Migrating Knowledge between Physical Scenarios based on Artificial Neural Networks

Yurui Qu,[1,2,*] Li Jing,[1,†] Yichen Shen,[1] Min Qiu,[2,3] and Marin Soljačić[1]

[1] *Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*
[2] *State Key Laboratory of Modern Optical Instrumentation, College of Optical Science and Engineering, Zhejiang University, Hangzhou 310027, China*
[3] *Institute of Advanced Technology, Westlake Institute for Advanced Study, Westlake University, Hangzhou 310024, China*

(Dated: September 5, 2018)

Deep learning is known to be data-hungry, which hinders its application in many areas of science when datasets are small. Here, we propose to use transfer learning methods to migrate knowledge between different physical scenarios and significantly improve the prediction accuracy of artificial neural networks trained on a small dataset. This method can help reduce the demand for expensive data by making use of additional inexpensive data. First, we demonstrate that in predicting the transmission from multilayer photonic film, the relative error rate is reduced by 46.8% (26.5%) when the source data comes from 10-layer (8-layer) films and the target data comes from 8-layer (10-layer) films. Second, we show that the relative error rate is decreased by 22% when knowledge is transferred between two very different physical scenarios: transmission from multilayer films and scattering from multilayer nanoparticles. Finally, we propose a multi-task learning method to improve the performance of different physical scenarios simultaneously in which each task only has a small dataset.

Deep learning is a powerful machine learning algorithm that discovers representations of data with multiple levels of abstraction based on multiple processing layers [1]. Recently, deep learning has received an explosion of interest because it continuously pushes the limit of traditional image recognition, machine translation, decision-making as well as many other applications [2–5]. Meanwhile, deep learning is also penetrating into other disciplines such as drug design [6, 7], genetics [8, 9], material science [10] and physics, including classification of complex phases of matter [11, 12], electromagnetic inverse problem [13, 14], nanostructure design [15, 16], high-energy physics [17] and quantum physics [18–20]. One drawback is that deep learning is a data-hungry method and can only work well if fed with massive data. However, collecting a large amount of data is slow and expensive for many numerical simulations, such as bands of three-dimensional (3D) photonic crystals [21], and even much more difficult for the experiment, since it might require fabricating tens of thousands of samples [22] or doing tens of thousands of measurements [23]. Similar situations are also common in other scientific areas in which collecting a large amount of simulated or experimental data is difficult. Direct learning from a small dataset results in under-represented features, which leads to poor performance. There has yet to emerge a solution to improve deep learning performance for scientific problems with a small dataset.

Transfer learning has attracted growing interest in recent years because it can significantly improve the performance in the target task through the transfer of knowledge from the source task that has already been learned [24–27]. Transfer learning is a useful method when the source dataset is large and inexpensive and target dataset is small and expensive. Jason Yosinski et al. demonstrated on ImageNet that transferred layers can improve classification accuracy by 2% on a new task after substantial fine-tuning [28]. Andrei A. Rusu et al. showed that learning from a simulation and transferring the knowledge to a real-world robot can solve the problem that training models on a real robot is too slow and expensive [29]. However, unlike classic transfer learning that only cares about doing well on one particular target task or domain, another method called multi-task learning can do well on all related tasks which are trained simultaneously and benefit from each other. Multi-task learning has been used successfully across many applications such as natural language processing [30], speech recognition [31, 32] and computer vision [33].

In this Letter, we propose a deep neural network architecture with transfer learning ability that can significantly improve the performance of physical problems even if their datasets are small. Although we focus here on a particular photonic problem of transmission from multilayer film and scattering from nanoparticle, the approach presented can easily be generalized to many other scientific problems. The deep neural network with transfer learning is investigated in several cases: (1) the source and target data come from similar physical problems, for example, transmission from multilayer films with different number of geometric layers. The validation error continuously decreases as more layers of the neural network are transferred. Note that the former "layers" is used for photonic structures and the latter "layers" is used for weights and biases in neural networks. The relative error reduction is 46.8% (26.5%) when the source data comes from 10-layer (8-layer) film and the target data comes from 8-layer (10-layer) film; (2) The source and target data come from very different physical problems. The source data are scattering cross-sections from

