SK hynix i-TAP
반도체 Data Scientist를 위한
ML/DL 심화 커리큘럼

특 2강

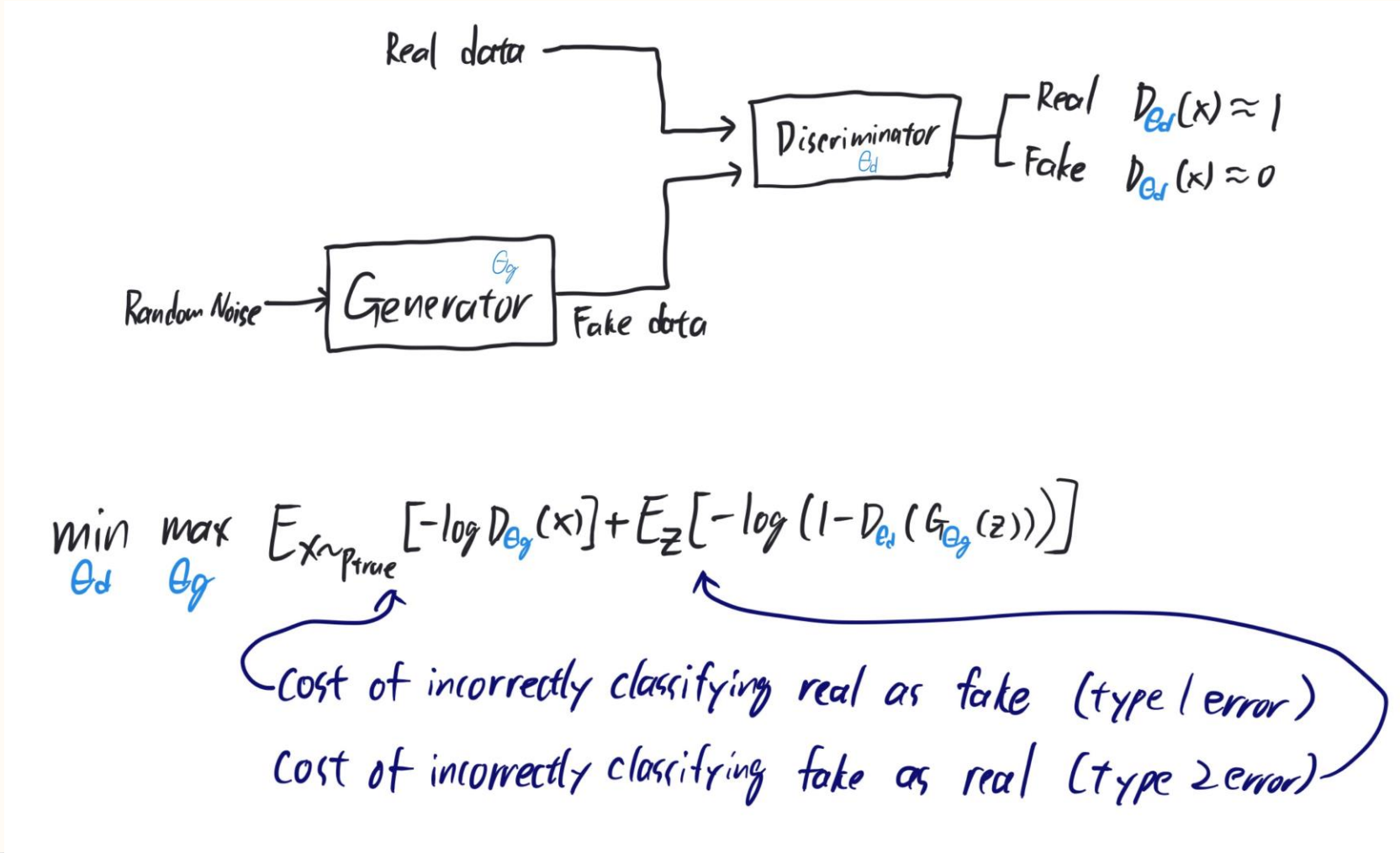
Ernest K. Ryu (류경석)

2020.12.2

특 2강

- GAN-based data augmentation
- Data anonymization
- Transformation-Invariant clustering

Generative Adversarial Networks



Are GANs Useful?

- GANs have produced visually impressive results.
- The GANs itself has no applications.
- However, adversarial training has become incredibly influential.

GAN-Based Data Augmentation

Key idea: Given data $X_1, \ldots, X_N$, (i) train a GAN (ii) use trained generator to create data $\tilde{X}_1, \ldots, \tilde{X}_M$, where $M \gg N$ (iii) train other model $\mathcal{M}$ with the larger dataset $\tilde{X}_1, \ldots, \tilde{X}_M$.

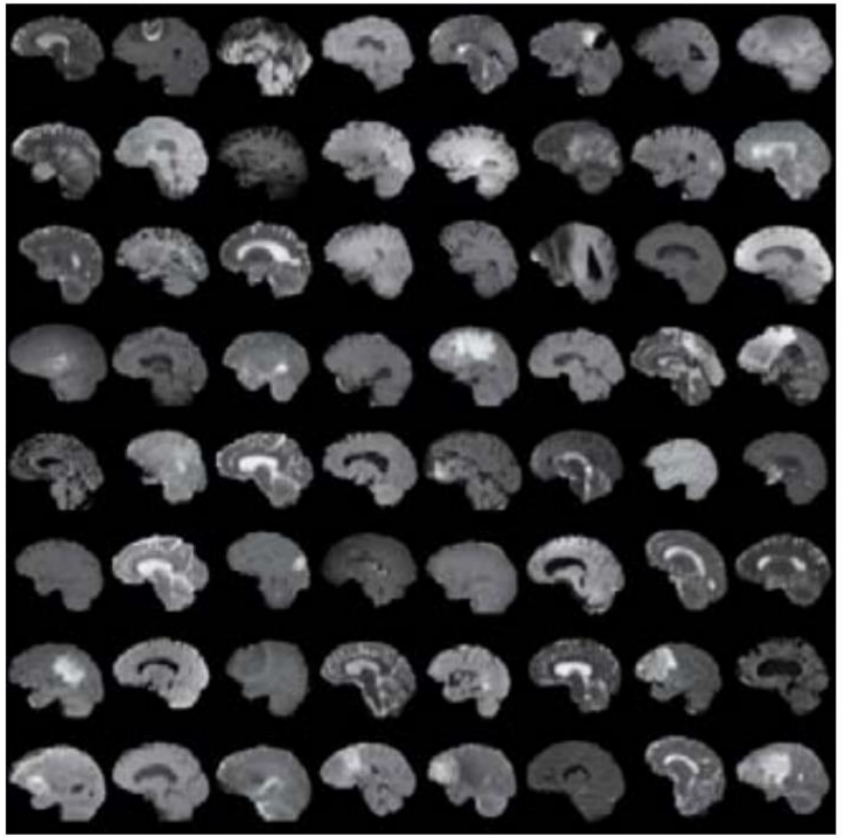
Goal: Achieve performance better than $\mathcal{M}$ trained with $X_1, \ldots, X_N$.

Question: Does this work?

GAN-Based Data Augmentation Failure (Cautionary Tale)

Medical data is very scarce. (Acquisition cost and privacy.) This makes machine learning training difficult.

