Knowledge Guided Deep Learning for General-Purpose Computer Vision Applications

Youcef Djenouri[1,2], Ahmed Nabil Belbachir[2], Rutvij H. Jhaveri[3], Djamel Djenouri[4]

[1] University of South-Eastern Norway, Kongsberg, Norway
[2] NORCE Norwegian Research Centre, Norway
[3] Pandit Deendayal Energy University, India
[4] University of the West of England, UK
youcef.djenouri@usn.no, {yodj,nabe}@norceresearch.no, rutvij.jhaveri@sot.pdpu.ac.in, djamel.djenouri@uwe.ac.uk

Abstract. This research targets general-purpose smart computer vision that eliminates reliance on domain-specific knowledge to reach adaptable generic models for flexible applications. It proposes a novel approach in which several deep learning models are trained for each image. Statistical information of each trained image is then calculated and stored with the loss values of each model used in the training phase. The stored information is finally used to select the appropriate model for each new image data in the testing phase. To efficiently select the appropriate model, a kNN (k Nearest Neighbors) strategy is used to select the best model in the testing phase. The developed framework called KGDL (Knowledge Guided Deep Learning) was evaluated and tested using two computer vision benchmarks, 1) ImageNet for image classification, and 2) COCO for object detection. The results reveal the effectiveness of KGDL in terms of accuracy and competitiveness of inference runtime. In particular, it achieved 94% of classification rate in ImageNet, and 92% of intersection over union in COCO dataset.

Keywords: Knowledge-based Learning · Ensemble Learning · Computer Vision · General-Purpose Artificial Intelligence.

1 Introduction

Deep learning has achieved outstanding results in a wide range of applications, including medical applications [1], and intelligent transportation systems [5]. In the domain of computer vision deep learning has been inspired by biological vision in mimicking visual descriptions and learning into computer vision algorithms. However, current deep learning techniques did not yet reach the flexible, general-purpose intelligence that biological systems have. Currently, each model is built using the domain knowledge of the application in question. This motivates researchers and data scientists to investigate this challenging topic. Since the learner does not have to infer the information from the data, integrating

a priori knowledge into the learning framework is an efficient way to deal with sparse data. Several solutions have been explored for the domain knowledge in the learning process. To perform semantic face editing using pretrained StyleGAN, Hou et al. [7] presented a novel learning framework called GuidedStyle. It also made it possible for a StyleGAN generator's attention mechanism to select a single layer for style alteration in an adaptive manner. Therefore, StyleGAN may make disentangled and controllable changes to various features, including attractiveness, mustache, eyeglasses, smiling, and gender. A cooperatively boosting framework (CBF) was proposed [10] to combine the knowledge-guided ontological reasoning module and the data-driven deep learning module. The DSSN architecture is used by the deep learning module, which integrates the original image and inferred channels as input. Branching for intra- and extra-taxonomy reasoning is also included in the module for ontology reasoning. The intra-taxonomy reasoning corrects the wrong classifications made by the deep learning module based on domain knowledge. Dash et al. [4] reviewed the existing solutions that explore domain knowledge. They reported that these solutions have a major limitation in that each model is made using the knowledge that is unique to the application in question. To overcome this limitation, we propose in this paper a novel framework called KGDL (Knowledge Guided for Deep Learning) as an alternative solution for the current computer vision deep learning architectures. To the best of our knowledge, this is the first piece of work that thoroughly examines the information gleaned from the training data to effectively address computer vision difficulties. The main contributions of this research work are:

1. We propose a novel framework called KGDL (Knowledge Guided Deep Learning) that explores the knowledge extracted from the data to efficiently select the best model for each testing data towards general-purpose learning.
2. We develop an intelligent strategy for the inference step in which the statistical information of each image in the testing is first calculated, and then compared with the images of the knowledge base created in the training phase using kNN to select the best model in the inference phase.
3. We evaluate the proposed KGDL framework on two computer vision benchmarks, 1) ImageNet for image classification, and 2) COCO for object detection, using classification accuracy and intersection over union metrics. The results show that the suggested framework outperforms the baseline solutions in terms of the quality of the outcomes at a reasonable cost in inference runtime.

