Joint Learning of Neural Networks via Iterative Reweighted Least Squares

Zaiwei Zhang
UT Austin

Xiangru Huang
UT Austin

Qixing Huang
UT Austin

Xiao Zhang
Google Inc.

Yuan Li
Google Inc.

Abstract

In this paper, we introduce the problem of jointly learning feed-forward neural networks across a set of relevant but diverse datasets. Compared to learning a separate network from each dataset in isolation, joint learning enables us to extract correlated information across multiple datasets to significantly improve the quality of learned networks. We formulate this problem as joint learning of multiple copies of the same network architecture and enforce the network weights to be shared across these networks. Instead of hand-encoding the shared network layers, we solve an optimization problem to automatically determine how layers should be shared between each pair of datasets. Experimental results show that our approach outperforms baselines without joint learning and those using pretraining-and-fine-tuning. We show the effectiveness of our approach on three tasks: image classification, learning auto-encoders, and image generation.

1. Introduction

Lack of training data remains one of the fundamental challenges in training effective deep neural networks for various visual recognition tasks. One potential solution is to perform transfer learning from other relevant datasets, which essentially amplifies the scale of the training data when training a particular network. A common strategy for transfer learning is to share network weights, i.e., either across the entire network or through a few hand-encoded layers. This strategy has proven to be effective in settings where the input datasets are similar, or where there is prior knowledge informing which layers of features to share. However, in cases where the input datasets exhibit significant variation, this strategy becomes sub-optimal, as it becomes unclear how to determine the shared network layers.

In this paper, we consider the problem of jointly learning neural networks from a collection of datasets that exhibit significant variations in content and appearance (See Figure 1). We show that despite such significant differences among the input datasets, they still present useful mutual information, which we can use to boost the performance of learning each individual network. We achieve this goal by introducing three novel ways for regularizing the network weights across all datasets. First, instead of hand-encoding the shared network weights, we formulate an optimization problem to assign them automatically. Moreover, rather than enforcing that shared network weights be identical, we use soft constraints to penalize the differences between pairs of shared network layers, allowing us to account for domain shifts between datasets. Finally, we jointly optimize the consistency among network weights using a robust norm, which allows us to extract layer-wise dataset clusters for weight-sharing.

Specifically, our approach takes a collection of datasets for the same task as input and outputs a learned network for each individual dataset. These networks share the same architecture but have different, yet correlated, layers. We integrate learning shared network layers and learning network

Figure 1: Results on black dress image generation with our approach. (Left) Results of an image generator trained from 20K dress images. (Right) Results of an image generator trained from 20K dress images and 200K face images.
weights through a unified optimization problem. The objective function combines a task-specific loss and a consistency term. The consistency term uses a robust norm to automatically determine how each layer should be shared. We also introduce a simple formulation to prioritize that the sharing scheme is consistent across adjacent layers. Our formulation admits effective optimization via iterative reweighted least squares (or IRLS).

We demonstrate the effectiveness of our approach across three diverse tasks: fine-grained image classification, learning auto-encoders, and image generation. These tasks range from predicting a single label (i.e., fine-grained image classification) to dense predictions (i.e., learning auto-encoders and image generation), and from supervised tasks (i.e., fine-grained image classification) to unsupervised tasks (i.e., learning auto-encoders and image generation). Across these tasks, we show that our approach is significantly better than learning each network in isolation, using an $L_2$-norm to share weights, as well as pretraining-and-finetuning. In particular, our approach extracts useful mutual information from datasets that seem to be visually uncorrelated, making our approach suitable across a broad range of settings.

In summary, we present the following contributions in this paper:

- We propose a robust norm to automatically determine how to share layers between each pair of datasets.
- We demonstrate the effectiveness of our approach on three tasks, including image classification, learning auto-encoders, and image generation.