Key idea: Train GAN on brain MRI images and generate new data.



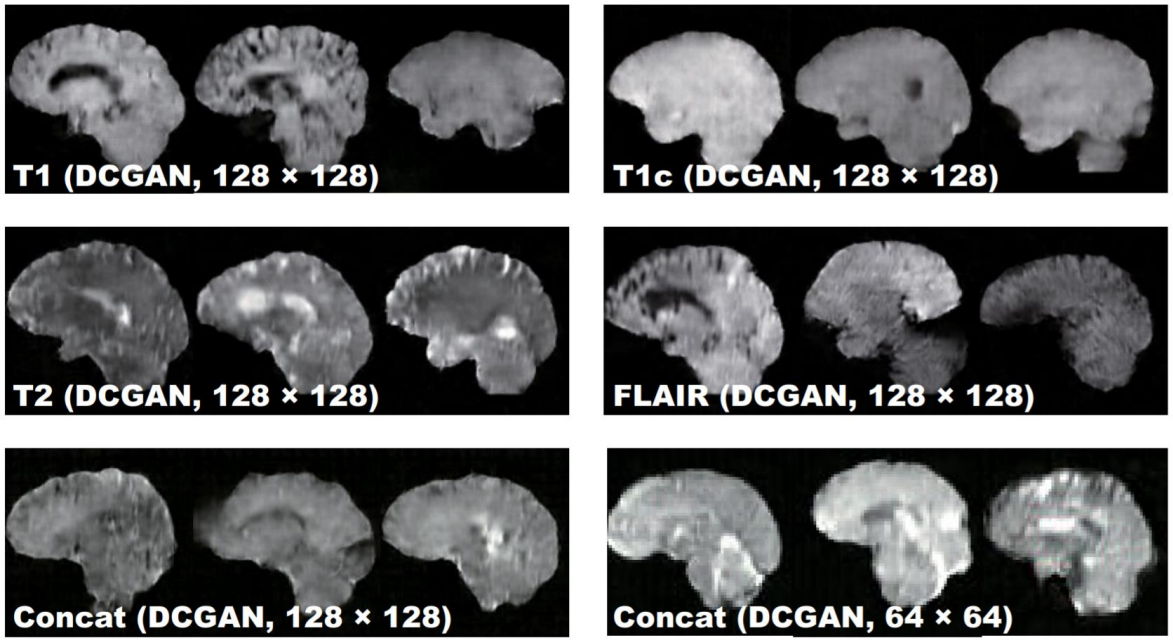
Original Brain MR Images

Han, Hayashi, Rundo, Araki, Shimoda, Muramatsu, Furukawa, Mauri, Nakayama, GAN-based synthetic brain MR image generation, International Symposium on Biomedical Imaging, 2018.

GAN-Based Data Augmentation Failure (Cautionary Tale)

Results: Look good.

Evaluation metric: “Visual Turing test”. Human doctors were unable to distinguish real from fake.



Evaluation metric is fundamentally flawed.
Is the generated data actually augmented? Perhaps they are essentially copies of the originals?
Is the generated data actually useful? Does it help with any application?

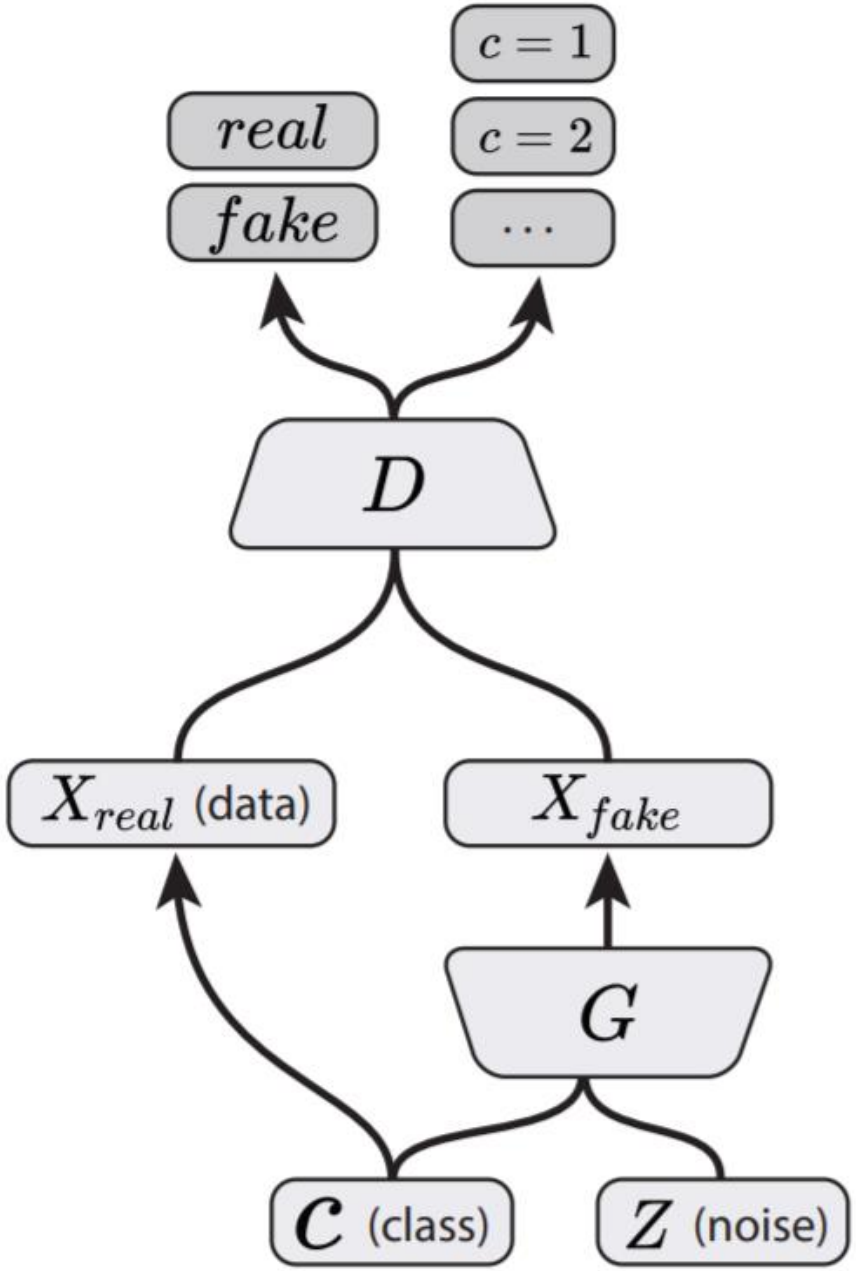
Han, Hayashi, Rundo, Araki, Shimoda, Muramatsu, Furukawa, Mauri, Nakayama, GAN-based synthetic brain MR image generation, International Symposium on Biomedical Imaging, 2018.

GAN with Labels

Goal: Given data $(X_1, Y_1), \ldots, (X_N, Y_N)$, where $Y_1, \ldots, Y_N$ are labels, generate $(\tilde{X}_1, \tilde{Y}_1) \ldots, (\tilde{X}_M, \tilde{Y}_M)$.

Key idea: Make discriminator network also predict labels.

