Toward Fusing Domain Knowledge with Generative Adversarial Networks to Improve Supervised Learning for Medical Diagnoses

Fu-Chieh Chang[†], Jocelyn J. Chang[‡], Chun-Nan Chou[†], and Edward Y. Chang[†]

Abstract— **This paper addresses the challenges of small training data in deep learning. We share our experiences in the medical domain and present promises and limitations. In particular, we show through experimental results that GANs are ineffective in generating quality training data to improve supervised learning. We suggest plausible research directions to remedy the problems.**

Index Terms— **Deep learning, knowledge-adaptive GANs, generative adversarial networks, transfer learning.**

I. Introduction

Recent advancements in artificial intelligence (AI) have allowed for novel methods in facilitating medical diagnoses in the healthcare domain (e.g., [17], [26]). Medical diagnosis, the process by which a disease or condition is linked to a patient's corresponding signs and symptoms, can prove challenging because many signs and symptoms are non-specific and occur similarly across multiple disorders. A variety of procedures are therefore employed during the diagnostic process, including pattern recognition, differential diagnosis, medical algorithms, and clinical decision support systems (CDSS), to narrow down the possibilities explaining a patient's condition [20], [35].

Deep learning has the potential to promote the procedure of pattern recognition, further aiding medical diagnoses. Pattern recognition is used to diagnose conditions in which the disease is "obvious" because its correlating set of symptoms is specific [20]. For example, although rashes are a common symptom of many skin disorders, shingles rashes appear in strips on strictly one side of the patient's torso, stopping abruptly at a line along the spine. As a result of the disease's unique pattern, dermatologists can quickly identify shingles without much further testing and prescribe the anti-virals needed to treat the condition. In the case of otitis media (OM) (further discussed in Section II), an inflammatory disease of the middle ear, distinctive visual changes in the eardrum such as redness or calcification can be useful markers in diagnosing OM. In the case of thoracic diseases (further discussed in Section III), chest X-ray images may reveal unique patterns of abnormality between the neck and the abdomen. The specific pathological patterns of OM and thoracic diseases allow for deep learning to learn their features for use in effective diagnoses.

The aim of this paper is to evaluate the promises and limitations of current AI exuberance for pattern-based disease diagnosis. Numerous papers have been published since 2016 at major medical-imaging related conferences [37]. However, most works appear to lack sufficient training data in terms of quantity and diversity, which we argue is required to ensure that data-driven deep learning is effective, useful, and deployable. We use OM and thoracic disease classifications as examples while discussing small-data learning strategies for cases where training data is insufficient.

A. Big Data Powering AI Resurgence

The history of the AI resurgence outside the healthcare domain helps us understand how we may improve supervised learning in the domain. The current level of "intelligence" achieved by the recent wave of AI exuberance is arguably similar to that of the last wave. The last wave of AI exuberance was fueled by the success of IBM Deep Blue, which defeated the reigning world chess champion Garry Kasparov in 1997. The current AI exuberance started with the success of AlexNet [19], but was largely perpetuated by the superior performance of AlphaGo [33]. Both AlphaGo and Deep Blue succeeded for the same key reason: the ability of a computer to evaluate a large number of candidate positions and make the subsequent best decision given the state of the game board.

In Deep Blue and AlphaGo, the intelligence of the system lies in generating virtually all possible "experiences" and evaluating which are valuable to keep. Based on these two well known systems, it appears that a major contributor to AI exuberance is the ability of a system to ensure all possibilities are covered by processing large scales of training data in both volume and diversity. Indeed, when we examine the success of AlexNet, even though both its employed CNN model and SGD algorithm were developed in the 80s, its success was only achieved after ImageNet [8] was available in 2012, when the scale of training data allowed CNNs to be effective.

B. Small Data in Real World

In most real-world scenarios, a large pool of labeled data does not exist. For example in healthcare, although raw data may be abundant, high-quality, high-volume labeled data may not be available for most diseases [28].

While DeepMind has successfully created algorithms to win many games, real world applications are limited due to challenges in synthesizing training data. Take symptom checking or disease diagnosis as examples. The presentation of [5] points out differences in three aspects between diagnosing diseases and playing Go.

- Input certainty: Go's every move during a game provides discrete input with certainty. On the contrary,

[†] HTC Research & Healthcare (DeepQ), Taipei & San Francisco
[‡] Johns Hopkins University, Baltimore

a patient's symptoms can be difficult to explain and quantify (e.g., severity of a headache), and values such as the degree of a fever are real numbers.

- Output possibilities: While a game ends with a win or loss, the possible diseases a patient may have can be n (a typical n is one, but can be two, three, or more in rare cases) out of the 800 possible diseases listed by the CDC.
- Data availability: AlphaGo can self-play to explore previously unknown moves and evaluate their effectiveness. Medicine does not allow for many avenues of exploration; any treatments not FDA certified cannot be evaluated without an approved IRB[1] that ensures clinical safety.