## 2. Related Works

### Domain adaptation.

Our problem falls in the general category of domain adaptation. It is beyond the scope of this paper to provide a comprehensive review of the literature. We refer to [24, 9, 10, 22, 32] for recent advances and to [7] for a recent survey on this topic. For discriminative tasks (e.g., image classification and image segmentation), visual domain adaptation techniques fall into supervised domain adaptation techniques [6, 27] and unsupervised domain adaptation techniques [32, 16, 28]. Our approach falls into the supervised category. A common approach for supervised techniques is to share weights across networks. This strategy works for the case where the domain shifts are relatively small, but tends to break when the domain shifts are large. A potential solution is to share weights across a subset of predefined layers (e.g., [30]). However, this method requires prior knowledge of the mutual information across datasets, and we found that such information is not obvious, particularly between visually dissimilar datasets. In addition, it becomes extremely difficult to hand-encode weights across multiple networks. Instead, we solve a joint optimization problem to determine the matched layers. Matching network weights under the $L_2$ norm has been considered in a recent work [27]. The difference in this work is that we propose to use robust norms to automatically extract matched network weights between pairs of datasets, and we do so across multiple datasets in a consistent manner.

State-of-the-art unsupervised domain adaptation techniques build maps across the domains [16, 28, 35]. In particular, the inter-domain maps are enforced to be consistent. These maps are usually trained using generative adversarial networks [11] and training procedures described in follow-up works [29, 1]. Although we do not consider unsupervised domain adaptation in this paper, our approach can be potentially used for joint training of the discriminators used in each domain.

For synthesis tasks such as image generation, Liu et al. [19] proposed a method for training a pair of generative adversarial networks. Their strategy is similar to supervised domain adaptation for image classification/segmentation, i.e., by sharing hand-encoded layers. In contrast, our approach automatically learns matched layers, and we do so across multiple domains.

**Joint object matching.** Joint linking of corresponding layers across multiple networks is related to a recent line of work on joint optimization of object maps among image and shape collections [14, 13]. Similar to our setting, the central theme of these methods is to enforce the consistency of maps along cycles, so that a noisy map between two different objects can be computed by composing maps along a path of similar objects. In this paper, we apply this methodology to link layers of neural networks. Another related line of work [34, 16] focuses on utilizing the cycle-consistency constraint for regularization when training neural networks. Our problem differs from these works in that the neural networks optimized in our setting are associated with each dataset, and we establish consistent correspondences between network weights. In contrast, in these works the neural networks are defined between pairs of datasets.

## 3. Problem Setup

In this Section, we formally define the joint neural network learning problem we consider in this paper. Suppose we are given $n$ datasets for a specific task (e.g., image classification) and a neural network $G_{h}$ designed for it. Without losing generality, we assume $G_{h}$ is a feed-forward neural network that consists of $L$ layers $\theta^1, \ldots, \theta^L$, and each layer $l$ has $n_l$ parameters. However, our approach can be easily adapted for more sophisticated networks that connect layers using a graph. Our goal is to learn $n$ network parameters
We will adopt the BEGAN [2] architecture, where we share where we use the adversarial loss to optimize the generative models: The second task is training an auto-Encoder Loss. Classification Loss. In this setting, each dataset is given by a set of labeled images \( \mathcal{I}_i = \{(I_i, y_i)\} \), where \( y_i \) is the labeled associated with \( I_i \). The corresponding data dependent loss function is then given by

\[
    f_i(\theta_i) = \frac{1}{|\mathcal{I}_i|} \sum_{(I, y) \in \mathcal{I}_i} \log \left( \frac{1}{\mathcal{G}(\theta_i)(I)} \right),
\]

where we set \( \log \left( \frac{1}{\mathcal{G}(\theta_i)(I)} \right) \) as the cross entropy loss between predicted labels and ground-truth labels.

