Evaluating the Effect of User-Given Guiding Attention on the Learning Process

Richard Nordsieck*, Michael Heider†, Andreas Angerer* and Jörg Hähner†
*XITASO GmbH IT & Software Solutions, Augsburg, Germany
{richard.nordsieck, andreas.angerer}@xitaso.com
†Organic Computing Group, University of Augsburg, Augsburg, Germany
{michael.heider, joerg.haehner}@informatik.uni-augsburg.de

***Abstract*—Most current supervised learning systems require large quantities of labelled data, limiting their applicability in domains where labelled data is scarce and hard to obtain. We introduce a novel approach for incorporating additional, user-given areas of interest during training by which the learning process can be guided. The provided guiding attention is incorporated in the training phase as a form of data augmentation, which ensures that input dimensions do not vary between train and test/deployment time, when no guiding attention is present. We evaluate this approach by extending the CIFAR-10 dataset with prototypical information and ascertain, that our approach reduces the required amount of samples by up to 44.89%, when combined with traditional data augmentation techniques. This would enable the use of learning systems in parts of manufacturing such as commissioning, where additional samples are scarce and costly to obtain while providing guiding attention is a matter of seconds.**

***Index Terms*—expert knowledge, artificial neural networks, data augmentation, guiding information**

I. Introduction

In the producing industry, there is a continuous interest in self-adapting and self-optimizing production processes [1]. The scope spans from single machines to complete production lines, thriving to produce parts with the best possible quality, using the least possible amount of resources. In recent years, data-driven methods have received increased interest in the production domain [2], [3]. Machines produce, besides the physical products, also large amounts of unstructured data and thus employing data driven methods seems promising. In today's factories, the products are usually inspected by the machines' operators (or specialized quality assurance personnel; we subsume the different roles as 'operator' here for the sake of simplicity), cf. Figure 1. The data produced by machines is oftentimes also inspected by operators on the shopfloor, e.g. using supervisory control and data acquisition (SCADA) systems for aggregation and visualization. The use of reinforcement learning methods is being investigated in the manufacturing context [4]. In theory, this approach (cf. Figure 2) could take operators out of the loop, creating fully autonomous self-x systems. However, the introduction is challenging in practice, as it either involves blocking machines for a significant amount of time and wasting production resources during the training phase, or needs simulation environments with adequate precision, which is often not feasible for complex production processes.

This paper focuses on the use of supervised learning approaches, addressing the unique constraints present in the manufacturing domain. The goal is to work on data and products generated during the normal production duty cycle and/or during commissioning of machines. The challenge here is often the limited availability of the *right* data—in particular, that of problematic production runs—in sufficient quality and quantity. Since supervised learning systems operate solely on problems and their solutions, they require vast amounts of labelled data. This is in strong contrast to human learning, where psychological studies have illustrated the importance of strong guidance of the learner [5] through explanations, such as explanatory solution paths [6] and previous knowledge [7]. Drawing inspiration from these aspects of human learning, we propose to get humans to provide additional information which can serve to explain relationships between input data and its label, thereby guiding the model's learning process in the right direction.

Perspectively, we aim to build a socio-technical system consisting of human operators and a supervised learning agent that can give increased decision assistance to the operator over time, while improving its performance based on labels and guiding information provided by the operator (cf. Figure 3). The explicit integration of the operator will lead to an increased robustness of the system compared to current SCADA systems. Specifically, it will exhibit self-improving characteristics in regards to prediction quality, which is improved by the operator's guiding information. Self-adaptiveness is increased through the interplay between the learning system's recommendations to the operators and the operator's expert knowledge that is passed to the learning system.

In this paper, we fundamentally examine the effect of employing guiding information—more specifically guiding attention—in a supervised learning setting. To get an initial grasp on what is achievable by this kind of augmentation in principle, we chose the setting for our study as follows:

- Guiding attention is supplied during the data labelling process, where experts give additional information about their attention focus during the labelling. The long-term goal is to collect such information during the operators'

This work is supported by the German Federal Ministry for Economic Affairs and Energy (BMWi).