The small data problem has been researched for many decades. One intuition behind the requirement for a large training dataset can be explained by linear algebra. If D variables are to be solved, solving them requires $N = D$ non-colinear equations, where each equation (or image) is a linear combination of variables (or features). When the number of equations or training data is insufficient or $D >> N$, *dimension reduction* attempts to reduce D to D', where $D' \approx N$.

Both linear and non-linear dimension reduction techniques, such as PCA and manifold learning, have not shown to be effective in real-world applications. PCA can embed data in a lower dimensional space, but that space may not be universally good for all target semantics. Manifold learning can learn a sub-space for each target class, but it is difficult to learn a low-dimensional manifold from a small amount of data.

To remedy this dimensionality-curse problem, support vector machines (SVMs) with kernel functions [7] have filled in the gap since the early 90s. SVMs address the $D >> N$ problem by taking advantage of the duality in quadratic optimization. Instead of dealing directly with D features and variables, SVMs deal with N training instances. Regardless how small N is, SVMs form a grand matrix of $N \times N$, which quantifies the pair-wise similarity between N training instances. SVMs convincingly address the small data problem by avoiding the dimensionality curse, and enjoy the global optimal solution that quadratic optimization can solve for when the grand matrix is positive semi-definite (similarity between instances is a non-negative value). SVMs can be considered as a compromise when training data is insufficient.

AlexNet and subsequent CNN models [1], [6] demonstrate that the power of CNNs lies outside of the classification phase (in fact, the classification is merely a logistic regression in its final stage). The key to the success of CNNs is their ability to learn representations from data [4]. The representations learned from big data ($N >> D$) have proven to be able to achieve much higher classification accuracy in several vision tasks. Though a CNN model does not have the explicit notion of dimensionality, one could consider D as the number of weighting parameters that ought to be "learned" from training data of large volume and adequate diversity.

In this article, we discuss two approaches to working with deep learning to make $N' > D$ when the available training instances $N << D$. We discuss related work and present plausible research directions.

- Transferring knowledge from some source domains to the target domain: Section II.
- Generating training data via generative adversarial networks (GANs): Section III.
- Fusing knowledge with GANs to expand diversity of training data: Section IV.

II. Transfer Learning

Transfer learning transfers knowledge learned from some source domains to a target domain. The knowledge learned from the source domains is attained through supervised, unsupervised, or other learning paradigms. The common practice of transfer representation learning is to pre-train a CNN on a very large dataset (called the source domain) and then to use the pre-trained CNN either as an initialization or a fixed feature extractor for the task of interest (called the target domain) [9]. The work of [38] experimentally quantifies transferability of neurons in each layer of a deep CNN. Recently, [40] showed that by using the transfer learning dependencies between various visual tasks in a latent space, the spacial structure of various visual tasks can be modeled.

We used otitis media (OM) diagnosis to perform a case study [6] to understand the effectiveness and shortcomings of transfer learning. We first show the results that have been published in [31]. We then further use these prior results to explain the effectiveness and ineffectiveness of using generative adversarial networks (GANs) to generate training data.

The available training data is comprised of $1,195$ OM images collected by seven otolaryngologists at Cathay General Hospital[2] [30]. The source domain from which representations are transferred to our target disease is ImageNet [8]. The transfer representation learning experiments consist of the following five steps:

1) Unsupervised codebook construction: We learned a codebook from ImageNet images, and this codebook construction is "unsupervised" with respect to OM.
2) Encode OM images using the codebook: Each image was encoded into a weighted combination of the pivots in the codebook. The weighting vector is the feature vector of the input image.
3) Supervised learning: Using the transfer-learned feature vectors, we then employed supervised learning to learn two classifiers from the $1,195$ labeled OM instances.

[1] An IRB is an appropriately constituted group that has been formally designated to review and monitor biomedical research involving human subjects.

[2] The dataset was used under a strict IRB process. The dataset was deleted by April 2015 after our experiments were completed.

TABLE I

OM CLASSIFICATION EXPERIMENTAL RESULTS.

Method	Accuracy(std)	Sensitivity	Specificity
Heuristic w/ seg	80.11%(18.8)	83.33%	75.66%
Heuristic w/o seg	76.19%(17.8)	79.38%	71.74%
Transfer w/ seg (pool5)	87.86%(3.62)	89.72%	86.26%
Transfer w/o seg (pool5)	88.37%(3.41)	89.16%	87.08%
Transfer w/ seg (fc6)	87.58%(3.45)	89.33%	85.04%
Transfer w/o seg (fc6)	88.50%(3.45)	89.63%	86.90%
Transfer w/ seg (fc7)	85.60%(3.45)	87.50%	82.70%
Transfer w/o seg (fc7)	86.90%(3.45)	88.50%	84.90%
Fine-tune	90.96%(0.65)	91.32%	90.20%

4) Feature fusion: We also combined some heuristic features of OM (published in [30]) with features learned via transfer learning.
5) Fine tuning: We further fine-tuned the weights of the CNN using labeled data to improve classification accuracy.

A. *Empirical Study*

Please consult [31] for the detailed algorithm and experimental settings. This section summarizes the results and findings, which can help us compare transfer learning with GANs, discussed in the next section.

