. "Deep learning: a primer for radiologists". In: *Radiographics* 37.7 (2017), pp. 2113–2131.

[5] NLST. "Reduced lung-cancer mortality with low-dose computed tomographic screening". In: *New England Journal of Medicine* 365.5 (2011), pp. 395–409.

Opportunity & Initiative

- Recently, an increasing number of publicly available datasets became available thanks to the collective efforts of imaging data archives and international competitions
- Exactly how such a great number of dissociated datasets can be harnessed and organized is a critical problem
- Need 1→Data 1, Need 2→Data 2, Need 3→Data 3, . . . , Need N→Data N
- Assembling Data 1–N for Needs 1–N
- We introduce a new initiative "data, assemble" that
  - Explores the full potential of an assembly of those datasets with partial labels
  - Ultimately curates a large-scale, fully-labeled dataset

Assembling Your Data (Strictly Non-overlapping)

- Assembling data with strictly non-overlapping labels substantially improves the performance of each class
  - $\mathcal{D}_0$={1, 3, 5, 7, 9} (odd number)
  - $\mathcal{D}_1$={0, 2, 4, 6, 8} (even number)
  - $\mathcal{D}_0$ and $\mathcal{D}_1$ are taken from MNIST

Dataset	$\mathcal{D}_0$	$\mathcal{D}_1$	$\mathcal{D}_0$ & $\mathcal{D}_1$
Num 1	99.6	-	99.6 (↑ 0.0)
Num 3	96.7	-	97.9 (↑ 1.1)
Num 5	96.8	-	97.2 (↑ 0.4)
Num 7	98.4	-	98.8 (↑ 0.4)
Num 9	93.8	-	95.8 (↑ 2.0)
Num 0	-	99.4	99.7 (↑ 0.3)
Num 2	-	97.4	98.4 (↑ 1.0)
Num 4	-	97.6	97.2 (↓ 0.4)
Num 6	-	99.0	99.0 (↑ 0.0)
Num 8	-	91.0	94.2 (↑ 3.2)

Assembling Your Data (Strictly Non-overlapping)

- Assembling data with strictly non-overlapping labels substantially improves the performance of each class
  - $\mathcal{D}_0$={bird, cat, deer, dog, frog, horse} (animal)
  - $\mathcal{D}_1$={airplane, automobile, ship, truck} (transportation)
  - $\mathcal{D}_0$ and $\mathcal{D}_1$ are taken from CIFAR

Dataset	$\mathcal{D}_0$	$\mathcal{D}_1$	$\mathcal{D}_0$ & $\mathcal{D}_1$
Bird	86.4	-	87.4 (↑ 1.0)
Cat	85.1	-	86.1 (↑ 1.0)
Deer	88.9	-	89.9 (↑ 1.0)
Dog	89.4	-	89.4 (↑ 0.0)
Frog	94.3	-	94.8 (↑ 0.5)
Horse	92.8	-	93.7 (↑ 0.9)
Airplane	-	93.0	93.5 (↑ 0.5)
Automobile	-	95.9	96.1 (↑ 0.2)
Ship	-	95.8	95.9 (↑ 0.1)
Truck	-	93.3	94.5 (↑ 1.2)

Assembling Your Data (Strictly Non-overlapping)

- Assembling data with strictly non-overlapping labels substantially improves the performance of each class
  - $\mathcal{D}_0$={Cardiomegaly, Pneumonia, Atelectasis, Edema}
  - $\mathcal{D}_1$={Effusion, Consolidation, Pneumothorax}.
  - $\mathcal{D}_0$ and $\mathcal{D}_1$ are taken from the CheXpert and ChestXray datasets, respectively

Dataset	$\mathcal{D}_0$	$\mathcal{D}_1$	$\mathcal{D}_0$ & $\mathcal{D}_1$
Cardiomegaly	72.7	-	75.5 (↑ 2.8)
Pneumonia	46.3	-	56.3 (↑ 10.0)
Atelectasis	58.0	-	74.9 (↑ 16.9)
Edema	83.0	-	87.3 (↑ 4.3)
Effusion	-	77.1	79.0 (↑ 1.9)
Consolidation	-	67.3	68.7 (↑ 1.4)
Pneumothorax	-	65.2	77.1 (↑ 1.9)

Assembling Your Data (Some Overlapping)

- Assembling data with some overlapping labels can also improve the performance of each class
  - $\mathcal{D}_0$={1, 2, 3, 4, 5}
  - $\mathcal{D}_1$={3, 4, 5, 6, 7}
  - $\mathcal{D}_0$ and $\mathcal{D}_1$ are taken from MNIST

Dataset	$\mathcal{D}_0$	$\mathcal{D}_1$	$\mathcal{D}_0$ & $\mathcal{D}_1$
Num 1	95.6	-	95.9 (↑)
Num 2	92.0	-	92.9 (↑)
Num 3	89.6	89.0	92.8 (↑↑↑)
Num 4	90.6	89.5	94.3 (↑↑↑)
Num 5	87.0	85.2	94.0 (↑↑↑)
Num 6	-	95.2	96.1 (↑)
Num 7	-	93.9	95.0 (↑)

Assembling Your Data (Some Overlapping)

- Assembling data with some overlapping labels can further benefit from advancements in semi-supervised learning (e.g., pseudo labeling and consistency constraints)
- Learning from a mixture of partial labels performs on par with that from full labels
  - $\mathcal{D}_0$={1, 2, 3, 4, 5}
  - $\mathcal{D}_1$={3, 4, 5, 6, 7}
  - $\mathcal{D}_0$ and $\mathcal{D}_1$ are taken from MNIST

n.s.