Auto-Encoder Loss. The second task is training an auto-encoder for a collection of images. We will evaluate auto-encoders indirectly, e.g., through their reconstruction loss on testing images and in the application of image completion. In this setting, each dataset \( \mathcal{I}_i = \{I\} \) is given by a collection of unlabeled images. Following [17], we directly use the regression loss to define the auto-encoder loss:

\[
    f_i(\theta_i) = \frac{1}{|\mathcal{I}_i|} \sum_{I \in \mathcal{I}_i} \|I - \mathcal{G}(\theta_i)(I)\|_2^2.
\]

Generative Adversarial Loss. The third task is training a generative adversarial network for a collection of real images. We will adopt the BEGAN [2] architecture, where we share the auto-encoders across different domains, and also adopt the DCGAN [26] architecture, where we share all the network parameters across different domains. Following [12], we use the adversarial loss to optimize the generative models:

\[
    f_i(\theta_i) = \frac{1}{|\mathcal{I}_i|} \sum_{I \in \mathcal{I}_i} \left( L(D_\theta(I)) - L(D_\theta(G_\theta(z_{z \in \mathcal{N}}))) \right),
\]

where \( L(\cdot) \) depends on the neural networks used. and are discussed in Section 5.

4. Approach

We proceed to present the proposed approach for joint learning of neural networks. In Section 4.1, we describe the proposed formulation. Then in Section 4.2, we show how to effectively solve the induced optimization problem.

\[
    f_{\text{data}} = \sum_{i=1}^{n} f_i(\theta_i).
\]

The regularization term \( f_{\text{cons}} \) forces the network parameters to match. In the presence of diverse datasets, the desired layer-wise network parameters \( \theta^l_i, 1 \leq i \leq n, 1 \leq l \leq L \) shall possess the following properties:

- For each layer \( l \), the input datasets form clusters so that for two datasets \( I_i \) and \( I_j \) that belong to the same cluster, \( \theta^l_i \) and \( \theta^l_j \) shall be close to each other. The motivation comes from the success of sharing bottom or top layers between a pair of networks for various domain adaption tasks (e.g., [19, 30]). In this paper, we generalize it to multiple networks. However, we do not assume the underlying clusters are given.

- The cluster structures are consistent between consecutive layers. This property is also motivated from the common practice of sharing a block of consecutive layers between a pair of networks (e.g., [19]). Again, we do not assume these blocks are given.

Our formulation of \( f_{\text{cons}} \) is motivated from [25], which perform data clustering using robust fusion penalties. Specifically, given a set of points \( p_1, \ldots, p_n \in \mathbb{R}^d \), these approaches solve the following problem to find perturbed points \( x_1, \ldots, x_n \in \mathbb{R}^d \) (c.f. [25]):

\[
    \min_{x_1, \ldots, x_n} \sum_{i=1}^{n} \|p_i - x_i\|^2 + \sum_{1 \leq i < j \leq n} \rho(||x_i - x_j||),
\]

where \( \rho(\cdot) \) is a robust norm. In [25], the authors have shown that the perturbed points of the same cluster tend to be identical.

We adapt this formulation to prioritize that the network parameters at each layer form clusters. To ensure that the cluster structures are consistent between adjacent layers, our key idea is to apply the formulation on concatenated layer-wise parameters \( (\theta^l_i, \theta^{l+1}_i) \). In other words, we would like to cluster \( (\theta^l_i, \theta^{l+1}_i) \) together, which implicitly forces the cluster structures to be consistent across adjacent layers:

\[
    f_{\text{cons}} = \sum_{1 \leq i < j \leq n} \sum_{l=1}^{L-1} \rho(||(\theta^l_i, \theta^{l+1}_i) - (\theta^l_j, \theta^{l+1}_j)||, \sigma^l).
\]