Table I compares OM classification results for different feature representations. All experiments were conducted using 10-fold cross validation. The measures of results reflect the discrimination capability of the learned features.

The first two rows in Table I show the results of human-heuristic features, followed by our proposed transfer-learned approach. The eardrum segmentation, denoted as seg, identifies the eardrum by removing OM-irrelevant information such as ear canal and earwax from the OM images [30]. The best accuracy achieved by using human-heuristic methods is 80.11% with segmentation.

Rows three to eight show results of applying transfer representation learning. All results outperform the results shown in rows one and two, suggesting that the features learned from transfer learning are superior to that of human-crafted ones.

Interestingly, segmentation does not help improve accuracy for learning representation via transfer learning. This indicates that the transfer-learned feature set is already discriminative. Among three transfer-learning layer choices (layer five (pool5), layer six (fc6) and layer seven (fc7)), fc6 yields slightly better prediction accuracy for OM. We believe that fc6 provides features that are more general or fundamental to transfer to a novel domain than pool5 and fc7 do. (Section II-B presents qualitative evaluation and explains why fc6 is ideal for OM.)

We also directly used the 1,195 OM images to train a new AlexNet model. The resulting classification accuracy was only 71.8%, much lower than that achieved by applying transfer representation learning. This result confirms our hypothesis that even though CNN is a good model, with merely 1,195 OM images (without the ImageNet images to facilitate feature learning) it cannot learn discriminative features.

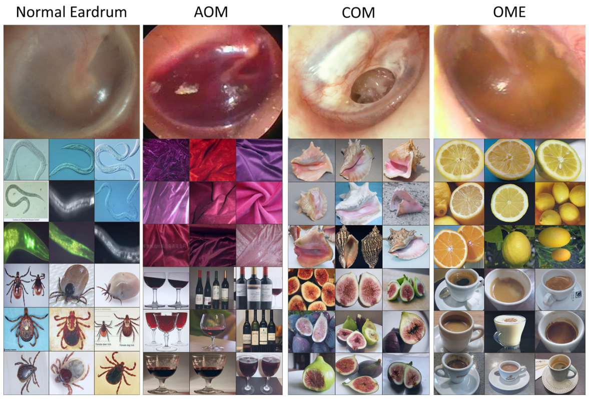


Fig. 1. The visualization of helpful features from different classes corresponding to different OM symptoms (from left to right: Normal eardrum, AOM, COM, and OME).

Finally, we used OM data to fine-tune the AlexNet model, which achieves the highest accuracy. For fine-tuning, we replaced the original fc6, fc7 and fc8 layers with the new ones and used OM data to train the whole network without freezing any parameters. In this way, the leaned features can be refined and are thus more aligned to the targeted task. This result attests that the ability to adapt representations to data is a critical characteristic that makes deep learning superior to other learning algorithms.

B. *Qualitative Evaluation - Visualization*

In order to investigate what kinds of features are transferred or borrowed from the ImageNet dataset, we utilized a visualization tool to perform qualitative evaluation. Specifically, we used an attribute selection method, SVMAttributeEval [13] with Ranker search, to identify the most important features for recognizing OM. We then mapped these important features back to their respective codebook and used the visualization tool from [39] to find the top ImageNet classes causing the high value of these features. By observing the common visual appearances shared by the images of the disease classes and the retrieved top ImageNet classes, we were able to infer the transferred features.

Fig. 1 depicts the qualitative analyses of four different cases: the normal eardrum, Acute Otitis Media (AOM), Chronic Otitis Media (COM) and Otitis Media with Effusion (OME), which we will now proceed to explain in turn:

1) Normal eardrum: Nematodes and ticks are all similarly almost gray in color with a certain degree of transparency.
2) AOM: Purple-red cloths and red wine have deep red colors, which are an obvious common attribute in ears affected by AOM.
3) COM: Seashells have a similar visual texture and color to a calcified eardrum, a prominent symptom of COM.
4) OME: Oranges and lattes possess colors very similar to those of an eardrum affected by OME.

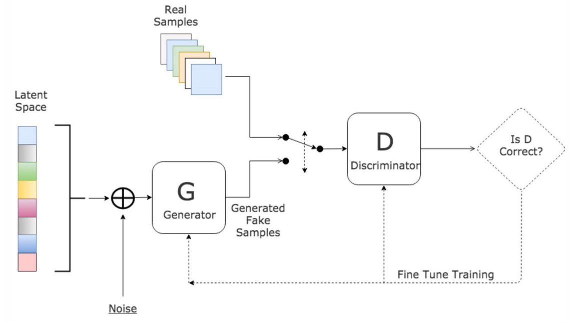


Fig. 2. The Vanilla GANS by [12]; figure credit: Hunter Heidenreich [14].