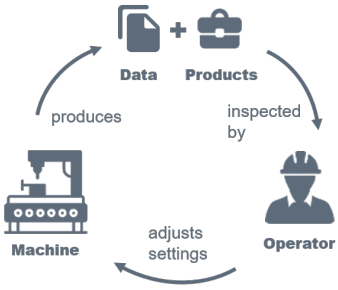

Fig. 1. Operator-in-the-loop in today's productions.

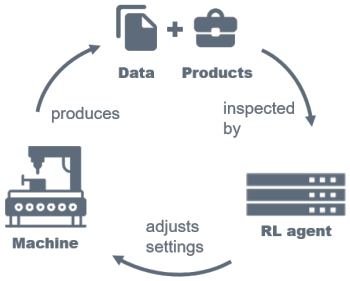

Fig. 2. Self-x production using agent trained with reinforcement learning.

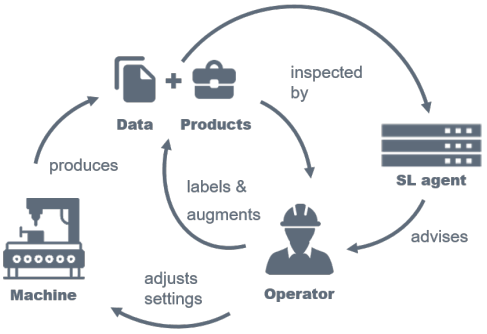

Fig. 3. Assisted production using agent trained with supervised learning during operation.

idle times or extract it from interaction data between operators and machines.

- We start with a standard image dataset (CIFAR-10) [8] and thus extend an image labelling workflow. We chose CIFAR-10 because it seemed easy for a laymen to provide the required guiding attention. Once we have a good understanding of the possibilities and limitations of this approach, we aim for a transfer to production datasets and labelling workflows of our industrial partners. The proposed approach is not limited to images but could also be applied to time-series, natural language texts and parameter sets among others, and is directly applicable to industrial settings e.g. by assisting the operator during quality assurance.
- The focus is explicitly *not* on competing with state of the art accuracy on this kind of dataset. Instead, we purely focus on the effect of additional guiding attention applied as data augmentation.

We address the following research question in this work: Can we compensate the lack of more training data by augmenting the existing data with information about where the expert's attention was focused when deriving the label?

The remainder of this work is structured as follows: Section II gives an overview of related approaches. Section III presents the methodology used in the presented approach. The dataset used for evaluation is explained in Section IV, while Section V details the experimental setup used. Section VI goes into detail about the achieved results until now. An outlook and conclusions are given in Section VII and Section VIII.

II. Related Work

To the best of our knowledge, this approach of incorporating more labels to be used as an augmentation technique has not been investigated before with research typically focusing on: (1) Boosting the available labelled data through various means such as data augmentation [9] (rotations, mirroring, gray-scaling etc.); (2) utilizing unlabelled additional samples, e. g. to obtain a better initialization by pre-training in an unsupervised setting [10] or requesting labels from experts for edge cases in Active Learning [11], [12], or (3) by learning representations on other related and similar datasets and transferring the knowledge onto the existing task [13]. However, there are multiple branches of research that are conceptually related.

There are multiple ways of combining expert knowledge and learning systems such as artificial neural networks. Most research has been focused on knowledge-base completion for which knowledge graph embeddings have been frequently used to solve tasks such as link prediction and entity classification [14]–[17]. Another approach are Graph Convolutional Networks, that directly operate on undirected graphs [18]. As knowledge-graphs are often directed to further specify relations between entities, they have been extended to operate on directed knowledge-graphs [19] and used for knowledge-base completion. Embeddings of knowledge bases have been successfully applied to increase performance of neural networks for text understanding [20], [21] and recommendations [22]–[25]. Tandon et al. have illustrated that common-sense knowledge is crucial to allow the application of learning systems to reasoning tasks [26]. All these works focus on relational rather than procedural or rule-based knowledge. Also, they depend on the complete additional knowledge to be present in a structured knowledge graph, that is assumed to be created before their application. This sets them apart from our approach, which is able to work with a fractioned knowledge representation and perspectively aims to acquire additional knowledge in an unobtrusive fashion.

Conceptually, our approach is related to attention mechanisms, that have proven beneficial in tasks such as image captioning [27], machine translation [28] and tracking [29]. However, in contrast to these works the attention present in our approach is given by humans.