* yuruiqu@mit.edu
† ljing@mit.edu

8-layer core-shell nanoparticles and the target data are transmissivities from 8-layer films. To our surprise, the relative error rate still deceases by 22% after transferring knowledge from the nanoparticle scattering problem to the multilayer film transmission problem; (3) Multiple tasks are 8-layer, 10-layer, 12-layer, 14-layer films. The performance of multi-task learning outperforms the direct learning trained only on a specific group of data. The neural network with transfer learning can significantly improve the performance of neural networks with only a small dataset, which could benefit the application of deep learning in many physical problems and other fields.

In classic deep learning, a model can be well trained for some task and domain if sufficient labeled data are provided. We will later define in more detail what exactly a task and a domain are. At this moment, let us assume that a task is the objective that our model aims to perform, e.g. predict the transmission spectra of 10-layer films, as shown in Fig. 1(a). We can now train a model on such a dataset and expect it to perform well on unseen data of a 10-layer film. However, this classic deep learning breaks down when we do not have sufficient labeled data for this task or domain. Training on a small dataset will cause collapse in performance because of severe overfitting problem. Transfer learning allows us to deal with this problem by leveraging the existing labeled data from some related task or domain, e.g. transmissivities from 8-layer films, or even a very different task like scattering cross-sections from core-shell nanoparticles. We try to store the knowledge gained in solving the source task in the source model and apply it to the target model to help the target task (see Fig. 1(b)).

Here, we introduce some notations to give the definition of transfer learning. First, we give the definitions of a domain and a task, respectively. A domain D has two components: a feature space χ and a marginal probability distribution $P(X)$, where $X = \{x_1, . . . , x_n\} \in \chi$. For example, if our learning task is to predict the transmission spectrum of a 10-layer film, x_i is the i^{th} term one-dimensional vector representing the thickness of each film. χ is the space of all term vectors, and X is a particular learning sample. Given a specific domain, $D = \{X, P(X)\}$, a task has two components: a label space Y and an objective predictive function $f(\cdot)$ (denoted by $T = \{Y, f(\cdot)\}$), which can be learned from the training data. Training data consist of pairs $\{x_i, y_i\}$, where $x_i \in X$ and $y_i \in Y$. The function $f(\cdot)$ can be used to predict the corresponding label, $f(x)$, of a new instance x. $f(x)$ can be written as $P(y|x)$ from a probabilistic viewpoint. In our example, Y is the set of all labels, which are transmissivities in the visible spectrum from 400 nm to 800 nm.

We denote the source domain data as $D_s = \{(x_{s1}, y_{s1}), ..., (x_{sn}, y_{sn})\}$, where $x_{si} \in X_s$ is the data instance and $y_{si} \in Y_s$ is the corresponding class label. Similarly, we denote the target domain data as $D_t = \{(x_{t1}, y_{t1}), ..., (x_{tn}, y_{tn})\}$, where the input $x_{ti} \in X_t$