Many of the visual features of OM are difficult to capture with only hand-crafted methods. Here, transfer learning works to recognize OM in a fashion analogous to how *explicit similes* are used in language to clarify meaning. The purpose of a simile is to provide information about an object by comparing it to another object with which one is more familiar. For instance, if a doctor says that OM displays redness and certain textures, a patient may not be able to comprehend the doctor's description exactly. However, if the doctor explains that OM presents with an appearance similar to that of a seashell, or with colors similar to that of red wine, oranges, or lattes, the patient is conceivably able to envision the appearance of OM at a much more precise level. At level fc6, transfer representation learning works like finding similes that can help explain OM using the representations learned in the source domain (ImageNet).

In summary, transfer representation learning can potentially remedy two major challenges of medical image analysis: labeled data scarcity and medical domain knowledge shortage. Representations of OM images can be effectively learned via transfer learning.

III. Generative Adversarial Networks (GANs)

Generative adversarial networks (GANs) [12] are a special type of neural network model where two networks are trained simultaneously. Figure 2 depicts that the generator (denoted as G) focuses on producing fake images and the discriminator (denoted as D) centers on discriminating fake from real. The goal is for the generator to produce fake images that can fool the discriminator to believe they are real. If an attempt fails, GANs use backpropagation to adjust network parameters. Since the introduction of the initial GAN model [12], there have been several variants depending on how the input, output, and error functions are modeled. GANs can be primarily divided into four representative categories based on the input and output (the error function is discussed in Section III-A), and their applications are as follows:

- Conditional GAN (CGAN) [22]: CGAN adds to GAN an additional input, y, on which the models can be conditioned. Input y can be of any type, e.g., class labels. Conditioning can be achieved by feeding y to both the generator $G(z|y)$ and the discriminator $D(x|y)$, where x is a training instance and z is random noise in latent space. The benefit of conditioning on class labels is that it allows the generator to generate images of a particular class. (Application: text to image.)
- Pixel-to-Pixel GAN (Pix2Pix) [16]: Pix2Pix GAN is similar to CGAN. However, conditions are placed upon an image instead of a label. The effect of such conditioning is that it allows the generator to map images of one style to another, e.g. mapping a photo to the painting style of an artist or mapping a sketch to a colored image. (Application: image to image translation, supervised.)
- Progressive-Growing GAN (PGGAN) [18]: PGGAN grows both the generator and discriminator progressively; starting from low resolution, it adds new layers that model increasingly fine details as training progresses. PGGAN can generate high-resolution images through progressive refinement. (Application: high-resolution image generation.)
- Cycle GAN [42]: Pix2Pix GAN requires paired training data to train. Cycle GAN is an unsupervised approach for learning to translate an image from a source domain X to a target domain Y without training examples. The goal is to learn two mappings from X to Y (i.e., G) and from Y to X (i.e., F) such that the distributions $G(X)$ is indistinguishable from the distribution Y, and the distributions $F(Y)$ is indistinguishable from the distribution X, respectively. Cycle GAN introduces a cycle consistency loss to approximate $F(G(X))$ to X and also $G(F(Y))$ to Y. (Application: image to image translation, unsupervised.)

A. Shortcomings of GANs

Though GANs have demonstrated interesting results, there are both micro and macro research issues that need to be addressed.

The micro issues are related to the formulation of the model's loss function to achieve good generalization. But this generalization goal has been cast into doubt by the empirical study of [3], which concludes that training of GANs may not result in good generalization properties.

The GAN loss formulation is regarded as a saddle point optimization problem and training of the GAN is often accomplished by gradient-based methods [12]. G and D are trained alternatively so that they evolve together. However, there is no guarantee of balance between the training of G and D with the KL divergence. As a consequence, one network may inevitably be more powerful than the other, which in most cases, is D. When D becomes too strong in comparison to G, the generated samples become too easy to differentiate from real ones. Another well known issue is that the two distributions are in high probability located in disjoint lower dimensional manifolds without overlaps. The work of WGAN [2] addresses this issue by introducing the Wasserstein distance. However, WGAN still suffers from unstable training, slow convergence after weight clipping (when the clipping window is too large), and vanishing gradients (when the clipping window is too small). Whereas the above micro issues have been studied by the community in order to propose novel models and optimization techniques, we are more concerned with the macro issue: can GANs help generate large-volume and diversified training data to improve validation and testing accuracy? As stated in the introductory section, deep learning depends on the scale of training data to succeed, but most applications do not have ample training data.

Specifically in medical imaging, GANs have been mainly used in five areas: image reconstruction, synthesis, segmentation, registration, and classification, with hundreds of papers published since 2016 [37]. A recent report [28] summarizes the state of applied AI in the field of radiology and conveys that promising results have been demonstrated, but the key challenge of data curation in collection, annotation, and management remains. The work of [10] uses GANs to generate additional samples for liver lesion classification and claims that both the sensitivity and specificity are improved. However, the total number of labeled images is merely 182, which is too small a dataset to draw any convincing conclusions. The work [29] applies a similar idea to thoracic disease classification and achieves better performance. The work uses human experts to remove noisy data, but fails to report how many noisy instances were removed and how much of the accuracy improvement was attributed to human intervention. The paper also claims that additional data contributes in making training data of all classes balanced to mitigate the imbalanced training data issue. Had the work demonstrated that generating additional data using GANs helps despite imbalanced distribution, the improved result would have been more convincing.