- Auxiliary Classifier GAN (AC-GAN)

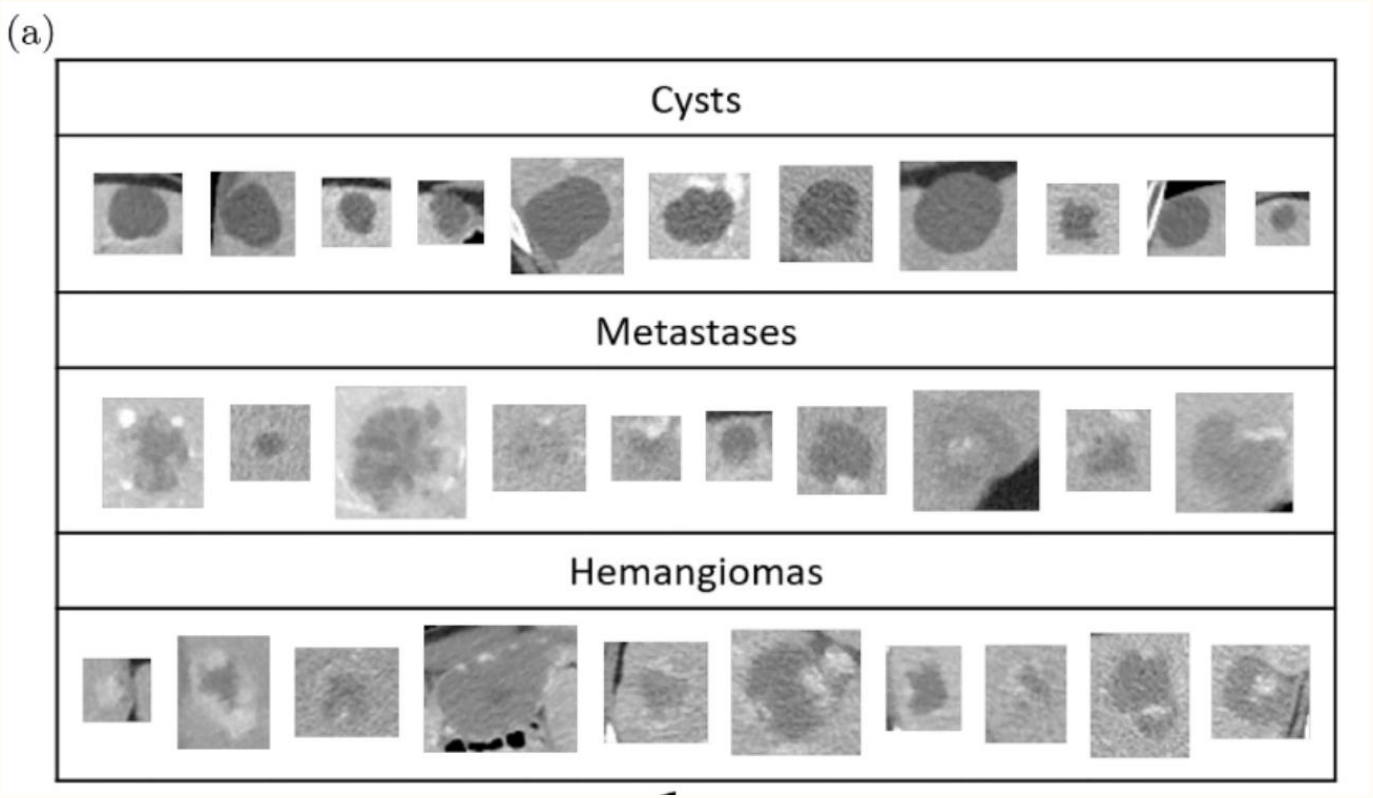


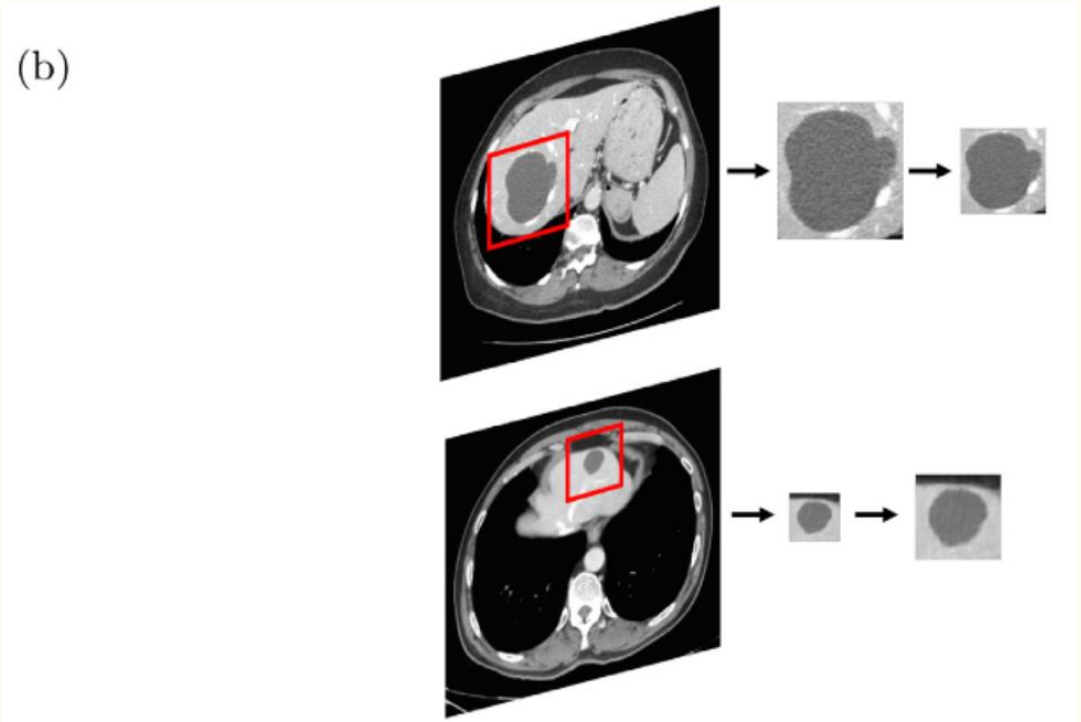
AC-GAN

Conditional Image Synthesis with Auxiliary Classifier GANs, ICML, 2017.