On a very high level of abstraction, our approach is related to active learning as described by Settles [12], insofar as active learning approaches query information during the learning process and consequently also deal with incomplete information.

III. Methodology

A supervised learning system is generally defined as a function $f_\theta : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X} \subseteq \mathbb{R}^{D_\mathcal{X}}$ and $\mathcal{Y} \subseteq \mathbb{R}^{D_\mathcal{Y}}$ with $D_\mathcal{X}, D_\mathcal{Y} > 0$ are input- and output space, respectively.

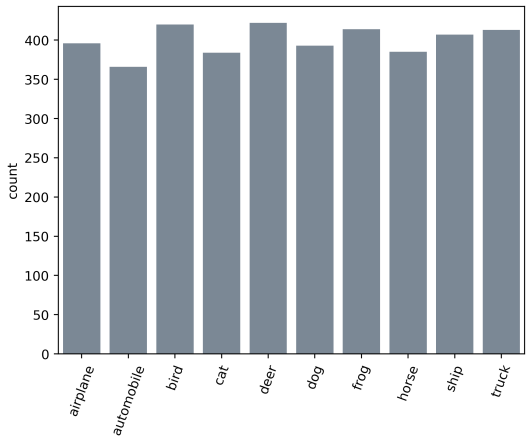

Fig. 4. Distribution of user-given guiding attention by class name.

Importantly, it is assumed that $\mathcal{X}$ and $\mathcal{Y}$ remain unchanged throughout training and test time (and also after deployment). In the scenario as outlined in Section I this is not directly the case as guiding attention is provided voluntarily and therefore inherently intermittent. To be able to build upon the achievements of learning systems research, we propose to introduce the guiding attention through augmentations of the training set. Let $X_A = \{x \,|\, a_x \in A\}$ denote the subset of the complete input data X for which guiding attention $a \in \mathcal{A} \subseteq \mathbb{B}^{D_\mathcal{X}}$ is present, then $X' = \{X, aug(X_A, A)\}$ denotes the new - extended - training data and $aug : (\mathcal{X}, \mathcal{A}) \to \mathcal{X}$ a suitable function for augmenting each $x \in X_A$ with the corresponding $a \in A$. This leads to an increase in training data by $|A|$, $|X^{'}| = |X| + |A|$. Note that, $X_A \subseteq X$ and therefore $x \in X$ for which no guiding attention might be available can exist and still be incorporated into the learning process as the input space $\mathcal{X}$ remains unchanged. Also, $aug(x, a)$ is specific to the chosen domain and data format. As such, it is described in Section V.

IV. Dataset

To evaluate our methodology we extended the CIFAR-10 dataset [8] with user-given guiding attention. The original dataset incorporates 60000 32x32 pixel color images, evenly distributed over ten distinct classes. There has been a plethora of research regarding CIFAR-10 with state of the art results at around 99% correctly classified images[1]. Some quite recent examples forming this impressive state of the art include BiT [13], GPipe [30] and EfficientNet [31]. However, in contrast to our neural architecture (cf. Section V) they rely on very intricate, vastly more complex designs and/or prior knowledge (transfer learning) rather than just the dataset itself.

We expanded on the original CIFAR-10 data by having two persons independently give guiding attention for 2000 samples each, highlighting areas which they deemed relevant for the respective classification, thereby mimicking an attention mechanism. The general idea is to point the learning system to where it should derive its decision from through *guiding attention*, limiting the influence of noisy or irrelevant background features. The selection of an area of interest was done by selecting the pixel that constitutes the center of the respective area of interest. The amount of areas of interest was not restricted per image and could be corrected by deleting previously highlighted areas of interest, if applicable adjusting their position by re-placing them. After an image was confirmed the users had to choose between correcting it and continuing with the next image. This information was appended to the dataset, which now has the properties X, y, A, where $a \in A$ is the user given guiding attention for an image.

The distribution of the 4000 samples including guiding attention is shown in Figure 4. The slight imbalance between classes that can be observed there is deliberately included as slight class imbalances are very common in real-world use-cases. To alleviate a strong impact, the disparity is kept to a relatively low amount of samples with guiding attention. Columns (a) and (b) in Figure 5 show the original image and a visualization of the selected centers of areas of interest, respectively. Overall an average of 13.82 areas of interest were selected per image.