Combining 3D model simulation with GANs seems to be another plausible alternative to reaching the same goal of increasing training instances. The work of [34] presents a framework that can generate a large amount of labeled data by combining a 3D model with GANs. Another work [32] combines a 3D simulator (with labels) with unsupervised learning to learn a GAN model that can improve the realism of the simulating labeled data. However, this combining scheme does not work for some tasks. For example, our AR platform Aristo [41] experimented with these methods and did not yield any accuracy improvements in its gesture recognition task. Moreover, most medical conditions have lacked exact 3D models so far, which makes the combining scheme difficult to apply.

B. Empirical Study

This section reports our experiments in generating training data using GANs to improve the accuracy of supervised learning.

Section II shows that adding images unrelated to OM can improve classification accuracy due to representation transfer in the lower layers of the model and representation analogy in the middle layers of the model. This leads us to the following questions: Can GANs produce useful labeled data to improve classification accuracy? If so, which CNN layers can GANs strengthen to achieve the goal and how do GANs achieve this classification accuracy improvement? Our experiments were designed to answer these questions.

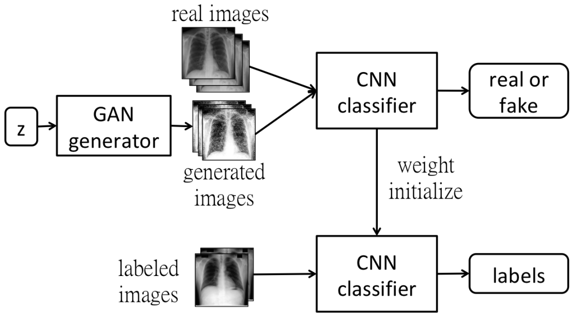


Fig. 3. Pre-Trained on generated images

1) Experiment Setup: We used the NIH Chest X-ray 14 [36] dataset to conduct our experiments. This dataset consists of 112,120 labeled chest X-ray images, from over 30,000 unique patients corresponding to 14 common thoracic disease types, including atelectasis, cardiomegaly, effusion, infiltration, mass, nodule, pneumonia, pneumothorax, consolidation, edema, emphysema, fibrosis, pleural thickening, and hernia. The dataset is divided into training, validation, and testing sets, containing 78,468, 11,219. and 22,433 images, respectively[3]. Our experiments were designed to examine and compare four training methods:

1) Random initialization: Model parameters were randomly initialized.
2) *Pre-trained by using ImageNet*: Similar to what we did with transfer learning in Section II, the network was pre-trained by using ImageNet.
3) *Pre-trained with additional data generated by unsupervised-GAN*: The method is shown in Figure 3. First, the GAN generated the same number of fake images as we had real images. Second, the CNN classifier was trained to differentiate between real and fake images. Third, the weights were used to initialize the subsequent classification task.
4) *Trained with additional data generated by supervised-GAN*: By adding the generated images, the size of the dataset was expanded to $2x$ and $5x$ (the size of original dataset is x). In order to show whether GAN can produce labeled data to directly improve classification accuracy instead of indirectly, we changed the configuration of GAN in Method 3 so that it could generate labeled images.

To establish a yardstick for these four methods, we first measured the "golden" results that supervised learning can attain using 100% training and validation data. We then dialed back the size of the training and validation data to be 50%, 20%, 10%, and then 5%. We used each of the four methods to either increase training data or pre-train the network. We used PGGAN[4] as our GAN model to generate images with 1024×1024 pixel resolution. For

[3] We followed the dataset splits in `https://github.com/zoogzog/chexnet/tree/master/dataset`

[4] We used a publicly available implementation of PGGAN via `https://github.com/tkarras/progressive_growing_of_gans`. This implementation has an auxiliary classifier [24] and hence can generate images conditionally (for Method 4) or unconditionally (for Method 3).

our CNN classifier, we employed DenseNet121 [15], and used AUROC[5] as our evaluation metric. Intuitively, our conjectures before seeing the results were as follows:

- Method 1 will perform the worst, since it does not receive any help to improve model parameters.
- Method 4 will perform the best, since it produces more training instances for each target class.
- Method 3 will outperform 2 as the training data generated, though unlabeled, is more relevant to the target disease images than ImageNet is.