Here we choose a variant of the Huber loss to define the robust norm

\[
    \rho(x, \sigma) = \frac{x^2}{\sigma^2} \cdot \mathbb{1}_{|x| \leq \sigma} + \frac{\sigma^2}{2} \cdot \mathbb{1}_{|x| > \sigma}.
\]
we also find that due to the usage of the robust norm, the
\[ \sigma \]
where\( \sigma \) where
\[ \| x \|_2 \] where
\[ \| x \|_2 \]
not intuitive, such as Truck(6) and Fighter Aircraft(7), the
connecting two similar datasets, such as Dog(2) and Wolf(3) or
Figure 2: “Social network” formed by 8 datasets in two rep-
coder and decoder networks, we construct a graph by adding
between similar datasets, such as Dog(2) and Wolf(3) or
datasets share similar parameters. For middle layers of en-
\[ L \]
\[ L \]
\[ \lambda \]
\[ \lambda \] in Section 4.2.3.
Let \( \theta_i^{(0)}, 1 \leq i \leq n, 1 \leq l \leq L \) be the resulting network parameters from (6). We set the layer-wise hyper-parameter as
\[ \sigma^l = \text{mean}_{l \leq i \leq n, j \neq i} \min \| (\theta_i^{(0)}, \theta_i^{l+1(0)}) - (\theta_j^{(0)}, \theta_j^{l+1(0)}) \|, \]
\[ \text{which works well in all of our experiments.} \]
\[ \text{4.2.2 Reweighted Least Squares Regularization} \]
we proceed to apply IRLS to minimize the objective function
\[ \min \{ \theta_i \} \sum_{i=1}^{n} f_i(\theta_i) + \lambda \sum_{1 \leq i,j \leq n} \sum_{l=0}^{L-1} \| (\theta_i^l, \theta_j^{l+1}) - (\theta_j^l, \theta_j^{l+1}) \|^2 \]
\[ \text{(4)} \]
\[ \text{where } c_{ijl} > 0 \text{ are constants. As we will see later, (6) can}
\[ \text{also be used for solving intermediate steps of IRLS. To avoid}
\[ \text{breaking the flow of the paper, we defer the technical details}
\[ \text{for solving (6) to Section 4.2.3.} \]
\[ \text{Let } \theta_i^{(0)}, 1 \leq i \leq n, 1 \leq l \leq L \text{ be the resulting network parameters from (6). We set the layer-wise hyper-parameter as}
\[ \sigma^l = \text{mean}_{l \leq i \leq n, j \neq i} \min \| (\theta_i^{(0)}, \theta_i^{l+1(0)}) - (\theta_j^{(0)}, \theta_j^{l+1(0)}) \|, \]
\[ \text{which works well in all of our experiments.} \]
\[ \text{4.2.2 Reweighted Least Squares Regularization} \]
we proceed to apply IRLS to minimize the objective function
\[ \min \{ \theta_i \} \sum_{i=1}^{n} f_i(\theta_i) + \lambda \sum_{1 \leq i,j \leq n} \sum_{l=0}^{L-1} \| (\theta_i^l, \theta_j^{l+1}) - (\theta_j^l, \theta_j^{l+1}) \|^2 \]
\[ \text{(4)} \]
\[ \text{where } c_{ijl} > 0 \text{ are constants. As we will see later, (6) can}
\[ \text{also be used for solving intermediate steps of IRLS. To avoid}
\[ \text{breaking the flow of the paper, we defer the technical details}
\[ \text{for solving (6) to Section 4.2.3.} \]
\[ \text{Let } \theta_i^{(0)}, 1 \leq i \leq n, 1 \leq l \leq L \text{ be the resulting network parameters from (6). We set the layer-wise hyper-parameter as}
\[ \sigma^l = \text{mean}_{l \leq i \leq n, j \neq i} \min \| (\theta_i^{(0)}, \theta_i^{l+1(0)}) - (\theta_j^{(0)}, \theta_j^{l+1(0)}) \|, \]
\[ \text{which works well in all of our experiments.} \]