GAN-Based Data Augmentation Success

Real liver lesion data with 3 classes





Frid-Adar, Diamant, Klang, Amitai, Goldberger, Greenspan, GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification

GAN-Based Data Augmentation Success

Train GAN and generate synthetic data

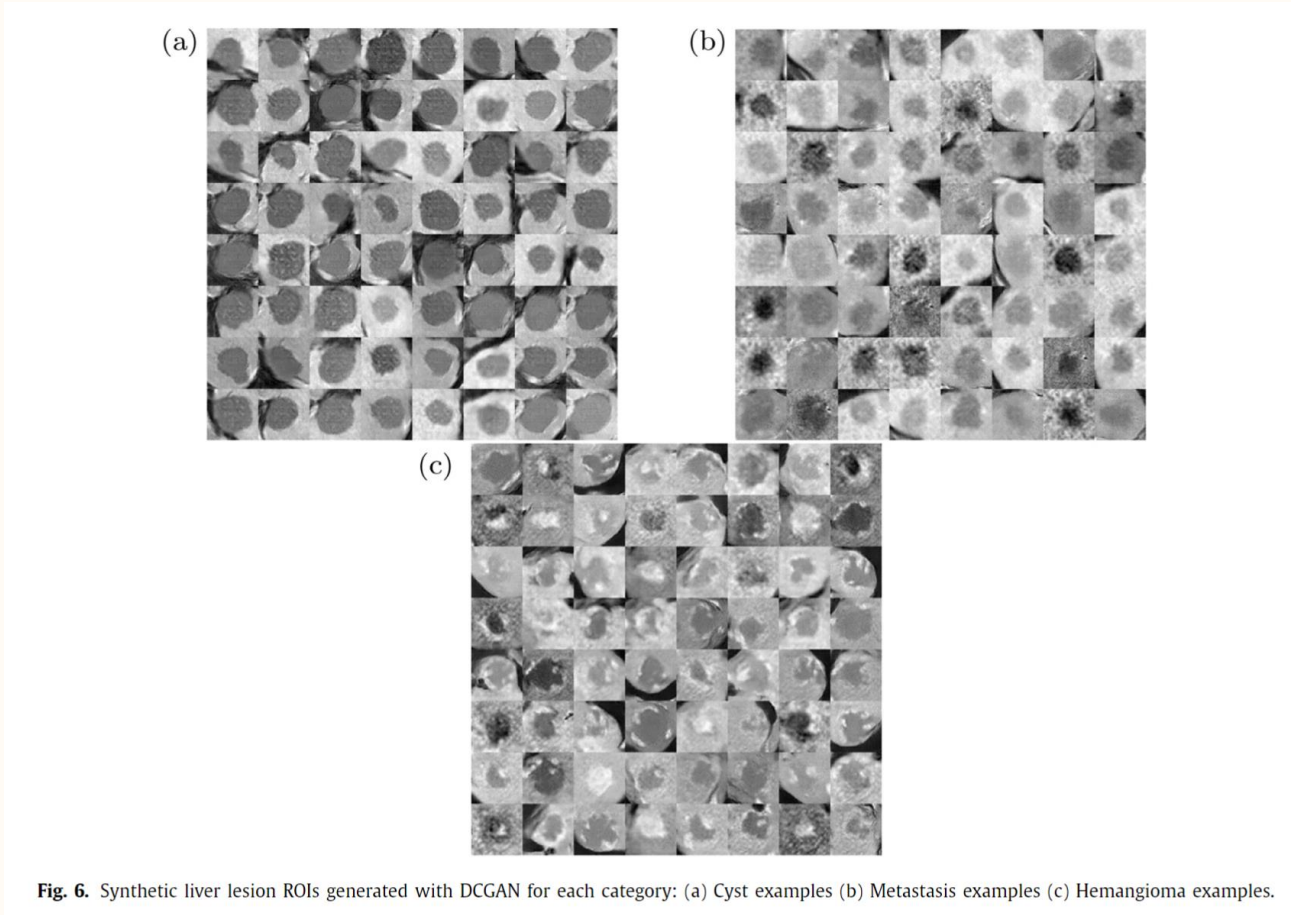


Fig. 6. Synthetic liver lesion ROIs generated with DCGAN for each category: (a) Cyst examples (b) Metastasis examples (c) Hemangioma examples.

Frid-Adar, Diamant, Klang, Amitai, Goldberger, Greenspan, GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification

GAN-Based Data Augmentation Success

Train classifier on synthetic and real data.

Addition of synthetic data improves performance.

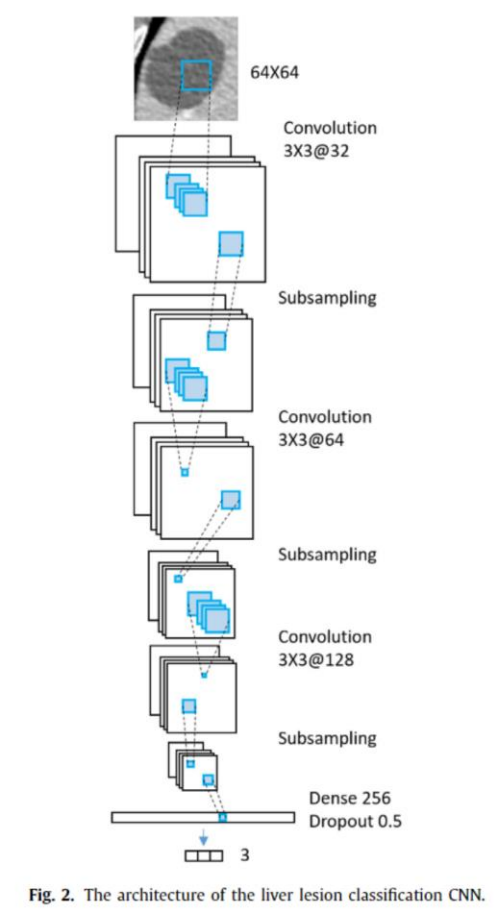
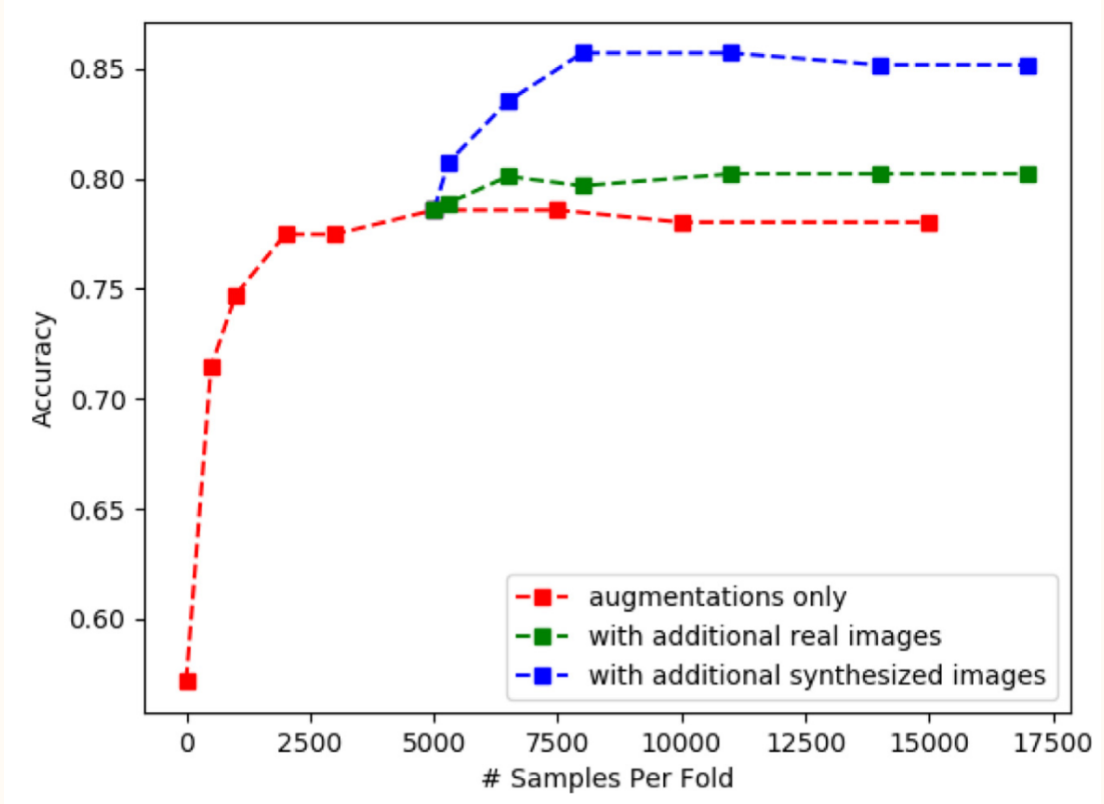


Fig. 2. The architecture of the liver lesion classification CNN.