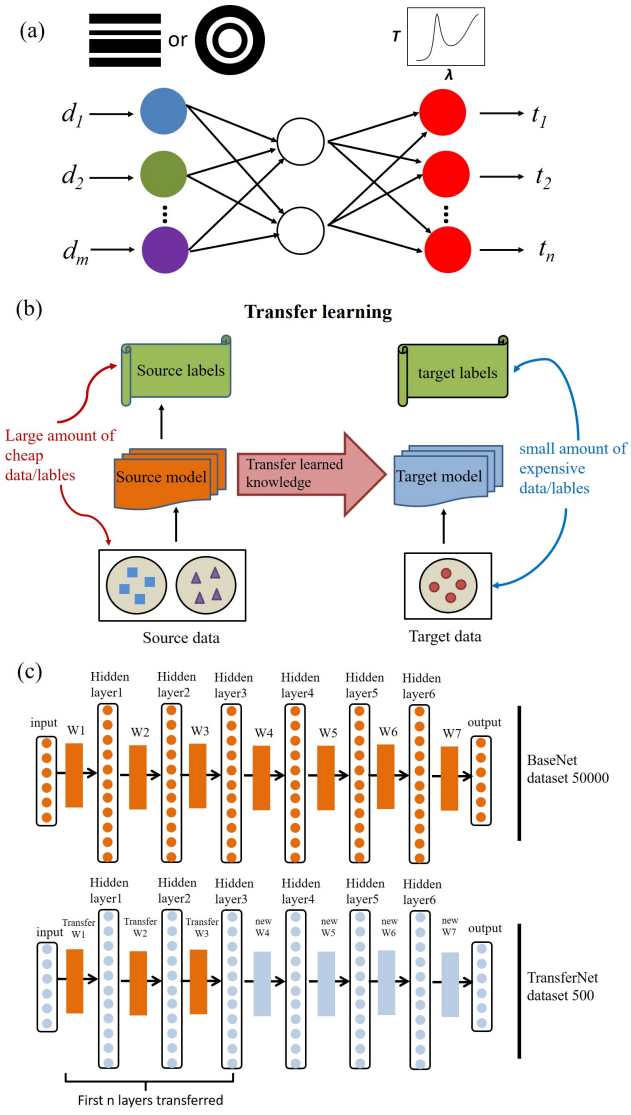

FIG. 1. (a) Illustration of the neural network architecture. The input is the thickness of each film or each shell of the nanoparticle, and the output is the transmission or scattering spectrum. Only one hidden layer is drawn here for convenience. Our actual neural network has six hidden layers. (b) Learning process of transfer learning. Transfer learning techniques can transfer the learned knowledge from the source model to the target model to improve the performance of the target task. Generally, the source domain has a large amount of inexpensive data while the target domain only has a small amount of expensive data. (c) Neural network structure of transfer learning. Top row: The base network (BaseNet) learns from scratch on a large source dataset of 50000 examples. Bottom row: The transfer network copies the first n layers from the BaseNet as the initialization weights and biases, and then the entire network is trained (fine-tuned) on the small target dataset of 500 examples.

and $y_{ti} \in Y_t$ is the corresponding output. The definition of transfer learning is as follows: given a source domain D_s and learning task T_s, a target domain D_t and learning task T_t , transfer learning can help learn the target predictive function $f_t(\cdot)$ better in D_t using the knowledge from D_s and T_s, where $D_s \neq D_t$, or $T_s \neq T_t$.

Artificial neural network structure of our transfer learning method is shown in Fig. 1(c). The input data are the thicknesses of each film or shell (the materials were fixed), and the output data are the transmissivities sampled at points between 400 nm to 800 nm. The thicknesses are between 30 nm to 70 nm, and materials are SiO_2 and TiO_2 for alternating layers of multilayer films and of core-shell nanoparticles. We train a fully connected neural network called BaseNet, with 6 hidden layers and 200 neurons per layer, on the source domain with a dataset of 50000 examples. 80% of the data are used for training the network and the other 20% are the validation data and the test data (10% each), which are the same for the target task. BaseNet learns from scratch, which can also be called direct learning. On the other hand, TransferNet has the same network structure as BaseNet, while TransferNet copies the first n layers from the BaseNet as the initialization of weights and biases of the first n layers. The rest higher layers of TransferNet are initialized randomly, and the entire TransferNet are fine-tuned simultaneously. TransferNet is trained on target domain with a small dataset of 500 examples. Next, we are going to compare transfer learning to direct learning on the same dataset to demonstrate that transfer learning can truly improve the performance.

The most general method of calculating the transmittance of a multilayer film is based on a matrix formula [34] of the boundary conditions at the film surfaces derived from Maxwells equations (see Method in Supporting Information). We use a neural network with direct learning and also with transfer learning, respectively, to approximate this transfer matrix formula, and we compare the performance in these two cases. We first explore the transfer learning between 8-layer films and 10-layer films, and try to transfer knowledge in both directions, as shown in Fig. 2(a) and (c). The spectrum error in this paper is defined as the average difference between the prediction and the exact result per spectrum point:

$$Error = \frac{1}{n}\sum_{i=1}^{n}\frac{|T_{prediction}(\lambda_i) - T_{exact}(\lambda_i)|}{T_{exact}(\lambda_i)} \qquad (1)$$

where n is the number of the spectrum points. In our case, we sampled between 400 nm to 800 nm with 2 nm step, so n=200 (not including 800 nm).