2 Related Work

Yin et al. [15] suggested a new model called Domain Knowledge Guided Recurrent Neural Networks (DG-RNN), which explicitly incorporated domain knowledge from the medical knowledge graph into an RNN architecture. The authors addressed the integration of domain knowledge by dynamically utilizing complex medical knowledge (such as relations between clinical occurrences). Liu et

al. [12] suggested a prior knowledge-guided deep learning-enabled (PK-DL) synthesis method that makes use of the conditional deep convolutional generative adversarial network (cDCGAN) algorithm. Prior information, including familiarity with basic electromagnetic theorems and expertise in antenna design, was purposefully employed early in the proposed process. By directing the image production process with a knowledge network, Hou et al. [7] introduced Guided-Style to perform semantic face editing on pretrained StyleGAN. Additionally, it enabled a StyleGAN generator's attention mechanism to adaptively choose a single layer for style manipulation. As a result, StyleGAN can execute disentangled and controllable modifications along various attributes, such as attractiveness, mustache, eyeglasses, smiling, and gender. Dong et al. [6] suggested a deep HSI denoiser-based iterative hyperspectral image super-resolution (HSISR) approach to take advantage of both deep image prior and domain knowledge likelihood. They demonstrated how to develop an iterative HSISR method into a unique model-guided deep convolutional network by taking the observation matrix of HSI into consideration during the end-to-end optimization (MoG-DCN). The unfolded deep HSISR network may also operate in various HSI scenarios thanks to the representation of the observation matrix by subnetworks, which increases the adaptability of MoG-DCN. For the classification of land cover, Li et al. [11] presented a novel domain knowledge-guided deep collaborative fusion network (DKDFN) with performance-boosting for minority categories. More specifically, a multihead encoder and a multibranch decoder structure are used by the DKDFN. The encoder's architecture makes it likely that enough complementary information may be gleaned from several modalities. The multibranch decoder performs semantic segmentation and reconstructs multimodal remote sensing indices to enable land cover categorization in a multitask learning setup. Li et al. [10] suggested a cooperatively boosting framework (CBF) to iteratively integrate the knowledge-guided ontology reasoning module and the data-driven deep learning module. The deep learning module utilizes the DSSN architecture and uses the DSSN's input to integrate the original image and inferred channels. The module for ontology reasoning also includes branches for intra- and extra-taxonomy reasoning. More particularly, the intra-taxonomy reasoning which is essential to enhance classification performance, directly corrects misclassifications made by the deep learning module based on domain knowledge. To replicate the workflow of radiologists, Mingjie et al. [8] suggested an Auxiliary Signal-Guided Knowledge Encoder Decoder (ASGK). Particularly, the external linguistic signals assist the decoder in better mastering prior information during the pre-training phase, while the auxiliary patches are investigated to increase the frequently used visual patch features before being provided to the transformer encoder. Yang et al. [14] suggested the SEmantic Guided Attention (SEGA) mechanism, in which semantic knowledge is used to direct visual perception top-down regarding which visual cues should be paid attention to when separating one category from the others. As a result, the novel class embedding can be more discriminative even with small sample sizes. To put it more specifically, a feature extractor is trained to transfer visual prior knowledge from base

classes to a few images of each novel class and integrate them into a visual prototype. Then, they developed a network that converted semantic information into category-specific attention vectors, which will be applied to feature selection to improve the visual prototypes.

According to this succinct literature analysis, the key problem with the present deep learning methods is that each model is created using knowledge specific to the application in question. This requires data scientists to be well knowledgeable about the particular application domain. The trained model in this study is created without the assistance of a domain expert, using a general deep learning approach that explores knowledge of the trained data.