\[ \text{4.2.2 Reweighted Least Squares Regularization} \]
we proceed to apply IRLS to minimize the objective function
\[ \min \{ \theta_i \} \sum_{i=1}^{n} f_i(\theta_i) + \lambda \sum_{1 \leq i,j \leq n} \sum_{l=0}^{L-1} \| (\theta_i^l, \theta_j^{l+1}) - (\theta_j^l, \theta_j^{l+1}) \|^2 \]
\[ \text{(4)} \]
\[ \text{where } c_{ijl} > 0 \text{ are constants. As we will see later, (6) can}
\[ \text{also be used for solving intermediate steps of IRLS. To avoid}
\[ \text{breaking the flow of the paper, we defer the technical details}
\[ \text{for solving (6) to Section 4.2.3.} \]
\[ \text{Let } \theta_i^{(0)}, 1 \leq i \leq n, 1 \leq l \leq L \text{ be the resulting network parameters from (6). We set the layer-wise hyper-parameter as}
\[ \sigma^l = \text{mean}_{l \leq i \leq n, j \neq i} \min \| (\theta_i^{(0)}, \theta_i^{l+1(0)}) - (\theta_j^{(0)}, \theta_j^{l+1(0)}) \|, \]
\[ \text{which works well in all of our experiments.} \]

\[ \text{4.2.2 Reweighted Least Squares Regularization} \]
we proceed to apply IRLS to minimize the objective function
\[ \min \{ \theta_i \} \sum_{i=1}^{n} f_i(\theta_i) + \lambda \sum_{1 \leq i,j \leq n} \sum_{l=0}^{L-1} \| (\theta_i^l, \theta_j^{l+1}) - (\theta_j^l, \theta_j^{l+1}) \|^2 \]
\[ \text{(4)} \]
\[ \text{where } c_{ijl} > 0 \text{ are constants. As we will see later, (6) can}
\[ \text{also be used for solving intermediate steps of IRLS. To avoid}
\[ \text{breaking the flow of the paper, we defer the technical details}
\[ \text{for solving (6) to Section 4.2.3.} \]
\[ \text{Let } \theta_i^{(0)}, 1 \leq i \leq n, 1 \leq l \leq L \text{ be the resulting network parameters from (6). We set the layer-wise hyper-parameter as}
\[ \sigma^l = \text{mean}_{l \leq i \leq n, j \neq i} \min \| (\theta_i^{(0)}, \theta_i^{l+1(0)}) - (\theta_j^{(0)}, \theta_j^{l+1(0)}) \|, \]
\[ \text{which works well in all of our experiments.} \]
We can understand the behavior of IRLS as follows. For layers with similar parameters, the corresponding weights are close to 1, while for layers with dissimilar parameters, the corresponding weights are close to 0. With such weights, solving (8) would push the layers with similar weights to be even closer to each other. In the mean time, layers with dissimilar weights are likely to be pulled away from each other due to the data terms. In the end, layers tend to form clusters.

Across all of our experiments, we found that IRLS converges favorably fast. In our implementation, we monitor

\[ \delta^{(k)} := \max_{1 \leq i < j \leq n, 1 \leq l \leq L} |w_{ijl}^{(k)} - w_{ijl}^{(k-1)}|, \]

and terminate the IRLS procedure when \( \delta^{(k)} \leq 10^{-2} \). In our experiments, we found 4-8 iterations are sufficient for convergence.

Figure 2 illustrates the links that connect layers with similar optimized parameters over eight datasets for the task of learning auto-encoders. We can see that the links reveal meaningful shared information across the datasets.

4.2.3 Joint Network Optimization

In this section, we describe the technical details for solving (6), which has been used in the variable initialization stage as well as the reweighted least squares stage. When \( n \) is large, it is hard to solve (6) directly, since it involves one copy of the network for each dataset. Motivated from the success of block-coordinate descent techniques for solving large-scale optimization problems (c.f. [4]), we propose to solve (6) by optimizing one network at a time while fixing the other networks. Specifically, suppose we pick the \( i \)-th network at the current iteration, and with \( \theta^i_j, j \neq i, 1 \leq l \leq L \) we denote the current parameters of other networks. Then it is clear that (6) reduces to

\[ \min_{\theta} f_i(\theta) + \lambda \sum_{l=1}^{L} \sum_{j \neq i} c_{ijl} \| \theta^i_l - \sum_{j \neq i} c_{ijl} \theta^j_l / \sum_{j \neq i} c_{ijl} \|^2 \]  (9)