Dataset	$\mathcal{D}_0$ & $\mathcal{D}_1$ (partial)	$\mathcal{D}_0$ & $\mathcal{D}_1$ (partial)	$\mathcal{D}_0$ & $\mathcal{D}_1$ (full)
Num 1	95.9	98.5 (↑ 2.6)	98.6 (↑ 2.7)
Num 2	92.9	96.5 (↑ 3.6)	96.5 (↑ 3.6)
Num 3	92.8	94.0 (↑ 1.2)	94.3 (↑ 1.5)
Num 4	94.3	94.9 (↑ 0.6)	95.3 (↑ 1.0)
Num 5	94.0	93.3 (↓ 0.7)	93.8 (↓ 0.2)
Num 6	96.1	98.1 (↑ 2.0)	98.0 (↑ 1.9)
Num 7	95.0	96.3 (↑ 1.3)	96.4 (↑ 1.4)

Semi-supervised learning

Assembling Your Data (Some Overlapping)

- Assembling data with some overlapping labels can further benefit from advancements in semi-supervised learning (e.g., pseudo labeling and consistency constraints)
- Learning from a mixture of partial labels performs on par with that from full labels
- Learning from Consolidation and Pneumothorax leads to a noticeable improvement of Cardiomegaly (64.6%→83.9%) and Pneumonia (46.1%→67.9%) classification.
  - $\mathcal{D}_0$={Cardiomegaly, Pneumonia, Atelectasis, Edema, Effusion}
  - $\mathcal{D}_1$={Atelectasis, Edema, Effusion, Consolidation, Pneumothorax}
  - $\mathcal{D}_0$ and $\mathcal{D}_1$ are taken from the CheXpert and ChestXray datasets, respectively

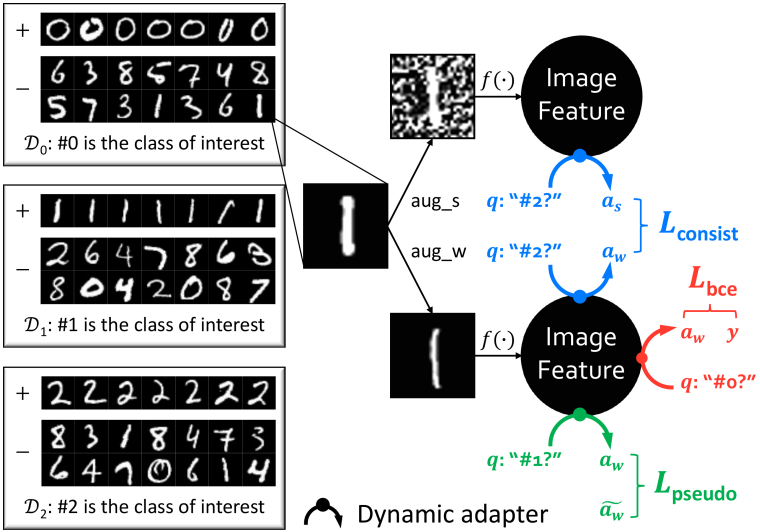
n.s.

Dataset	$\mathcal{D}_0$	$\mathcal{D}_0$ & $\mathcal{D}_1$ (partial)	$\mathcal{D}_0$ & $\mathcal{D}_1$ (partial)	$\mathcal{D}_0$ & $\mathcal{D}_1$ (full)
Cardiomegaly	**64.6**	75.0	**83.9** (↑ 19.3)	83.5 (↑ 18.9)
Pneumonia	**46.1**	62.9	**67.9** (↑ 21.8)	68.3 (↑ 22.2)

Semi-supervised learning

Data, Assemble

- How to exploit the (partially labeled) data assembly? Detailed in
  - Mintong Kang, Yongyi Lu, Alan Yuille, Zongwei Zhou. Data, Assemble: Leveraging Multiple Datasets with Heterogeneous and Partial Labels. https://arxiv.org/pdf/2109.12265.pdf



Data, Assemble

- The initiative of "data, assemble" is critical
- (I) Creating large-scale labeled datasets from scratch for each clinical need is difficult; assembling publicly available data is relatively easier—e.g., AbdomenCT-1K[6]
- (II) Learning from diverse labels can delimit the decision boundary of each individual class and enhance the performance—Example I
- (III) Assembling existing labels of related diseases is a more effective and efficient choice than narrowly pursuing extensive labels for positive examples—Example II

[6] Jun Ma et al. "Abdomenct-1k: Is abdominal organ segmentation a solved problem". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

Data, Assemble: Towards Efficient Medical Image Analysis

Zongwei Zhou, PhD

Johns Hopkins University

October 1, 2021