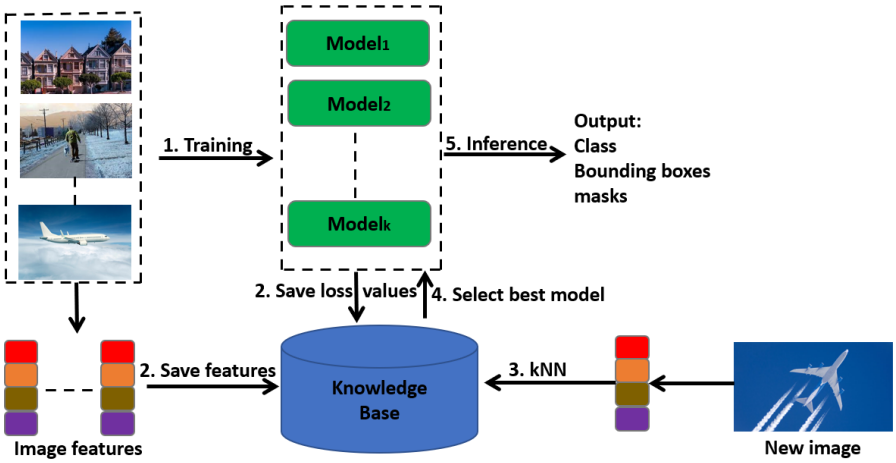

Fig. 1. KGDL Framework: The deep learning models are first trained. The training data is maintained in a knowledge base to accurately combine the results across the trained models. During the inference phase, the model's suitability for a specific set of test data is assessed using the kNN approach.

3 KGDL: Knowledge Guided Deep Learning

3.1 Principle

The KGDL framework is illustrated in Figure 1. It is based on deep learning, kNN, and relevant knowledge from the training and testing data. The main idea is that several deep learning models are trained in the training phase, and then the knowledge base is used to select the best model that will be used in the inference phase for each testing image. First, the data is extracted from the various images. Several deep learning architectures are then trained, and the pertinent data resulting from this training stage is preserved in a knowledge

base. The combined data is further utilized along with kNN to determine which model is appropriate for a given test dataset during the inference phase. Detailed description of the KGDL components is given in the following.

3.2 Training

We consider a set of l images used in the training, say $I = \{I_1, I_2...I_l\}$. The training is performed using the set of n models $\mathcal{M} = \{\mathcal{M}_1, \mathcal{M}_2, ...\mathcal{M}_n\}$. Each image I_i is plugged to each model $\mathcal{M}_j$ for the training. The loss value v_{ij} is determined by computing the error between the output of the model, $\mathcal{M}_j$, and the ground truth associated to the image, I_i. The features of I_i (denoted F_i) and the loss value, v_{ij}, are saved into the knowledge base. The standard back-propagation is also used to optimize the weights of the models in $\mathcal{M}$. At the end of this step, the following variables are created and saved:

1. n matrices, each one, say matrix $W^{(i)}$, represents the trained weights of the model, $\mathcal{M}_i$.
2. The knowledge base KB, which contains l rows. The i^{th} row contains the relevant information of the image I_i. It contains the features F_i, and the set of the n loss values $\{v_{i1}, v_{i2}...v_{in}\}$.

The loss values will be computed using the loss functions according to the considered problems. For instance, Binary Cross-Entropy Loss is used for classification problem as follows:

$$L(y, y^*) = -y \times log(y^*) - (1 - y) \times log(1 - y^*), \tag{1}$$

where y is the ground truth value, y^* is the predicted value by the model.
Diss loss could be used for segmentation problem, as follows:

$$L(y, y^*) = 1 - \frac{2 \times y \times y^* + 1}{y + y^* + 1}. \tag{2}$$

1 was appended to the numerator and denominator to prevent the function from being undefined in extreme cases, such as when $y = y^* = 0$.

For the hyperparameter optimization of the n models, we adopt the recent greedy search algorithm (GHO) [3]. In order to converge to the local optimal solution with the hope that this decision will result in a global optimal one, the GHO algorithm optimizes every hyperparameter while holding the others constant. Up until all of the hyperparameters are optimized, the local solution for each one is optimized iteratively. Therefore, the greedy algorithm reduces the exponential computational cost of the hyperparameter optimization.