Fig. 5. Samples of the dataset and the applied preprocessing visualized. Columns (a) and (b) show the original images and their user-given guiding attention, respectively. Column (c) shows the mask obtained by applying $d(x)$. The resulting images that are augmented to the training set are shown in Column (d).

V. Experimental Setup

To integrate the user-given guiding attention as data augmentation cf. Section III we need to find a suitable function $aug(x, a)$ to preprocess the data. Since the guiding information

[1]https://paperswithcode.com/sota/image-classification-on-cifar-10

in our case derives from the users' attention towards certain areas of interest, we decided to apply blur to the parts of an image the user deemed unimportant. This removes, or at least reduces, possibly distracting irrelevant information contained in the background of the object to be classified. Therefore, we compute a mask that encodes the differing degrees of importance of each pixel of an input image. This is done by determining the distance to the nearest center of a highlighted area $f_A(x_i) = min\{|x_i - a_i| : a_i \in a, a_i = 1\}$, and passing that distance to a function $d(x_i) = \max(0, -0.1x_i + 1)$. The resulting mask is then used to blur the original image by multiplying and scaling it with the average color information. The result can be seen in Figure 5, where Columns (c) and (d) show the computed mask and its application to the original image, respectively. Relevant areas are visible but the discernible area is substantially reduced and the background blurred, indicating that there are fewer edges to detect by the models, which we assume results in a reduction of complexity.

The model's architecture is a relatively shallow convolutional neural network (CNN)—implemented in Tensorflow [32]—with two subsequent blocks of Convolution (3x3 kernel), ReLu Activation, Convolution (3x3 kernel), ReLu Activation, MaxPooling (2x2 pools) and Dropout with 0.25. This is followed by a Fully Connected Layer, ReLu Activation, Dropout with 0.5, another Fully Connected Layer and a Softmax Activation to map the output to the respective classes. The model's architecture is deliberately kept comparatively simple, as we aim to investigate the fundamental effects of introducing guiding attention through data augmentation.

All models are trained for 15 epochs where convergence was reached most of the time. The learning rate was set to 0.001 and a batch size of 64. Hyperparameter optimization was omitted, since we want to illustrate the data augmentation's effect and therefore the model's exact performance is not relevant. Each experiment was executed 20 times to limit stochastic influences.

VI. Evaluation

To evaluate our approach we present three experiments. The first validates whether the dataset extension is viable at all, the second inspects how different amounts of normal to augmented data affect accuracy and the third evaluates the approach in combination with traditional data augmentation.

A. Validate Dataset Extension

To validate that the user-given guiding attention is indeed suited to be applied as data augmentation, this experiment compares the performance of the chosen architecture on both a subset of unaugmented, normal CIFAR-10 data, as well as a set where both training and test data are augmented with preprocessed data to contain guiding attention cf. Section V. Therefore, the respective datasets are limited to the 4000 samples which constitute X_A, that is for which guiding attention is present. Both were split to 2666 training and 667 validation and test data, respectively.

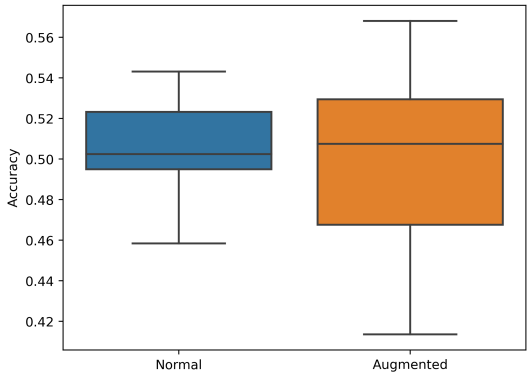

Fig. 6. Results of Experiment A: Accuracy for datasets consisting of only preprocessed images and normal, non-preprocessed images, respectively.

The results are shown in Figure 6. They are comparable with regards of medians of 50.74% and 50.24% accuracy, respectively. It has to be noted, however, that standard deviation for the augmented dataset is higher than for the unaugmented one with 4.10% and 2.26%, respectively.