Since \( \sum_{j \neq i} c_{ijl} \theta^j_l / \sum_{j \neq i} c_{ijl} \) is a constant vector when optimizing \( \theta^i_l \), (9) can be considered as a standard network training problem, and we apply stochastic coordinate descent for optimization. At each iteration, we train (9) with one epoch before moving to the next iteration.

5. Experimental Evaluation

In this section, we present the experimental evaluations of the proposed approach. We first describe the experimental setup. We then evaluate the benefits of the proposed approach on each specific task. Please refer to the supplemental material for additional results.

5.1. Experimental Setup

Datasets. We summarize the datasets we used for each specific task below. Please refer to the supplemental material for a detailed specification:

- Fine-grained image classification. For this task, we consider five fine-grained classification datasets: Cars Dataset [18], FGVC-Aircraft Benchmark [21], Stanford Dogs Dataset [15], Caltech-UCSD Birds-200-2011 [33], 102 Category Flower Dataset [23] for evaluating our joint learning approach. Note that the appearance of these images are considered visually dissimilar. We pick fine-grained classification tasks to
demonstrate that across different domains, our approach can still extract meaningful mutual information, leading to performance gains. For this task, we employ InceptionV3 [31] network architecture, with pretrained weights from ImageNet for initialization. In order to mitigate over-fitting, we carry out the Inception-style data augmentation methods: scale and aspect ratio variation, and image distortion.

- **Auto-encoder learning.** The second task considers image completion from multiple domains of images with auto-encoders. For this task, we employ a 9-layer auto-encoder architecture (details are deferred to the supplemental material). We picked images from 8 classes (image samples are shown in Figure 2), which are Bird, Dog, Wolf, Flower, Car, Truck, Plane, and Fighter Aircrafts, using 500 images for training and 100 for testing. All images are retrieved from ImageNet, and sampled from the five fine-grained classification datasets. For each image, we randomly cropped out a 22x22 square for the image completion task.

- **Image generation.** In the third task we consider image generation using a collection of images from different domains. For this task, we employ the experimental setting (e.g., network architecture) from both [26] and [2]. We have conducted three experiments which involve six datasets. The first dataset is comprised of video snapshots taken from [3], and contains 3K images of humans walking down a street, while the second dataset contains 10K synthetic images of humans walking, which we generated. The third dataset is CelebA [20], which contains around 200K images, and the fourth dataset is a collection of 20K black dress images crawled from the web. We additionally use the Flower dataset [23] and the Bird dataset [33].

Baseline Approaches. We consider the following four baselines to assess the proposed approach.

- **Baseline I: Isolated-Training.** The first baseline simply trains the network from each dataset independently. To make a fair comparison between the proposed approach and state-of-the-art approaches, we report the performance of state-of-the-art methods that use AlexNet with weights pre-trained on ImageNet.

- **Baseline II: L2-regularization** The second baseline replaces the robust-norm by the L2-norm for regularizing differences in layer-wise parameters. This baseline assesses the benefit of using the robust norm for regularization.

- **Baseline III: Shareall.** The third baseline simply shares the same network parameters across all the datasets. This baseline is introduced to assess the importance of sharing weights in an adaptive manner.

Baseline IV: Pretrain-Finetune. The fourth baseline is generalized from the popular pretraining-fine-tuning paradigm. In this case, we first train a joint model from all the datasets by sharing weights. We then fine-tune the joint model on each dataset independently. This baseline is introduced to assess the advantage of solving an optimization problem for joint training.