3.3 Inference

In the inference stage, the features F_{new} of the new image I_{new} are determined. The knowledge base KB is then explored to select the best model using kNN

algorithm. It calculates the separations between the features of the new image and the features of all training images in the knowledge base. Thus, it discovers the images that resemble the new image the most. A user-specified integer k determines the neighborhood size. The neighborhood of an image is defined in the space of the measured distances. The best model can then be chosen to be utilized, inferring the output of the new image by receiving majority votes or an average guess of the k nearby neighborhood, which are the k closest images in terms of distance. We propose a variant of kNN that computes the distances adjusted to accommodate the image features, as the conventional Euclidean distance measure would produce inaccurate distances for image data. The primary explanation is that the data drift problem has resulted in quite varied distributions for images from various classes. Therefore, it is hard to measure the similarities between images accurately. Instead of using the manually created similarity measures directly to solve this problem, we suggest an end-to-end similarity metric learning network. The proposed similarity metric consists of two modules:

1. **Similarity metric network**: It is a fully connected neural network that seeks to determine how similar the features of two images are. To assess the degree of similarity between the trained images and the new image, we use a fully connected neural network with a single hidden layer. The inputs of the subsequent similarity measurement function are the feature vector of the new image (F_{new}) and the feature vector of each trained image (F_{I_i}):

$$S(F_{new}, F_{I_i}) = 1 - \sigma(concat([F_{new}, F_{I_i}])C), \tag{3}$$

   C is is the coefficient of similarity metric function.
2. **Smooth similarity loss learning**: Backpropagation optimization of the similarity metric function is done by measuring the surrogate loss of the similarity network developed in the first step. Synthetic images are captured from several distributions when training the network. To compute the ground truth (the similarity value), we determine the similarity between the distributions of the images at hand.

Finally, the weights of the best model are used to infer the output of the new image. Indeed, if we assume $\mathcal{M}_{best}$ will be the best model, and $W^{(best)}$ will be the weights of the best model, the new input image is fed into the network's input layer of $\mathcal{M}_{best}$, which passes it through the network layer by layer. Each layer performs a weighted sum of the inputs by $W^{(best)}$ and applies an activation function to produce an output using the forward propagation mechanism.

4 Numerical Results

To evaluate the KGDL framework, intensive simulation have been carried out using well-known benchmarks to compare it with recent deep learning solutions in solving computer vision based applications.

Setting Details We will first go through the details of our experiment in this section. Then, we will compare our classification, and object detection results to those of baseline models. ImageNet and COCO are well-known computer vision benchmarked datasets[5],[6]. We chose these benchmarks and undertake experiments on the ImageNet 2012 ILSVRC challenge classification task and on the COCO challenge object detection task. We utilize a batch size of 2048 by default for labeled images, and we decrease the batch size when the model cannot fit in the memory. We discover that employing 512, 1024, or 2048 batch sizes result in the same speed. The batch size for labeled images is used to calculate the number of training epochs and the learning rate. With a dropout rate of 0.5, we apply dropout to the last layer of our framework and the baseline models.

Baseline Methods We compare the proposed KGDL framework with the following baseline methods: 1) Classification: We use two recent algorithms for comparison of the classification task, namely Revised RESNET [2] and MViTv2 [9]. 2) Object Detection: We use two algorithms for comparison regarding the object detection task, MViTv2 [9] and Improved Yolov5 [13].

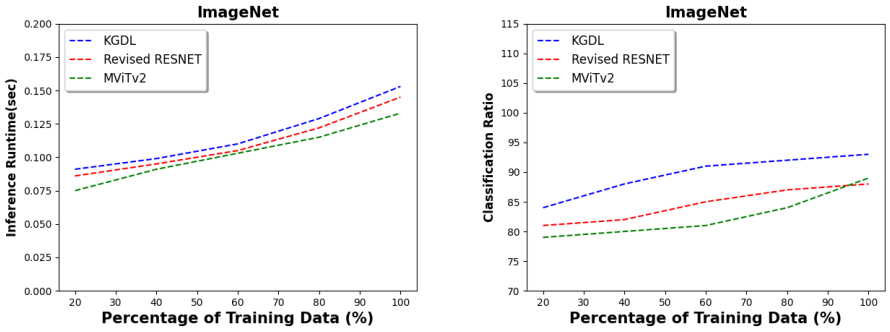

Fig. 2. Classification rate and Runtime of the proposed solutions and the SOTA models for different training samples of the ImageNet.