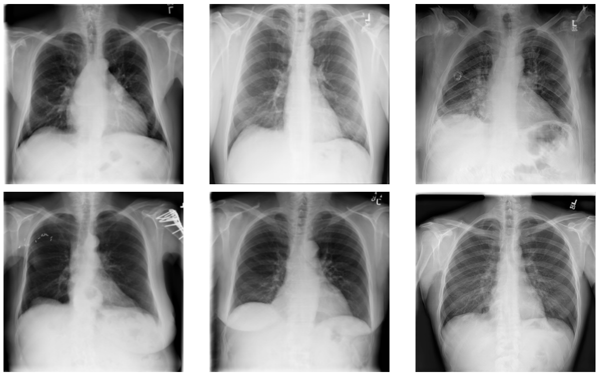
2) Experiment Results: Table II presents our experimental results. We report the AUROC of detecting 14 thoracic disease types using each of the four different training methods. These results are inconsistent with our conjectures:

- Method 2, which is equivalent to transfer learning, performs the best. No methods using GANs were able to outperform this method.
- Method 4 performs the worst. In Method 4, additional GAN-generated labeled images were used to perform training. We believe that the labeled images generated using GANs were too noisy. Therefore, when the generated images are increased ($5x$ vs. $2x$), the prediction accuracy is not always increased and sometimes even worse. This suggests that GANs do not produce helpful training instances and may in fact be counter-productive.
- Method 3 does not outperform method 2, even though ImageNet data used by method 2 is entirely irrelevant to images of thoracic conditions. We believe that the additional images generated by GANs used for initializing network parameters are less useful because of their low volume and diversity. After all, adding more low-quality similar images to an unlabeled pool cannot help the model learn novel features. Note that a recent keynote of I. Goodfellow [11] points out that GANs can successfully generate more unlabeled data (not labeled data) to improve MNIST classification accuracy. Table II reflects the same conclusion that method 3 outperforms method 1, which uses randomly-initialized weights. However, using GANs to generate unlabeled data may not be more productive than using ImageNet to pre-train the network.

Figure 4 samples real and GAN-generated images. The first column presents real images, the second column GAN-generated unsupervised, and the third GAN-generated supervised. The GAN-generated images may successfully fool our colleagues with no medical knowledge. However, as reported in [29], the GAN-generated labeled chest X-ray images must be screened by a team of radiologists to remove erroneous data (with respect to diagnosis knowledge). Without domain knowledge, incorrectly labeled images may be introduced by GANs into the training pool, which would degrade classification accuracy.

TABLE II
EXPERIMENTAL RESULTS

	Scale of Dataset				
	5%	10%	20%	50%	100%
Methods	AUROC (std)				
Method 1	0.708 (0.020)	0.757 (0.003)	0.780 (0.004)	0.807 (0.002)	0.829 (0.000)
Method 2	0.756 (0.006)	0.790 (0.002)	0.807 (0.005)	0.832 (0.001)	0.843 (0.000)
Method 3	0.726 (0.002)	0.765 (0.004)	0.789 (0.001)	0.817 (0.002)	0.828 (0.000)
Method 4 (2x)	0.713 (0.003)	0.724 (0.004)	0.768 (0.004)	0.809 (0.001)	0.824 (0.000)
Method 4 (5x)	0.693 (0.005)	0.727 (0.002)	0.774 (0.005)	0.798 (0.005)	0.813 (0.000)



Real GAN unsupervised GAN supervised

Fig. 4. Real vs. GAN-generated Images.

In summary, the study of [21] shows that pre-training with datasets that are multiple orders of magnitude larger than ImageNet can achieve higher performance than pre-training with only ImageNet on several image classification and object detection tasks. This result further attests that volume and diversity of data, even if unlabeled, helps improve accuracy. GANs may indeed achieve volume, but certainly cannot achieve diversity.

To explain why using ImageNet can achieve better pre-training performance than that achieved when using GAN-generated images, we perform layer visualizations using the technique introduced in [25]. Figure 5 plots the output layer of the first dense-block of DenseNet. Row one shows five filters of untrained randomly initialized weights. Row three shows five filters with more distinct features learned from the ImageNet pre-trained model. The unsupervised-GAN method (row two) produces filters of similar quality to that of row one. Qualitatively, unsupervised-GAN learns similar features akin to how the random-initialization method does, and does not yield more promising classification accuracy.

IV. Fusing Knowledge with GANs

The desired outcome of GANs after training is that samples formed by x_g approximate the real data distribution $pr(x)$. However, if the real data distribution is under-represented by the training data, the generated samples cannot "guess" beyond the training data. For instance, if the otitis media (OM) training data shown in Section II consists of only one type of OM, say AOM, GANs cannot generate the other two types of OM, COM and OME. As

[5] We used a publicly available implementation of ChexNet [27] from `https://github.com/zoogzog/chexnet`, which contains a DenseNet121 classifier, and used its evaluation metric. The metric is deriving by first summing up all AUROCs from each of the 14 classes and then dividing the summation by 14.

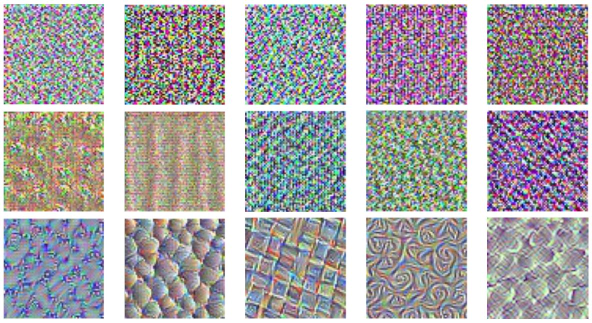


Fig. 5. CNN layer visualization of the first denseblock of DenseNet121. The top row is random weight, the second row is pre-trained by unsupervised-GAN method, and the third row is pre-trained by ImageNet.

another example, if a set of training data consists of a large number of red roses, and the aim of GANs is to generate entire categories of different colored roses, there would be no *knowledge* or *hint* for *G* or *D* to respectively achieve and tolerate diversity in color. On the contrary, the discriminator *D* would reject any roses that are not red and *G* would not be encouraged to expand beyond generating red colored roses. The nature of GANs treats exploration beyond the paradigm of the seen or known to be erroneous.