Is this a success?

- Question: Accuracy is very low. (3 classes 85% accuracy). Perhaps the GAN-based data augmentation is providing a basic regularization effect that other more simpler approaches can provide?

- Similar success stories of GAN-based data augmentation are rare.

- Why should this work? GANs cannot create new information not in the original dataset.

S+U data augmentation

Key idea: Use adversarial training to refine simulated images.

Simulated+Unsupervised (S+U) learning

Best paper award, CVPR, 2017.

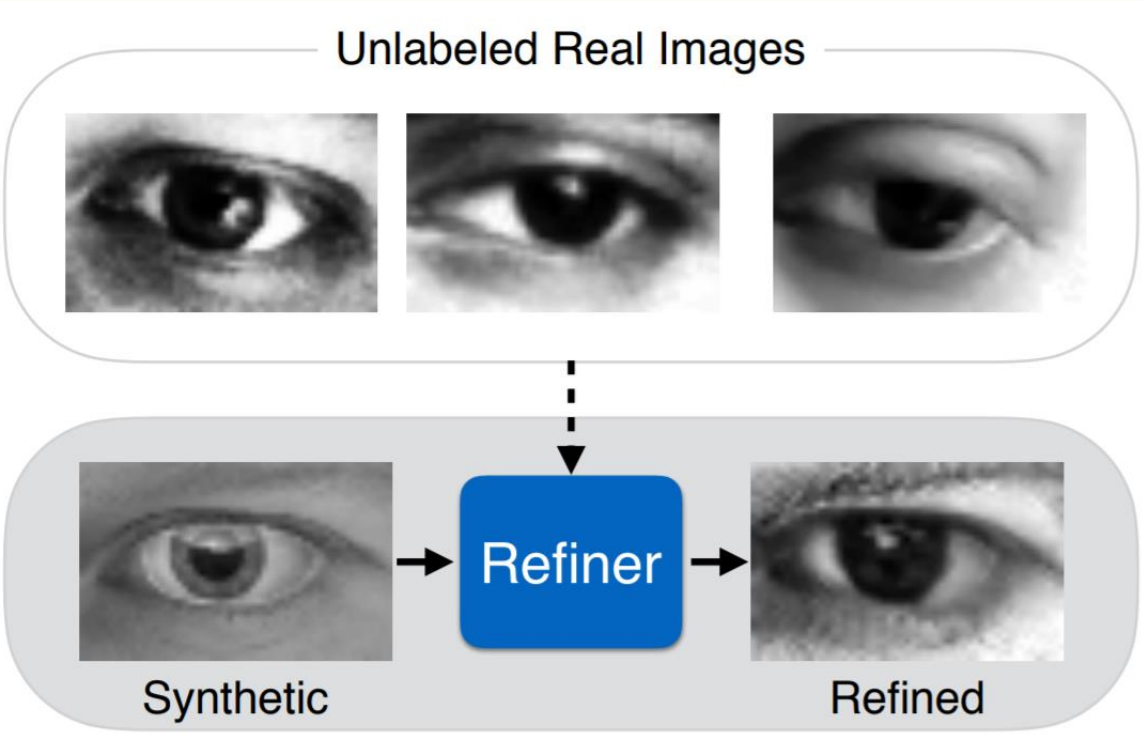
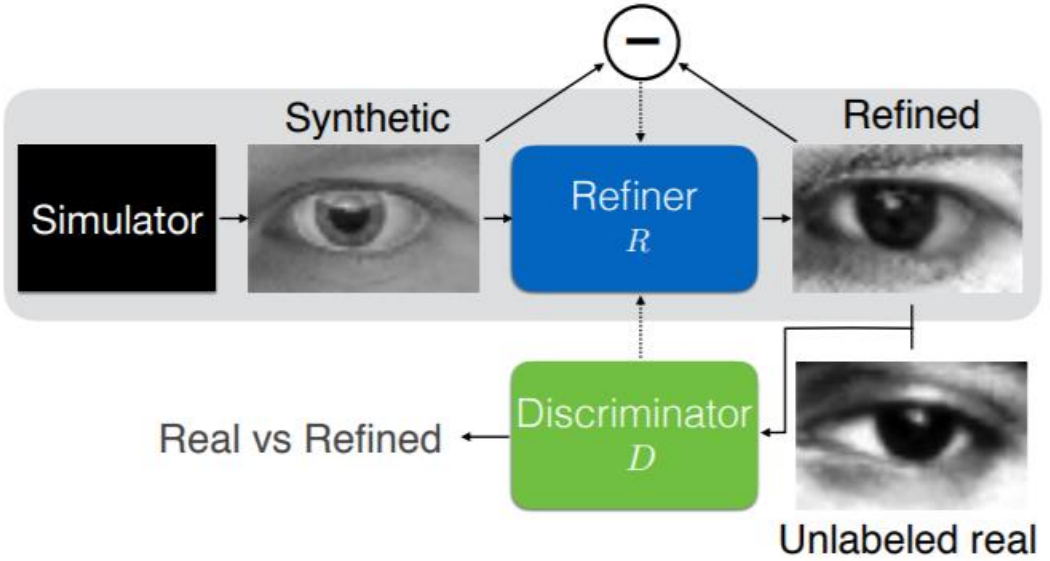


Figure 1. Simulated+Unsupervised (S+U) learning. The task is to learn a model that improves the realism of synthetic images from a simulator using unlabeled real data, while preserving the annotation information.

Shrivastava, Pfister, Tuzel, Susskind, Wang, Webb, Learning from Simulated and Unsupervised Images through Adversarial Training, CVPR, 2017.

S+U data augmentation



$$\mathcal{L}_R(\theta) = \sum_{i=1}^{N} -\log\left(1 - D_\phi(R_\theta(x_i))\right) + \lambda \left\| R_\theta(x_i) - x_i \right\|^2$$

← make D think refined images are real

→ make refined images not too different from original

$$\mathcal{L}_D(\phi) = \sum_{i=1}^{N} -\log\left(D_\phi(R_\theta(x_i))\right) + \sum_{j=1}^{M} -\log\left(1 - D_\phi(y_j)\right)$$

↗ cost of incorrectly classifying refined as real

↑ cost of incorrectly classifying real as refined

Shrivastava, Pfister, Tuzel, Susskind, Wang, Webb, Learning from Simulated and Unsupervised Images through Adversarial Training, CVPR, 2017.

S+U data augmentation

Significant improvement in performance.

Combines domain knowledge used in creating the model of the simulator with the information in the data.

Shrivastava, Pfister, Tuzel, Susskind, Wang, Webb, Learning from Simulated and Unsupervised Images through Adversarial Training, CVPR, 2017.