The validation spectrum errors over epochs have sharp decline occasionally, suggesting that the neural network is learning something about the data at each of those points (see Fig. S1). For target task of 8-layer film, direct learning with randomly initialized weighs and biases on 400 examples has a validation spectrum error of 7.7% (Fig. 2(b)). After transferring the first layer from trained BaseNet and retraining all the layers in TransferNet together, the spectrum error is reduced to 6.9%. With more layers of the BaseNet transferred, the spectra errors continuously decrease. When 6 layers of BaseNet are transferred, the spectrum error decreases to 4.1%, which is 46.8% relative reduction compared to direct learning. Transfer learning also works well when knowledge is transferred from 8-layer film to 10-layer film

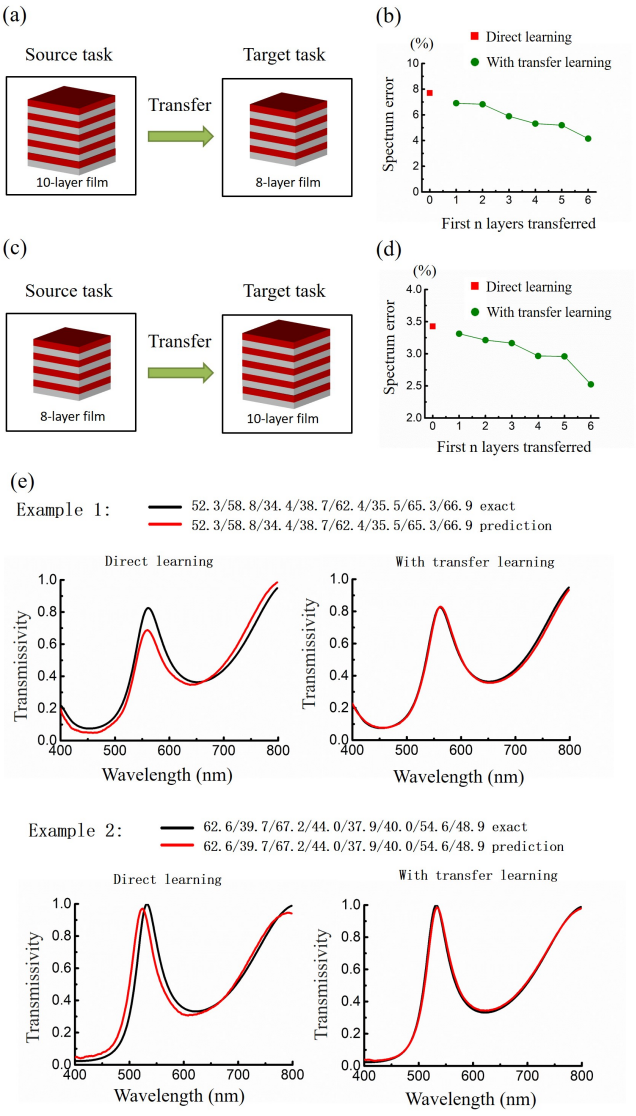

FIG. 2. (a) Illustration of transfer learning process and (b) spectrum error for the first n layers of BaseNet transferred when the source domain are 10-layer films and the target domain are 8-layer films. (c) and (d) are the case that the source domain are 8-layer films and the target domain are 10-layer films. (e) Two examples of transmission spectra for the case that the source domain are 8-layer films and the target domain are 10-layer films. Exact spectra are black lines and predicted spectra are red lines. Comparison of the direct learning to the transfer learning demonstrates that transfer learning can predict more accurate spectra than direct learning.