These results illustrate two things. Firstly, that the chosen preprocessing does not have a detrimental effect on the prediction task. In extension to this observation we can assume that the user-given attention indeed still contains valuable information. Secondly, since the uninteresting parts of an image are blurred in the augmented dataset and performance is still comparable we can conclude that the unaugmented model was not focusing on features of the images that do not pertain to the actual object in question, such as clouds in the background of birds.

B. Effect of Different Percentages

To evaluate the effect of additional information on the learning process we present this experiment, which compares achievable accuracy for separate amounts of normal images ($|X|$) and images that are augmented ($|X_A|$) by applying the preprocessing as described in Section V with the user-given guiding attention. As in industrial use-cases—on which this approach will be tested and realized in the future—low data availability is prevalent we chose to focus on comparatively small amounts of normal images. Specifically, $N = \{4000, 6000, 8000, 10000, 20000\}$ and $I = \{0, 1000, 2000, 3000, 4000\}$, where N is amount of normal images, and I the amount of preprocessed images serving as data augmentation. $i = 0$ serves as a baseline, since it indicates that no preprocessed images are included in the training data X', leaving it unmodified and unguided. All combinations are validated on the test portion of CIFAR-10, i.e. the last 10000 images, which are evenly distributed among classes.

The results of the runs for each configuration are shown in Figure 7. The x-axis displays the amount of normal images, while the y-axis illustrates the accuracy on the validation set achieved by the respective combinations of n and i. These are displayed as box-plots containing data of 20 runs using

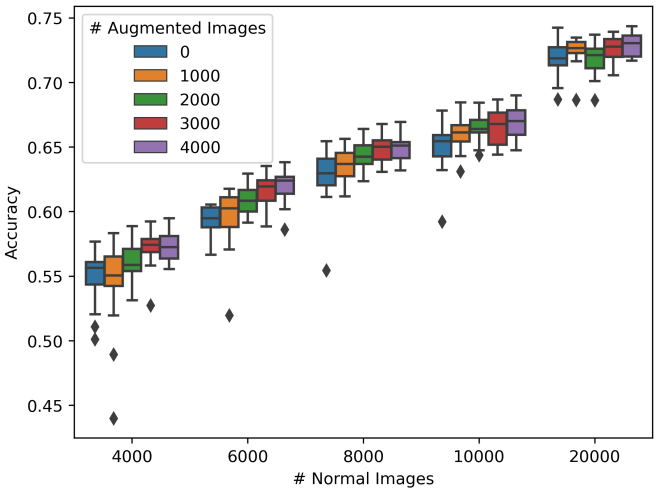

Fig. 7. Accuracy on test set for varying amounts of unaugmented, normal images by amount of images containing guiding attention. No traditional data augmentation was applied.

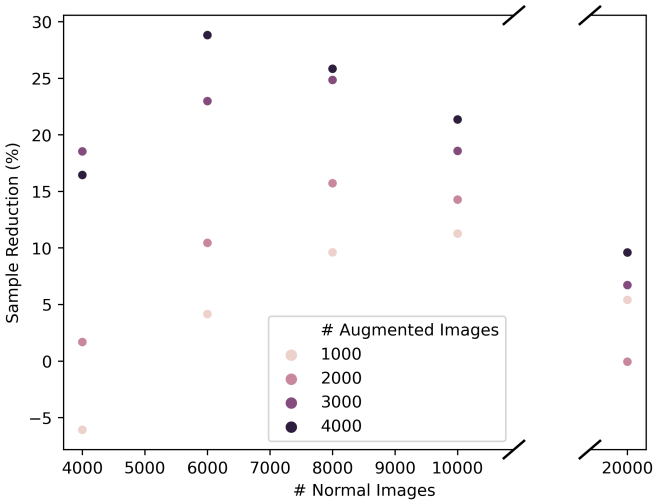

Fig. 8. Relative sample reduction in %, illustrating how much additional unaugmented data would be required to achieve a comparable result. No traditional data augmentation was applied.

different random seeds. One can observe that in general accuracy increases with increasing training size. This is to be expected since more samples (usually) offer a wider variety of data and as such better generalization to the test set. However, it should be noted that the x-axis steps are not completely evenly-spaced, therefore the jump in accuracy increase between $n = 10000$ and $n = 20000$ is not surprising. In fact, the increase can be characterized as asymptotic, slowing with greater n. If we consider the different quantities of preprocessed images, we can observe that generally the accuracy also increases with increasing amounts of guiding attention used in training. $n = 4000$ / $i = 1000$ and $n = 20000$ / $i = 2000$ are the only significant exceptions from this trend. The maximum increase in median accuracy for each $n \in N$ is limited to 1.8, 2.9, 2.2, 1.6 and 1.2 percentage points, respectively. For all combinations of n and $i > 0$ a mean median benefit on accuracy of 1.21 ± 0.68 percentage points was achieved over the baseline $i = 0$. The largest mean median increase in accuracy was at $i = 4000$ with 1.88 ± 0.68 percentage points.