Data Anonymization

Goals of data anonymization:

- Given dataset $(X_1, Y_1), \ldots, (X_N, Y_N)$, create $(\tilde{X}_1, \tilde{Y}_1), \ldots, (\tilde{X}_N, \tilde{Y}_N)$ so that $(\tilde{X}_1, \tilde{Y}_1), \ldots, (\tilde{X}_N, \tilde{Y}_N)$ does not reveal certain information about the dataset.
- The performance of model $\mathcal{M}$ trained on $(\tilde{X}_1, \tilde{Y}_1), \ldots, (\tilde{X}_N, \tilde{Y}_N)$ is similar to $\mathcal{M}$ trained on $(X_1, Y_1), \ldots, (X_N, Y_N)$.

Question: How secure is a data anonymization strategy?

Anonymization Failure Case Study

Netflix challenge: User i rated movie j. What unseen movie will user i like? Anonymized movie rating dataset released.

Netflix said: "No, all customer identifying information has been removed; all that remains are ratings and dates. This follows our privacy policy [. . .] Even if, for example, you knew all your own ratings and their dates you probably couldn't identify them reliably in the data because only a small sample was included (less than onetenth of our complete dataset) and that data was subject to perturbation."

Attack: Match entries with Internet Movie Database (IMDb). Successfully identified the Netflix records of known users, uncovering their apparent political preferences and other potentially sensitive information.

Narayanan, Shmatikov, Robust De-anonymization of Large Sparse Datasets, IEEE Symposium on Security and Privacy, 2008.

Membership Attack

(i) Alice produces data (X, Y) and gives it to Bob.

(ii) Bob uses Alice's and other people's data to train model $\mathcal{M}$.

(iii) Eve received trained model $\mathcal{M}$ from Bob. Can Eve determine whether Alice's data was used in training $\mathcal{M}$?

Example scenario: Alice tested for a disease, and Bob used the data to train $\mathcal{M}$. Even if Alice tested negative, she may consider the fact that she got tested to be sensitive private information.

Membership Attack on NN

Key idea: Overfitting is a privacy breach.

Attack: If $\mathcal{M}(X)$ is close to a 1-hot vector (very high confidence) then X was in training set. Otherwise, X was not in training set.

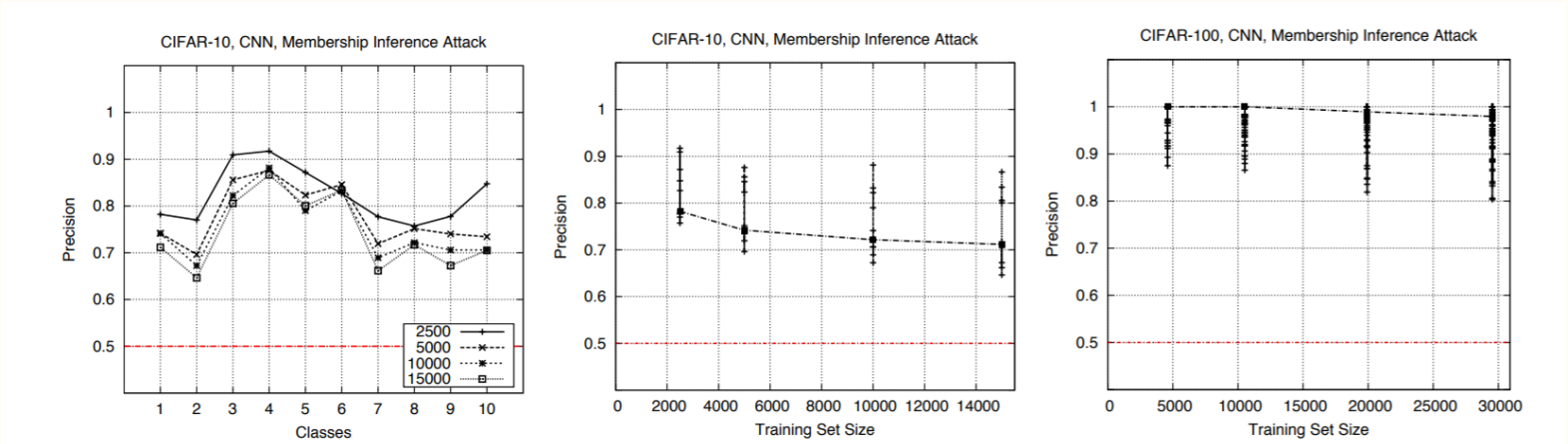


Fig. 4: Precision of the membership inference attack against neural networks trained on CIFAR datasets. The graphs show precision for different classes while varying the size of the training datasets. The median values are connected across different training set sizes. The median precision (from the smallest dataset size to largest) is $0.78, 0.74, 0.72, 0.71$ for CIFAR-10 and $1, 1, 0.98, 0.97$ for CIFAR-100. Recall is almost 1 for both datasets. The figure on the left shows the per-class precision (for CIFAR-10). Random guessing accuracy is 0.5.

Shokri, Stronati, Song, Shmatikov, Membership Inference Attacks against Machine Learning Models, IEEE Symposium on Security and Privacy, 2017.

Anonymization with GANs

Key idea: given data $X_1, \ldots, X_N$

(i) Train a GAN

(ii) Generate data $\tilde{X}_1, \ldots, \tilde{X}_M$ and release synthetic data.

There are ways to do this well, but simple approach has privacy breaches.

Membership Attack on GAN

Key idea: Overfitting is a privacy breach.

Attack with access to $\mathcal{D}$: If $\mathcal{D}(X)$ is close to a 1 (very high confidence that X is real) then X was in training set.

Attack with access to $\mathcal{G}$ or $\tilde{X}_1, \ldots, \tilde{X}_M$: Train $\mathcal{D}$ with synthetic data and brand new data $\hat{X}_1, \ldots, \hat{X}_K$. If $\mathcal{D}(X)$ is close to a 1 (very high confidence that X is synthetic) then X was in training set.

Hayes, Melis, Danezis, De Cristofaro, LOGAN: Membership Inference Attacks Against Generative Models, Proceedings on Privacy Enhancing Technologies, 2019.

Transformation-Invariant Clustering

Classical K-mean clustering

$$\underset{c_1,\ldots,c_k}{\text{minimize}}\, \mathcal{L}(c_1, \ldots, c_k) = \sum_{i=1}^{N} \ell(x_i, \{c_1, \ldots, c_k\})$$

Transformation-invariant clustering by Frey and Jojic

$$\underset{c_1,\ldots,c_k}{\text{minimize}}\, \mathcal{L}(c_1, \ldots, c_k) = \sum_{i=1}^{N} \min_{\beta_1,\ldots,\beta_k} \ell(x_i, \{\mathcal{T}_{\beta_1}(c_1), \ldots, \mathcal{T}_{\beta_k}(c_k)\})$$

Problem: minimizing with respect to $\beta_1, \ldots, \beta_k$ is costly and unreliable.