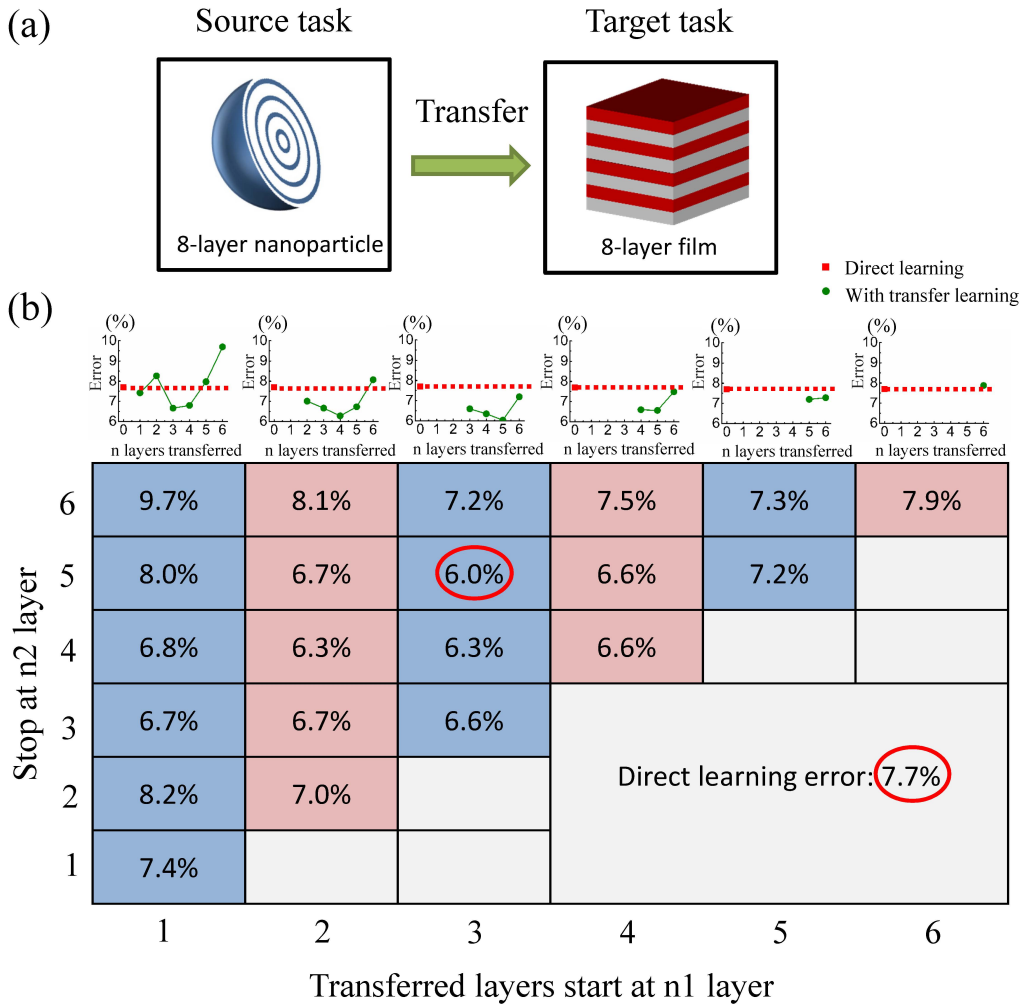

FIG. 3. (a) Illustration of the transfer learning process when the source domain are 8-layer nanoparticles and the target domain are 8-layer films. (b) Validation error when transferred layers begin at n1 layer and stop at n2 layer of trained BaseNet. The insets are spectra errors for each column.

(Fig. 2(c)). Direct learning for 10-layer film has the spectrum error of 3.4%. With 6 layers of BaseNet transferred, the spectrum errors is reduced to 2.5% , which is 26.5% relative error reduction (see Fig. 2(d)).

Two examples are presented in Fig. 2(e) to demonstrate that transfer learning can give a better prediction of the transmission spectrum compared to direct learning. Two examples come from the case that the source domain are 10-layer films and the target domain are 8-layer films. Black lines and red lines are theoretical and predicted transmission spectra, respectively. For the first example, the predicted spectrum using direct learning has lower peak transmissivity than the exact spectrum, and differences at other wavelengths are also obvious. For the second example, the entire predicted transmission spectrum based on direct learning shifts to shorter wavelength compared to the exact spectrum. However, the predicted spectra using transfer learning for both cases are much more accurate. This result is surprising because the spectra are predicted by the neural network which has only seen 400 training examples.