If we would like GANs to generate diversified samples to improve supervised learning, the new approach must address two issues:

- Guiding the generator to explore diversity *productively*.
- Allowing the discriminator to tolerate diversity *reasonably*.

The adverbs *productively reasonably*, convey exploration (beyond exploitation) with guidance and rules. In the case of playing games, rules and rewards are clear. In the case of generating roses beyond red colors or generating types of flowers beyond roses, guidance and rules are difficult to articulate. Supposing computer vision techniques can precisely segment petals of roses in an image, what colors can the generator use to replace red petals? For example, black roses do not exist, so this color would be deemed unreasonable and unproductive for generating realistic rose images. Exploration beyond training distribution should be permitted, but at the same time guided by knowledge. How can knowledge be incorporated into training GANs? We enumerate two schemes.

1) Incorporating a human in the loop: Placing a human in the loop instead of letting function *D* make the decision can ensure *D* is properly adjusted, due to human input. The work of [29] discussed in Section III implements a GAN to generate labeled chest X-ray images and then asks a team of radiologists to remove mislabeled images. We believe that merely removing "bad" images without productively generating new images with novel disease patterns may provide only limited help.
2) Encoding knowledge into GANs: We can convey to GANs knowledge about the information to be modeled via the knowledge layers/structures and/or via the knowledge graph/dictionary using natural language processing. We elaborate this scheme in the remainder of this section.

A. *Information from Knowledge Layers/Structures*

Considering the structure of information may improve the effectiveness of GANs. For instance, differentiating two types of strokes, ischemic and hemorrhagic, in order to provide proper treatment is critical for patient recovery. Ischemic strokes occur as a result of an obstruction within a blood vessel supplying blood to the brain. It accounts for 87 percent of all stroke cases. Hemorrhagic strokes occur when a weakened blood vessel ruptures inside or on the surface of the brain. Two types of weakened blood vessels usually cause hemorrhagic stroke: aneurysms and arteriovenous malformations (AVMs).

Without the above knowledge, GANs could generate data that flips the appearance of ischemic versus hemorrhagic strokes, which would blur the critical ability to differentiate between the two. Additionally, without knowledge of brain anatomy, GANs could generate obstructions and ruptures in clearly erroneous locations where no blood vessels are present. With the knowledge that the symptoms largely occur within and on blood vessels, multi-layer GANs may be able to impose anatomical constrains through layering information.

B. *Information from Knowledge Graph/Dictionary*

The possible colors of roses can be obtained from the following Wikipedia text via natural language processing (NLP) parsing:

"Rose flowers have always been available in a number of colours and shades; they are also available in a number of colour mixes in one flower. Breeders have been able to widen this range through all the options available with the range of pigments in the species. This gives us *yellow*, *orange*, *pink*, *red*, *white* and *many combinations of these colours*. However, they lack the *blue pigment* that would give a true *purple* or blue colour and until the 21st century all true blue flowers were created using some form of dye. Now, however, genetic modification is introducing the blue pigment."

Once possible colors and their combinations have been extracted using NLP, we can enhance the idea of text-adaptive GANs [23] to generate roses of these colors. The current text-adaptive GANs may borrow colors from any flower samples in the training pool, and this exhibits two problems. The flower colors in the training pool may be a superset or subset of rose colors. Text-adaptive GANs do not support exploration with knowledge as guidance.

V. Conclusions

This paper presented the challenges in using GANs to generate additional labeled data to improve the performance of supervised learning. Based on prior work in GANs and our case studies using transfer learning and GANs, we found the additional data generated by GANs may be counter-productive for improving supervised learning tasks. This is partly because the GAN generator cannot generate data of different patterns not seen in the training data, and partly because the GAN discriminator cannot tolerate new patterns not seen in the training data. We used OM and thoracic disease type classifications as clinical examples, and generated different types of roses from training data of a single rose color as an additional example to illustrate and validate the enumerated challenges in using GANs.

To properly allow GANs to explore beyond *the known* and *the seen* conveyed via the training data, we propose knowledge-adaptive GANs: incorporating GANs with information layers/structures and knowledge graphs. Our ongoing efforts are focused on conducting extensive empirical studies to validate the effectiveness of these methods.

References

[1] M. Z. Alom, T. M. Taha, C. Yakopcic, S. Westberg, M. Hasan, B. C. Van Esesn, A. A. S. Awwal, and V. K. Asari. The history began from alexnet: a comprehensive survey on deep learning approaches. *arXiv preprint arXiv:1803.01164*, 2018.