While these relatively small numbers might seem a bit disenchanting it is useful to have a look at what these numbers imply. As previously observed, the accuracy asymptotically increases with growing training sizes. Therefore, we calculate how many samples of unaugmented training data are corresponding to the achieved accuracies, where augmented samples were used. To achieve this, we first fit a logarithmic function to the median baseline of unaugmented runs. This results in $a(n) = -0.2895 + 0.1021 \ln(n)$, that maps a number of unaugmented images to its expected accuracy. We can reformulate this as $n^{'}(a) = \exp(\frac{a+0.2895}{0.1021})$, which gives us the expected required amount of unaugmented training images for a median given accuracy. Based on $n^{'}(a)$, the percentage of how many data samples that could be saved by applying our methodology can be calculated. The resulting relative percentages are shown in Figure 8.

We can observe that there seem to exist better performing ratios between i and n, as for all i the percentage improves up to a certain n and then begins to decrease, with higher accuracies for higher amounts of augmented guiding images. A probable cause for the increase is that for low n a substantial amount of preprocessed images is added. This could lead to worse accuracy as the weights that are adjusted during the learning process are likely to substantially adapt to the changed semantic of the input images, especially since the test-images are not augmented with guiding attention. Also, it has to be noted, that the proposed metric only measures the impact through the indirection of model performance and as such can be influenced by other hyperparameters that influence the models accuracy and scale differently with the number samples used in training. Most notable of those is the number of epochs, which we assume is responsible for the decreasing improvement with increasing amount of normal images since generally larger training sets can be trained for longer without overfitting. Nonetheless, we achieve a stable improvement of over 15% sample reduction in the lower regions of available sample data which fits our targeted domains. The *highest reduction* is recorded at $n = 6000$ / $i = 4000$ with 28.82%. We also observe an outlier at $n = 4000$ / $i = 1000$, that was also an outlier in the relative trend of improvement through augmentation.

C. Effect in relation to Traditional Data Augmentation

To evaluate the effect of our approach in combination and in contrast with traditional data augmentation techniques, we repeated Experiment B while including traditional data augmentation and compare the results. In the following, we will differentiate between the terms *augmented* and *traditionally augmented* which refer to augmentation via our approach and traditional data augmentation, respectively. Preliminary

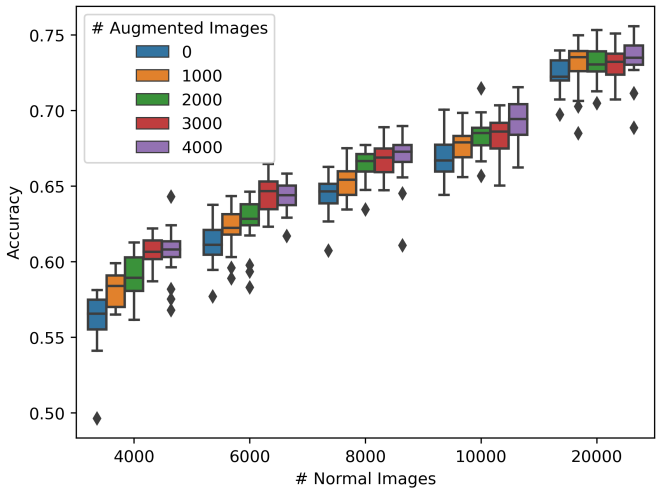

Fig. 9. Accuracy on test set for varying amounts of unaugmented, normal images by amount of images containing guiding attention. Traditional data augmentation was combined with our approach.