Next, we try to transfer knowledge between two very different tasks, the scattering from multilayer nanoparticles and the transmission from multilayer film, as shown in Fig. 3(a). Scattering cross-section from multilayer nanoparticle can be calculated using transfer matrix method, but in the forms of Bessel functions [35] (also see Method in Supporting Information). In Fig. 3(b), the first column of the table represents the spectrum error with the first n layers of the BaseNet transferred. The error of transfer learning is a little lower than direct learning (red dashed line) when only the first layer of the BaseNet is transferred before training the TransferNet. After transferring the first two layers of the BaseNet, the

error of transfer learning increases instead and surpasses the red dashed line, which is negative transfer that is harmful for learning the target task. Transferring the first 3 or 4 layers of the BaseNet can help to reduce the spectrum error by around 1% compared to direct learning. However, if the first 5 or 6 layers of the BaseNet are transferred together, the final performance will deteriorate sharply. From the results we can tell that some layers transferred from the BaseNet are specific to the nanoparticle scattering problem. These layers of the BaseNet will not help the target task learning at all, and can even be harmful for the final performance. Other layers of the BaseNet which are apparently general to both problems can be transferred and improve the target performance.

To find the layers of the BaseNet that contain the transferable physical knowledge, we study the performance of transferring layers starting from $n1$ ending at $n2$, as shown in Fig. 3(b). The validation error decreases to 6% when 3^{rd}, 4^{th} and 5^{th} layers of the BaseNet are transferred, with around 22% relative error reduction compared to direct learning. The results are quite surprising and enlightening. Even for the case when there is not enough available data from a similar task, we can still utilize data from a different task, which largely expands the areas where this transfer learning method can be applied. Through this process, we are able to isolate the shared physical knowledge learned by the neural networks to some extent while keeping out the scenario specific knowledge.

The transfer learning method described above requires a large amount of inexpensive data from related tasks. However, in some cases, there are several related tasks, but each of them has only a small amount of data. Classic transfer leaning method cannot work well in this case. Here, we introduce another knowledge transferring method through multi-task learning. As shown in Fig. 4(a), multi-task learning shares some hidden layers among all physical scenarios, while keeping several task-specific output layers. Each task can benefit from the knowledge learned in the other tasks, and the performance of each task can be improved compared to individual direct learning. This makes sense intuitively: the more tasks we are learning simultaneously, the more our model has to find a representation that captures all of the tasks and the less is our chance of overfitting the our original task. The neural network structure of multitask learning is shown in Fig. 4(b). The network shares the first n hidden layers and splits for the rest task-specific layers. Four target data are from 8-layer, 10-layer, 12-layer, 14-layer films, and each has a small dataset of 500 examples. The key to the successful learning is to train the model for all the tasks simultaneously. All the data are fed in each update of the model. The training algorithm needs to be adjusted slightly from the conventional backpropagation algorithm because of the split task-specific layers. When a training example is from 8-layer film, only the shared layers and the specific layers belonging to the 8-layer films task are updated. Other