[2] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[3] S. Arora, R. Ge, Y. Liang, T. Ma, and Y. Zhang. Generalization and equilibrium in generative adversarial nets (gans). In *International Conference on Machine Learning*, pages 224–232, 2017.

[4] E. Y. Chang. Perceptual feature extraction (chapter 2). In *Foundations of large-scale multimedia information management and retrieval: Mathematics of perception*, chapter 2, pages 13–35. Springer, 2011.

[5] E. Y. Chang. Deepq: Advancing healthcare through artificial intelligence and virtual reality. In *ACM Multimedia Conference*, pages 1068–1068. ACM, 2017.

[6] C.-N. Chou, C.-K. Shie, F.-C. Chang, J. Chang, and E. Y. Chang. Representation learning on large and small data. *arXiv preprint arXiv:1707.09873*, 2017.

[7] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.

[9] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International Conference on Machine Learning*, pages 647–655, 2014.

[10] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *arXiv preprint arXiv:1803.01229*, 2018.

[11] I. Goodfellow. Adversarial machine learning (keynote). In *AAAI Conference on Artificial Intelligence*, 2019.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[13] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422, 2002.

[14] H. Heidenreich. What is a generative adversarial network? `http://hunterheidenreich.com/blog/what-is-a-gan/`.

[15] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708. IEEE, 2017.

[16] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5967–5976. IEEE, 2017.

[17] H.-C. Kao, K.-F. Tang, and E. Y. Chang. Context-aware symptom checking for disease diagnosis using hierarchical reinforcement learning. In *AAAI Conference on Artificial Intelligence*, 2018.

[18] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[20] J. P. Langlois, M. B. Mengel, W. L. Holleman, and S. A. Fields. Making a diagnosis, chapter 10. In *Fundamentals of Clinical Practice* 2^{nd} *edition*, page 198. Kluwer Academic Publishers, 2002.

[21] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. van der Maaten. Exploring the limits of weakly supervised pretraining. *arXiv preprint arXiv:1805.00932*, 2018.

[22] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[23] S. Nam, Y. Kim, and S. J. Kim. Text-adaptive generative adversarial networks: manipulating images with natural language. In *Advances in Neural Information Processing Systems*, pages 42–51, 2018.

[24] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier gans. In *International Conference on Machine Learning*, pages 2642–2651, 2017.

[25] C. Olah, A. Mordvintsev, and L. Schubert. Feature visualization. *Distill*, 2(11):e7, 2017.

[26] Y.-S. Peng, K.-F. Tang, H.-T. Lin, and E. Chang. REFUEL: Exploring sparse features in deep reinforcement learning for fast disease diagnosis. In *Advances in Neural Information Processing Systems*, pages 7333–7342, 2018.

[27] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[28] E. Ranschaert. Artificial intelligence in radiology: hype or hope? *Journal of the Belgian Society of Radiology*, 102(S1):20, 2018.

[29] H. Salehinejad, S. Valaee, T. Dowdell, E. Colak, and J. Barfett. Generalization of deep neural networks for chest pathology classification in x-rays using generative adversarial networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 990–994. IEEE, 2018.

[30] C.-K. Shie, H.-T. Chang, F.-C. Fan, C.-J. Chen, T.-Y. Fang, and P.-C. Wang. A hybrid feature-based segmentation and classification system for the computer aided self-diagnosis of otitis media. In *IEEE International Conference on Engineering in Medicine and Biology Society*, pages 4655–4658. IEEE, 2014.

[31] C.-K. Shie, C.-H. Chuang, C.-N. Chou, M.-H. Wu, and E. Y. Chang. Transfer representation learning for medical image analysis. In *IEEE International Conference on Engineering in Medicine and Biology Society*, pages 711–714. IEEE, 2015.

[32] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. Learning from simulated and unsupervised images through adversarial training. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2107–2116. IEEE, 2017.

[33] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.

[34] L. Sixt, B. Wild, and T. Landgraf. Rendergan: Generating realistic labeled data. *Frontiers in Robotics and AI*, 5:66, 2018.

[35] K. B. Wagholikar, V. Sundararajan, and A. W. Deshpande. Modeling paradigms for medical diagnostic decision support: a survey and future directions. *Journal of medical systems*, 36(5):3029–3049, 2012.

[36] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3462–3471. IEEE, 2017.

[37] X. Yi, E. Walia, and P. Babyn. Generative adversarial network in medical imaging: A review. *arXiv preprint arXiv:1809.07294*, 2018.

[38] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014.

[39] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015.

[40] A. R. Zamir, A. Sax, W. Shen, L. J. Guibas, J. Malik, and S. Savarese. Taskonomy: Disentangling task transfer learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3712–3722. IEEE, 2018.

[41] Z. Zheng, B. Wang, Y. Wang, S. Yang, Z. Dong, T. Yi, C. Choi, E. J. Chang, and E. Y. Chang. Aristo: An augmented reality platform for immersion and interactivity. In *ACM Multimedia Conference*, pages 690–698. ACM, 2017.

[42] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision*, pages 2242–2251. IEEE, 2017.