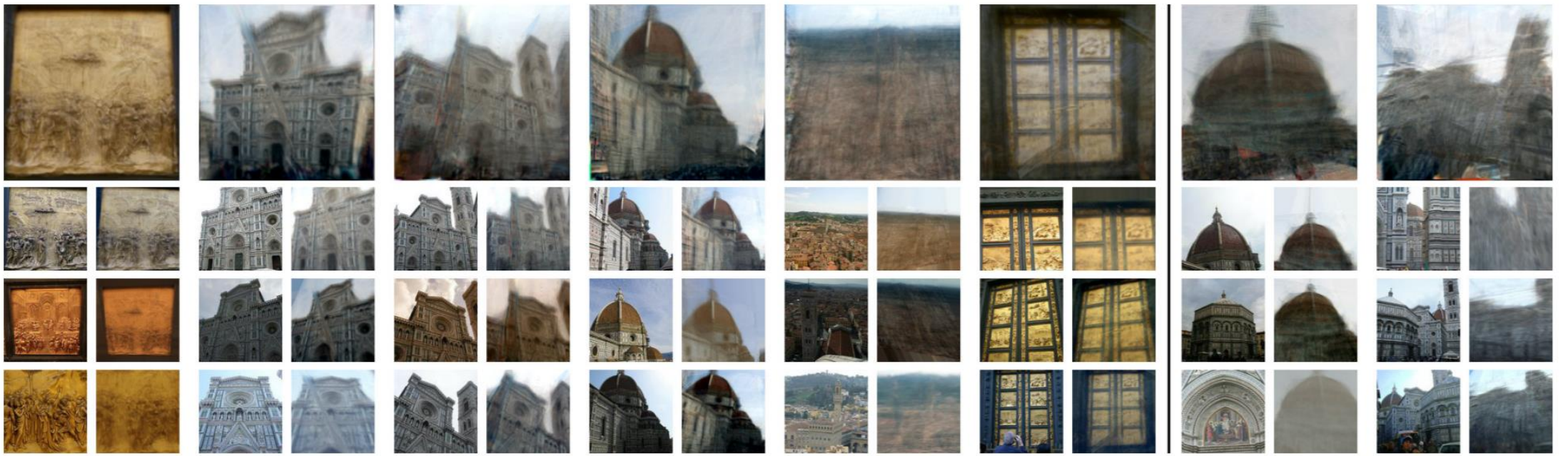
Deep Transformation-Invariant Clustering

Classical K-mean clustering

$$\underset{\theta, c_1, \ldots, c_k}{\text{minimize}}\, \mathcal{L}(\theta, c_1, \ldots, c_k) = \sum_{i=1}^{N} \ell\left(x_i, \left\{\mathcal{T}_{f_{1,\theta}(x_i)}(c_1), \ldots, \mathcal{T}_{f_{k,\theta}(x_i)}(c_k)\right\}\right)$$

$f_{j,\theta}(x_i)$ is a neural network predicting the transformation parameter to match x_i with prototype c_j.

Deep Transformation-Invariant Clustering



(b) Examples of cluster centers and aligned images with DTI GMM (20 clusters)

Deep Transformation-Invariant Clustering

- Paper also considers GMM.
- Paper only studies raw image comparison. Technique can probably be combined with clustering with extracted features.
- Takeaway: If there is a transformation we wish to ignore, allow the clustering algorithm to remove it through a neural network determining how to apply the transformation.

Domain-Adversarial Networks

Let $(X_1, Y_1), \ldots, (X_N, Y_N)$ from distribution $\mathcal{P}$ and $\tilde{X}_1, \ldots, \tilde{X}_M$ from distribution $\mathcal{Q}$ (no labels).

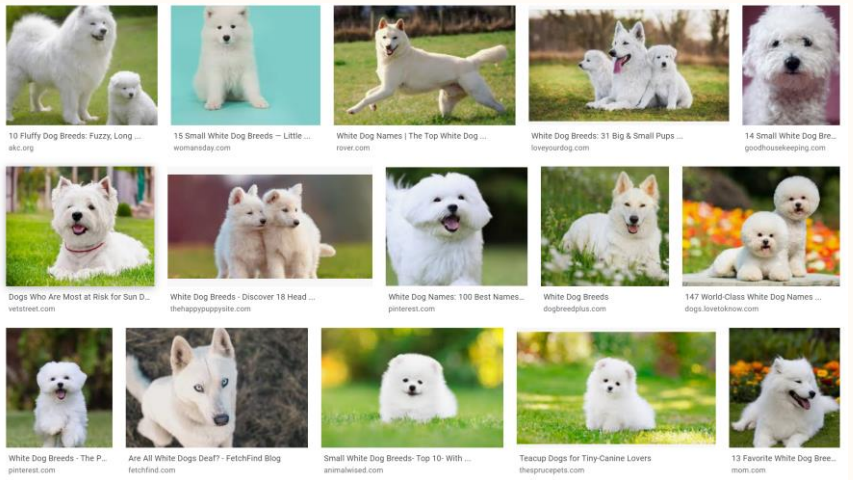
Can we train model $\mathcal{M}$ on $(X_1, Y_1), \ldots, (X_N, Y_N)$ and infer labels on $\tilde{X}_1, \ldots, \tilde{X}_M$? (Can we perform transfer learning?)

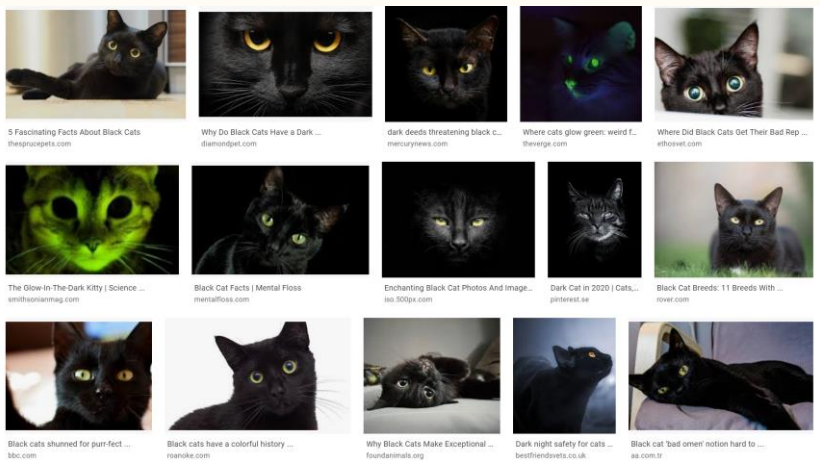
Answer: In general, of course not. But if $\mathcal{P}$ and $\mathcal{Q}$ are somehow related, maybe.

Ganin, Ustinova, Ajakan, Germain, Larochelle, Laviolette, Marchand, Lempitsky, Domain-Adversarial Training of Neural Networks, JMLR, 2016.