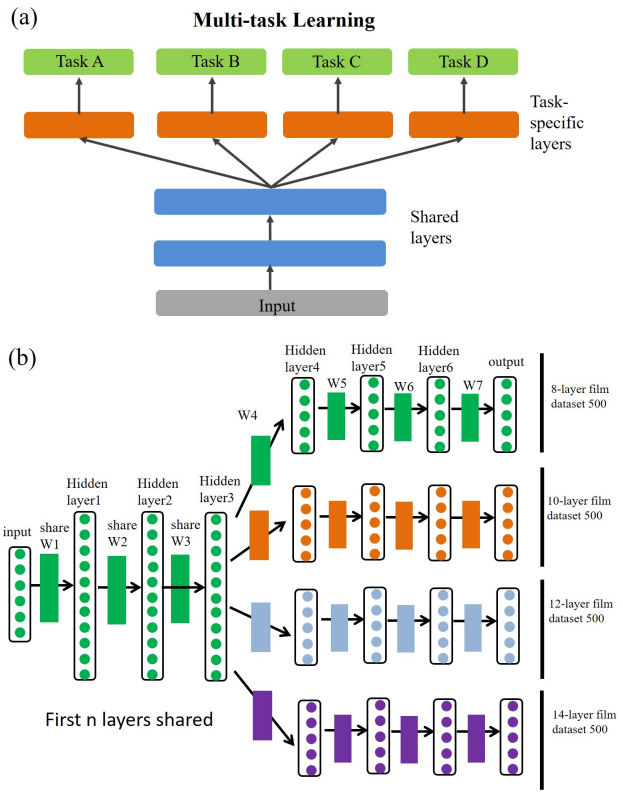

FIG. 4. (a) Learning process of multi-task learning. Several target tasks share part of hidden layers together and are learned simultaneously. Each target domain has a small dataset. (b) Neural network structure of multitask learning. The network shares the first n hidden layers and splits for the rest. Four target domains are 8-layer, 10-layer, 12-layer, 14-layer films, and each has a small dataset of 500 examples.

task-specific layers are kept intact. The same operation is done to train all four tasks.

We compare the spectrum errors of direct learning to that of multi-task learning in Table I. Even if we only use four tasks learned together, each with a small dataset of 500 examples, multi-task learning has lower spectrum error than direct learning in each of four tasks. The relative reduction of the spectrum error is 10.4%, 17.6%, 22.9% and 15.3% for 8-layer, 10-layer, 12-layer, 14-layer films, respectively. We expect that the performance can be better with more target tasks. The best neural network structure is different for each target task, as shown in Fig. S2. For 8-layer, 10-layer and 12-layer film, the best performance can be achieved when the first 2 hidden layers are shared. For 14-layer film, however, the best case is for the first 3 hidden layers shared. We can also see that the performance deteriorates sharply if too many layers are shared. The reason is that the last several layers are specific for each task and cannot transfer knowledge among different tasks.

In conclusion, we present two transfer learning methods to help with the fact that deep learning methods cannot work well with small datasets. We demonstrate

TABLE I. Comparison of direct learning error to multi-task learning error

Training dataset	8 layers	10 layers	12 layers	14 layers
Direct learning error	7.7%	3.4%	7.0%	11.1%
Multi-task learning error	6.9%	2.8%	5.4%	9.4%
Relative error reduction	10.4%	17.6%	22.9%	15.3%

that the neural network with transfer learning can give more accurate prediction compared to direct learning when trained on the same dataset. The knowledge in the neural network can be transferred not only between similar physical scenarios, such as transmission from multilayer films with different number of flims, but also between very different physical scenarios like scattering from core-shell nanoparticles and transmission from multilayer films. Multi-task learning, on the other hand, can improve the performance of several related tasks simultaneously even if each task only has a small dataset of 500 examples. The challenge of this transfer learning method is how to avoid negative transfer between two different tasks. Here we systematically select the general layers and specific layers in the neural network, and it would be important to investigate if this process can be done automatically in the future. Looking forward, neural networks with transfer learning could not only benefit the development of deep learning in many physical problems of which datasets are expensive and small, but also in other areas of science such as biology, chemistry and material science.

ACKNOWLEDGMENTS

We acknowledge discussions with John Peurifoy. Research was sponsored in part by the Army Research Office and was accomplished under Cooperative Agreement Number W911NF-18-2-0048. This material is based upon work supported in part by the National Science Foundation under Grant No. CCF-1640012. This material is based upon work supported in part by the Semiconductor Research Corporation under Grant No. 2016-EP-2693-B. Yurui Qu was supported by Chinese Scholarship Council (CSC No. 201706320254).