Results on Image Classification Using the previously described ImageNet, the initial experiments compare the KGDL's accuracy against SOTA image classification methods (Revised RESNET, and MViTv2). Figure 2 demonstrates that KGDL surpasses the two baseline algorithms in terms of classification rate and it is very competitive in terms of inference runtime when the percentage of the number of images used as input is varied from 20% to 100%. Thus, the classification rate of the KGDL is 93% whereas the baseline methods go below 90% when

[5] https://www.image-net.org/
[6] https://cocodataset.org/

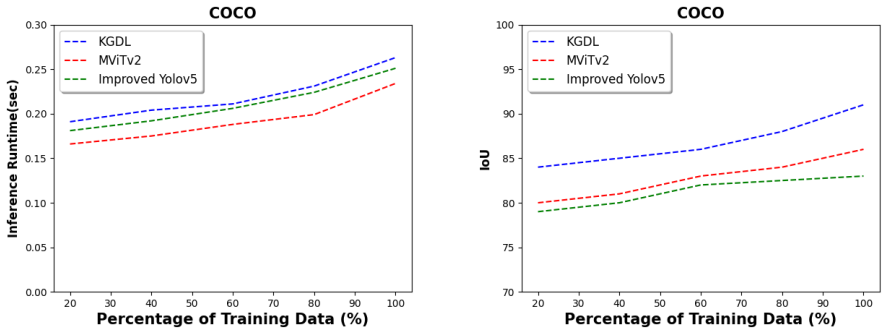

Fig. 3. Performance of the proposed solutions and the SOTA models for object detection use case using COCO dataset.

the entire ImageNet is processed in the training phase. These results are obtained thanks to the selective strategy used in the inference step, which explore the knowledge base to find the best model for each testing image.

Results on Object Detection Using the previously described COCO dataset, the next experiments compare the KGDL's accuracy against SOTA object detection methods (MViTv2, and Yolov5). Figure 3 demonstrates that KGDL surpasses the two baseline algorithms in terms of IoU (Intersection over Union) and it is very competitive in terms of inference runtime. The IoU of the KGDL is 91% whereas the baseline methods remains below 86% when the entire COCO dataset is processed in the training phase. These outcomes were again made possible by the inference step's selective strategy, which looked through the knowledge base to identify the most appropriate model for each testing image.

5 Discussion and Future Perspectives

In this section, we go over some current difficulties and major problems with the built KGDL framework, and by considering such framework as a foundation, we demonstrate potential future paths for computer vision applications. The first challenge of the KGDL is to find a smart way for adding human experience and knowledge to computer vision tasks. By examining the earlier works, we discover that the majority of studies which integrate the human experience, only concentrate on natural language processing. Understanding the causes makes it clear that adding human experience and knowledge to the model at every stage is difficult, with the exception of direct labeling. To solve this issue, we plan to integrate inverse reinforcement learning in the KGDL framework. It entails extrapolating another agent's hidden preferences from its observed behavior, avoiding the need to manually specify its reward function. Therefore, the interaction between the environment (human experience in our case), and the agent (KGDL framework

in our case) will be done automatically and without the need to manual assessment. The second challenge is how to design an evaluation benchmark for knowledge-guided deep learning. The existing solutions including KGDL framework consider standard benchmarks such as ImageNet, PASCAL VOC, CIFAR and MNIST. Creating a useful test benchmark is essential for the community's development of knowledge-guided deep learning. In order to effectively explore this research topic, it is crucial to discover how to create benchmarks and evaluation methodologies for knowledge-guided deep learning. We plan to investigate the use of attention diversification in building benchmarks specified for knowledge-guided deep learning. It consists to reassign appropriate attention to diverse task-related features for domain generalization. We will inspire attention diversification for designing both the training and testing data for evaluating the knowledge-guided deep learning-based frameworks. The third challenge is to make multi-task learning into practice. It is difficult to totally tackle a real-world task with just one categorization because it is complex and usually required intensive computation and intelligent learning processes. We have seen optimism for a universal model through human-in-the-loop fine-tuning with the emergence of a unified large-scale pre-training method. we plan to adopt a suitable method to incorporate human knowledge into huge models as existing machine learning models, in particular, are not as intelligent as humans.