Example:

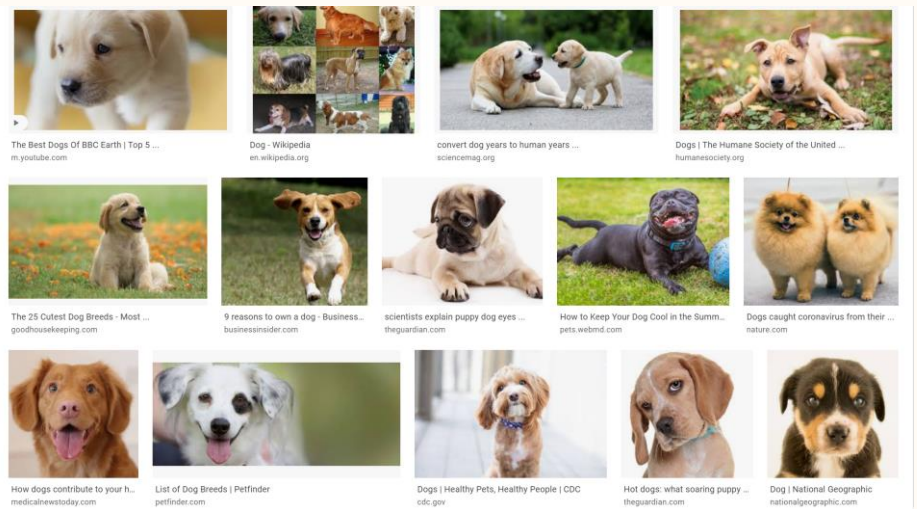
$\mathcal{P} =$

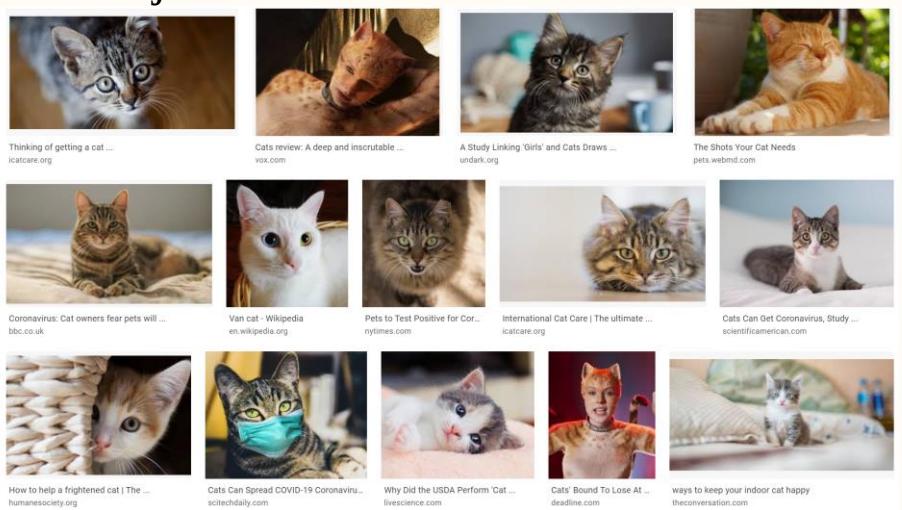




Classifier on $\mathcal{P}$ will look at brightness of object.

$\mathcal{Q} =$

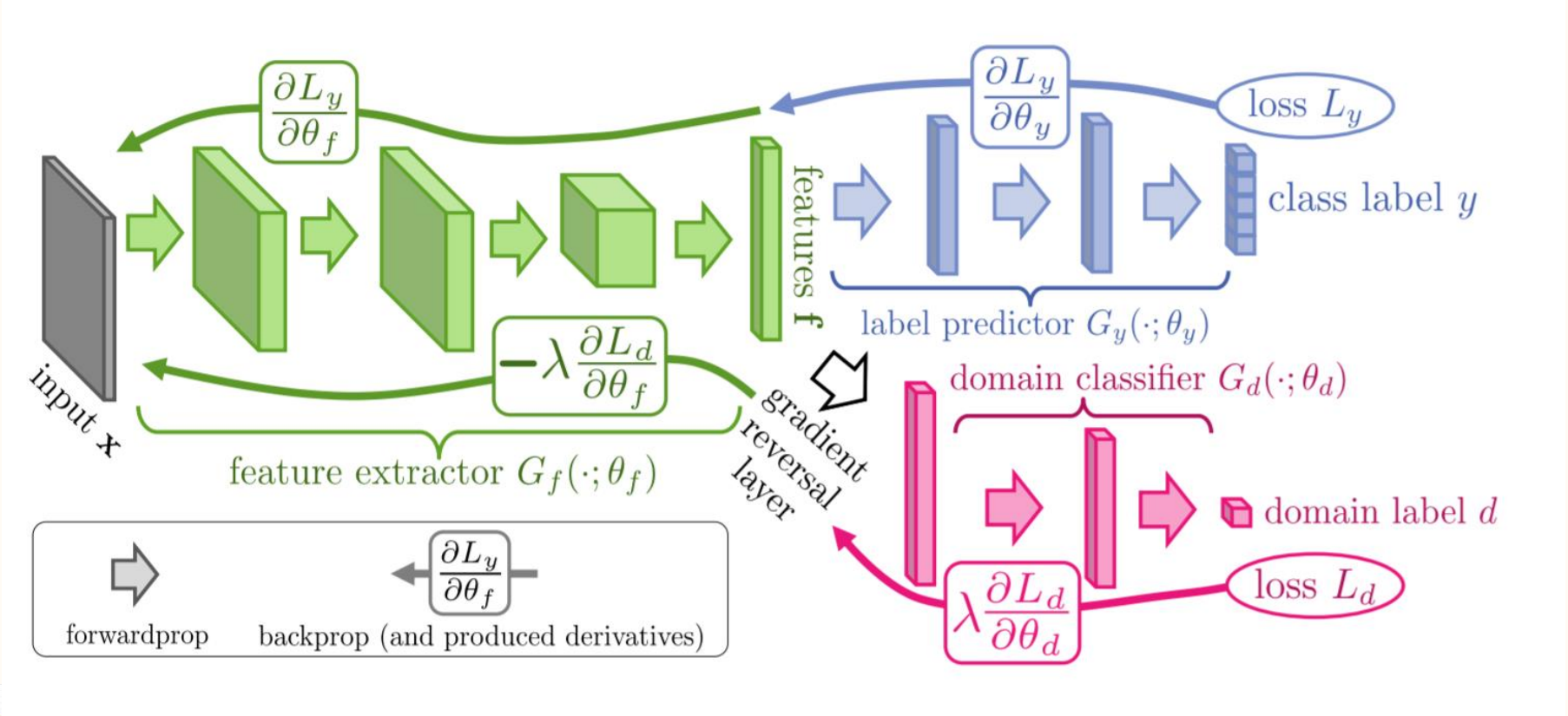




Classifier on $\mathcal{Q}$ cannot focus too much on brightness

Domain-Adversarial Networks

Use adversarial training to extract features f so that a discriminator cannot distinguish $\mathcal{P}$ from $\mathcal{Q}$.Train classifier to make decision only on these extracted features



Ganin, Ustinova, Ajakan, Germain, Larochelle, Laviolette, Marchand, Lempitsky, Domain-Adversarial Training of Neural Networks, JMLR, 2016.

Adversarially Fair Representations

Motivation: we want ML systems to be “fair”.

Example: NN determines whether to approve bank loan. No data scientist would provide skin color as a feature for NN. (Obvious target for lawsuit.) But NN can easily distinguish black people from white people from names, home addresses, music and entertainment preferences, etc.

If NN infers race of applicant and reject black people, that would be unethical. (Unfortunately, race correlates with family wealth and therefore can be a predictor of the probability of not repaying the debt.)

Madras, Creager, Pitassi, Zemel, Learning Adversarially Fair and Transferable Representations, ICML, 2018.

Adversarially Fair Representations

Takeaway: If you have designated set of features you wish to ignore, use adversarial training on the features.