[1] Y. LeCun, Y. Bengio, and G. Hinton, Nature **521**, 436 (2015).
[2] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, *et al.*, IEEE Signal processing magazine **29**, 82 (2012).
[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, in *Advances in neural information processing systems* (2012) pp. 1097–1105.
[4] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *et al.*, Nature **518**, 529 (2015).
[5] 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.*, Nature **529**, 484 (2016).
[6] J. Ma, R. P. Sheridan, A. Liaw, G. E. Dahl, and V. Svetnik, Journal of chemical information and modeling **55**, 263 (2015).
[7] E. Gawehn, J. A. Hiss, and G. Schneider, Molecular informatics **35**, 3 (2016).
[8] M. K. Leung, H. Y. Xiong, L. J. Lee, and B. J. Frey, Bioinformatics **30**, i121 (2014).
[9] H. Y. Xiong, B. Alipanahi, L. J. Lee, H. Bretschneider, D. Merico, R. K. Yuen, Y. Hua, S. Gueroussov, H. S. Najafabadi, T. R. Hughes, *et al.*, Science **347**, 1254806 (2015).
[10] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, and C. Kim, npj Computational Materials **3**, 54 (2017).
[11] S. S. Schoenholz, E. D. Cubuk, D. M. Sussman, E. Kaxiras, and A. J. Liu, Nature Physics **12**, 469 (2016).
[12] L. Wang, Physical Review B **94**, 195105 (2016).
[13] J. Peurifoy, Y. Shen, L. Jing, Y. Yang, F. Cano-Renteria, B. G. DeLacy, J. D. Joannopoulos, M. Tegmark, and M. Soljačić, Science Advances **4**, eaar4206 (2018).
[14] D. Liu, Y. Tan, E. Khoram, and Z. Yu, ACS Photonics **5**, 1365 (2018).
[15] I. Malkiel, A. Nagler, M. Mrejen, U. Arieli, L. Wolf, and H. Suchowski, arXiv preprint arXiv:1702.07949 (2017).
[16] W. Ma, F. Cheng, and Y. Liu, ACS Nano (2018).
[17] P. Baldi, P. Sadowski, and D. Whiteson, Nature Communications **5**, 4308 (2014).
[18] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, Nature **549**, 195 (2017).
[19] D.-L. Deng, X. Li, and S. D. Sarma, Physical Review X **7**, 021021 (2017).
[20] G. Carleo and M. Troyer, Science **355**, 602 (2017).
[21] J. D. Joannopoulos, S. G. Johnson, J. N. Winn, and R. D. Meade, *Photonic crystals: molding the flow of light* (Princeton university press, 2011).
[22] S. Noda, K. Tomoda, N. Yamamoto, and A. Chutinan, Science **289**, 604 (2000).
[23] T. Zahavy, A. Dikopoltsev, D. Moss, G. I. Haham, O. Cohen, S. Mannor, and M. Segev, Optica **5**, 666 (2018).
[24] S. J. Pan, Q. Yang, *et al.*, IEEE Transactions on knowledge and data engineering **22**, 1345 (2010).
[25] Y. Bengio, in *Proceedings of ICML Workshop on Unsupervised and Transfer Learning* (2012) pp. 17–36.
[26] S. Thrun and L. Pratt, *Learning to learn* (Springer Science & Business Media, 2012).
[27] L. Torrey and J. Shavlik, in *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques* (IGI Global, 2010) pp. 242–264.
[28] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, in *Advances in neural information processing systems* (2014) pp. 3320–3328.
[29] A. A. Rusu, M. Vecerik, T. Rothörl, N. Heess, R. Pas-

canu, and R. Hadsell, arXiv preprint arXiv:1610.04286 (2016).
[30] R. Collobert and J. Weston, in *Proceedings of the 25th international conference on Machine learning* (ACM, 2008) pp. 160–167.
[31] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on* (IEEE, 2013) pp. 7304–7308.
[32] L. Deng, G. Hinton, and B. Kingsbury, in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on* (IEEE, 2013) pp. 8599–8603.
[33] R. Girshick, in *Proceedings of the IEEE international conference on computer vision* (2015) pp. 1440–1448.
[34] M. Bass, C. DeCusatis, G. Li, V. Mahajan, and E. Stryland, "Handbook of optics. vol. iv. optical properties of materials, nonlinear optics, quantum optics," (2010).
[35] W. Qiu, B. G. DeLacy, S. G. Johnson, J. D. Joannopoulos, and M. Soljačić, Optics Express **20**, 18494 (2012).