5.2. Task I: Image Classification.

Table 4 illustrates the testing accuracy of our approach and baseline approaches. Our approach outperforms most baseline approaches across all the input datasets. The improvement over Isolated-Training is significant in most cases, especially on Birds and Dogs dataset. Figure 3 shows the performance gain during training for those two datasets, and only when consistency loss drops, there are significant boosts in testing accuracy for Joint method. This demonstrates the importance of sharing weights. L2-Regularization improves over Isolated-Training in some cases. However, falsely linking datasets together causes unstable behaviors, such as very low testing accuracy on Aircrafts. Using the robust norm not only lifts the testing accuracy but also insures at least similar performance compared to Isolated-Training. This shows the advantage of using a robust norm for regularization in the presence of diverse datasets. A surprising result is that Shareall leads to the lowest accuracy in some cases. This can be understood from the fact that the five datasets are diverse, and it is important to allow some difference across the networks (e.g., using L2-Regularization). Finally, Pretrain-Finetune improves over Shareall but only matches the performance of Isolated-Training due to the quality of the
5.3. Task II: Learning Auto-Encoders

Figure 5 illustrates the average reconstruction loss of our approach and baseline approaches on the testing data (which is left out during training). Compared to Isolated-Training and Pretrain-Finetune, our joint learning method leads to the smallest reconstruction error on the testing data. This again shows the importance of sharing weights in a soft manner and using a robust norm to filter out irrelevant layers. In particular, compared to Shareall, the performance gain is significant. Moreover, Pretrain-Finetune results in overfitting. The improvements over the other two baselines are also noticeable. For example, the generalization behavior of Isolated remains poor. Figure 2 illustrates links that connect layers with similar weights. We can see that our approach successfully recover the intrinsic similarities across the datasets, e.g., between the two animal datasets and between the two airplane datasets.

5.4. Task III: Image Generation

We conducted the following three experiments on the image generation task:

- Use DCGAN [26] to generate images of humans walking down a street by joint learning with 3K real and 10K synthetic images.
- Use BEGAN [2] to generate images of black dresses by joint learning with 20K dress and 200K face images.
• Use BEGAN [2] to generate images of flowers by joint learning with 2K flower, 6K bird, and 200K face images.

Figure 6 shows the qualitative results of the image generation task. We can see that our approach leads to the best general results across the input datasets. As for baseline approaches, Isolated training easily overfits the training data. As the datasets are quite diverse, Shareall, which uses one network to generate images, tends to generate images that interpolate across different categories. The resulting images are thus unrealistic. Starting from Shareall and then finetuning on each dataset certainly improves the visual quality. However, the results are still not as competent as our approach. L2-regularization, which uses a soft weight sharing scheme, avoids the issue of generating mixed-class images experienced by Shareall. However, the issue of L2-regularization is that it evenly distributes the error, resulting in overly smooth images.

6. Conclusions and Future Work

In this paper, we have introduced a method for joint learning of neural networks among a collection of relevant and diverse datasets for the task of transfer learning. The key idea behind our approach is to use a robust norm to automatically identify which layers should be matched between each pair of datasets. Our formulation also enforces consistency of matches between adjacent layers. The formulation can be easily optimized via iterative reweighted least squares. Experimental results show the advantage of our approach against four baseline approaches, namely, (1) training from each dataset in isolation, (2) sharing the same weights across all the networks, (3) finetuning from shareall, and (4) using the L2-regularization. The improvements are consistent among the three tasks introduced in this paper, namely, fine-grained image classification, learning auto-encoders, and image generation.

There are ample opportunities for future research. We would like to apply our method for other tasks such as depth prediction and semantic segmentation. Moreover, we would like to apply our approach to other types of neural networks such as recurrent neural networks and densely connected networks. Moreover, we have used the same network architecture across all datasets. It would be interesting to extend the approach to jointly learn different but correlated networks, e.g., networks for image segmentation and networks for image classification may share similar convolutional layers. Finally, we would like to extend our approach for unsupervised domain adaptation across multiple datasets.

References