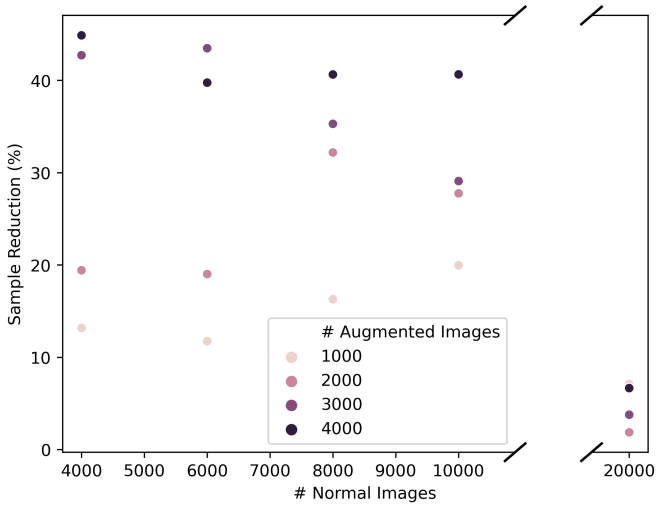

Fig. 10. Relative sample reduction in %, illustrating how much additional unaugmented data would be required to achieve a comparable result. Traditional data augmentation was combined with our approach.

experiments have shown that rotation and shear have a negative effect on the CIFAR-10 dataset, probably due to the small image sizes. Therefore, we limited traditional augmentation in this experiment to horizontal mirroring. Figures 9 and 10 illustrate the results. Due to the different results for the baseline, $n^{'}(a) = \exp(\frac{a+0.2358}{0.0974})$ was used as a basis to calculate sample reduction.

Using this technique, the median accuracy improves by an average of 2.01 ± 0.73 for all combinations of n and $i > 0$ over the baseline $i = 0$. The best increase over the baseline is achieved for $n = 4000$ / $i = 4000$ with 4.25 percentage points, which translates to a sample reduction of 44.89%. In general, the combination resulted in increased sample reduction rates for n compared to Experiment B. The only exception is $n = 20000$, which is likely due to the percentage of added guided attention getting smaller although sub-optimal hyperparameters are also a possible issue.

To quantify the effect of traditional data augmentation we compare the baseline obtained in Experiment C to the unaugmented baseline of Experiment B ($i = 0$; no traditional data augmentation). We observe a benefit in classification accuracy of 1.17 ± 0.49 percentage points averaged over all the medians of all training sizes. The largest increase in accuracy is at $n = 8000$ with 1.70 percentage points. To compare our approach to traditional data augmentation we note, that the average benefit of traditional data augmentation is slightly smaller than the mean median benefit of 1.21 ± 0.68 percentage points that was achieved with our approach in Experiment B, indicating that guided attention is able to achieve better results than the employed traditional technique. Both, our augmentation approach and traditionally augmented data combined lead to a mean median increase of 2.05 ± 0.08 percentage points for all combinations of n and $i > 0$ over the results from Experiment B. This translates to an increase in mean sample reduction of 11.77 ± 2.24 percentage points over the results from Experiment B.

VII. Future Work

We currently plan on investigating the effect of our approach on state of the art models and multiple data sets that allow more aggressive traditional data augmentation methods to be used. Also, we would like to further investigate the effect of varying numbers of areas of interest as well as providing users with the ability to define them without geometric constraints. Furthermore, the approach's effect on different model architectures should be researched to quantify our notion that it is model-agnostic.

The positive results presented in Section VI encourage us to try and adapt our approach to challenging domains, e. g. manufacturing, in the future. A first step could be the application to complex image recognition tasks, that require context knowledge. A second, more complex adaption is to non-image datasets, i. e. machine parameters, environmental conditions and resulting part quality. This would necessitate more fundamental changes to $aug(x, a)$. In the industrial context, the combination with eye trackers could be investigated to unobtrusively gather guiding attention.

To strengthen the socio-technical aspect of such systems and give the user a more fulfilling role than simply providing labels and parameters, the combination of our approach and active learning could be investigated. This would lead to interesting query strategies, since the learners would gain the ability to differentiate between requesting labels or explanations.