6 Conclusion

This work has addressed the challenges related to establishing general-purpose and flexible AI using the existing deep learning models and proposed a novel general-purpose deep learning approach for tackling generic computer vision applications. For each set of visual data, many deep learning models have first been trained. Following that, the statistical data for each trained image is computed and stored along with the loss values of each model used during the training. In the testing step, the right models for each new set of image data are ultimately chosen using the stored information. A kNN (k Nearest Neighbors) technique is employed to effectively choose the optimal model during the testing phase. ImageNet benchmark was used to evaluate the created knowledge-guided deep learning system. The outcomes presented validated the KGDL framework's higher accuracy and strong inference runtime competitiveness compared to the baseline methods. Since the runtime of the KGDL is critical, in particular for real-time processing based applications, we plan to improve the knowledge base exploration by investigating on kNN query processing techniques for finding the best model in the inference phase.

Acknowledgment

This work is funded in part by the Research Council of Norway's ULEARN "Unsupervised Lifelong Learning" project, which is co-funded under grant number 316080.

References

1. Belhadi, A., Djenouri, Y., Diaz, V.G., Houssein, E.H., Lin, J.C.W.: Hybrid intelligent framework for automated medical learning. Expert Systems **39**(6), e12737 (2022)
2. Bello, I., Fedus, W., Du, X., Cubuk, E.D., Srinivas, A., Lin, T.Y., Shlens, J., Zoph, B.: Revisiting resnets: Improved training and scaling strategies. Advances in Neural Information Processing Systems **34**, 22614–22627 (2021)
3. Chowdhury, A.A., Hossen, M.A., Azam, M.A., Rahman, M.H.: Deepqgho: Quantized greedy hyperparameter optimization in deep neural networks for on-the-fly learning. IEEE Access **10**, 6407–6416 (2022)
4. Dash, T., Chitlangia, S., Ahuja, A., Srinivasan, A.: A review of some techniques for inclusion of domain-knowledge into deep neural networks. Scientific Reports **12**(1), 1–15 (2022)
5. Djenouri, Y., Belhadi, A., Lin, J.C.W., Cano, A.: Adapted k-nearest neighbors for detecting anomalies on spatio–temporal traffic flow. IEEE Access **7**, 10015–10027 (2019)
6. Dong, W., Zhou, C., Wu, F., Wu, J., Shi, G., Li, X.: Model-guided deep hyperspectral image super-resolution. IEEE Transactions on Image Processing **30**, 5754–5768 (2021)
7. Hou, X., Zhang, X., Liang, H., Shen, L., Lai, Z., Wan, J.: Guidedstyle: Attribute knowledge guided style manipulation for semantic face editing. Neural Networks **145**, 209–220 (2022)
8. Li, M., Liu, R., Wang, F., Chang, X., Liang, X.: Auxiliary signal-guided knowledge encoder-decoder for medical report generation. World Wide Web pp. 1–18 (2022)
9. Li, Y., Wu, C.Y., Fan, H., Mangalam, K., Xiong, B., Malik, J., Feichtenhofer, C.: Mvitv2: Improved multiscale vision transformers for classification and detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4804–4814 (2022)
10. Li, Y., Ouyang, S., Zhang, Y.: Combining deep learning and ontology reasoning for remote sensing image semantic segmentation. Knowledge-Based Systems **243**, 108469 (2022)
11. Li, Y., Zhou, Y., Zhang, Y., Zhong, L., Wang, J., Chen, J.: Dkdfn: Domain knowledge-guided deep collaborative fusion network for multimodal unitemporal remote sensing land cover classification. ISPRS Journal of Photogrammetry and Remote Sensing **186**, 170–189 (2022)
12. Liu, P., Chen, L., Chen, Z.N.: Prior-knowledge-guided deep-learning-enabled synthesis for broadband and large phase shift range metacells in metalens antenna. IEEE Transactions on Antennas and Propagation **70**(7), 5024–5034 (2022)
13. Qu, Z., Gao, L.y., Wang, S.y., Yin, H.n., Yi, T.m.: An improved yolov5 method for large objects detection with multi-scale feature cross-layer fusion network. Image and Vision Computing **125**, 104518 (2022)
14. Yang, F., Wang, R., Chen, X.: Sega: semantic guided attention on visual prototype for few-shot learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1056–1066 (2022)
15. Yin, C., Zhao, R., Qian, B., Lv, X., Zhang, P.: Domain knowledge guided deep learning with electronic health records. In: 2019 IEEE International Conference on Data Mining (ICDM). pp. 738–747. IEEE (2019)