VIII. Conclusion

In this paper we investigated the effect of incorporating guiding attention into a classification task through data augmentation. Guiding attention is provided by humans and indicates where relevant information to derive the correct label from is located in a sample. We collected guiding attention

on 4000 randomly selected images of the CIFAR-10 dataset in form of a bit mask where centers of areas of interest result in a True value. We then used this guiding attention by increasingly blurring pixels based on their distance to the nearest center. In three experiments we showed: Firstly, that training with only those augmented images achieves comparable results to using the same 4000 images without the guiding attention. Secondly, that the proposed approach increases performance equal to using up to 28.82% additional traditionally labelled samples—an effect that was especially prominent for small numbers of available training samples. Lastly, we combined our augmentation with traditional data augmentation techniques, again showing that our approach to incorporate guiding attention improves results and can even save up to 44.89% standard samples.

The main application of this technique is in industrial settings where sample generation is often quite costly while machine operator knowledge can be provided mid-production without impacting throughput. E. g. when testing a parameter change results in producing a new batch of product, or quantifying the actual result involves substantial mechanical or biochemical analysis. However, given the results we showed in this preliminary work towards a socio-technical systems we are confident that the application of guiding attention can perform well on the small, labelled data sets prevalent in specific areas of manufacturing.

REFERENCES

[1] E. Permin, F. Bertelsmeier, M. Blum, J. Bützler, S. Haag, S. Kuz, D. Özdemir, S. Stemmler, U. Thombansen, R. Schmitt, C. Brecher, C. Schlick, D. Abel, R. Poprawe, P. Loosen, W. Schulz, and G. Schuh, "Self-optimizing Production Systems," *Procedia CIRP*, vol. 41, pp. 417–422, 12 2016.

[2] Y. Zhang, C. Qian, J. Lv, and Y. Liu, "Agent and cyber-physical system based self-organizing and self-adaptive intelligent shopfloor," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 2, pp. 737–747, 2017.

[3] E. Lughofer, C. Zavoianu, R. Pollak, M. Pratama, P. Meyer-Heye, H. Zörrer, C. Eitzinger, and T. Radauer, "Autonomous Supervision and Optimization of Product Quality in a Multi-stage Manufacturing Process based on Self-adaptive Prediction Models," *Journal of Process Control*, vol. 76, pp. 27–45, 04 2019.

[4] G. Schoettler, A. Nair, J. Luo, S. Bahl, J. A. Ojea, E. Solowjow, and S. Levine, "Deep Reinforcement Learning for Industrial Insertion Tasks with Visual Inputs and Natural Rewards," 2019.

[5] P. A. Kirschner, J. Sweller, and R. E. Clark, "Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching," *Educational psychologist*, vol. 41, no. 2, pp. 75–86, 2006.

[6] R. K. Atkinson, S. J. Derry, A. Renkl, and D. Wortham, "Learning from examples: Instructional principles from the worked examples research," *Review of educational research*, vol. 70, no. 2, pp. 181–214, 2000.

[7] J. R. Anderson, "ACT: A simple theory of complex cognition," *American psychologist*, vol. 51, no. 4, p. 355, 1996.

[8] A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images," Tech. Rep., 2009.

[9] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *Journal of Big Data*, 2019.

[10] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio, "Why Does Unsupervised Pre-Training Help Deep Learning?" *J. Mach. Learn. Res.*, vol. 11, p. 625–660, Mar. 2010.

[11] S. Das, W. Wong, T. Dietterich, A. Fern, and A. Emmott, "Incorporating Expert Feedback into Active Anomaly Discovery," in *2016 IEEE 16th International Conference on Data Mining (ICDM)*, 2016, pp. 853–858.

[12] B. Settles, "Active Learning Literature Survey," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 6, no. 1, pp. 1–114, 2012.

[13] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby, "Big Transfer (BiT): General Visual Representation Learning," 2020, arXiv:1912.11370v3.

[14] T. Dettmers, P. Minervini, P. Stenetorp, and S. Riedel, "Convolutional 2d knowledge graph embeddings," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[15] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu, "Learning entity and relation embeddings for knowledge graph completion," in *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

[16] T. Trouillon and M. Nickel, "Complex and holographic embeddings of knowledge graphs: a comparison," *arXiv preprint arXiv:1707.01475*, 2017.

[17] Z. Wang, J. Zhang, J. Feng, and Z. Chen, "Knowledge graph embedding by translating on hyperplanes," in *Twenty-Eighth AAAI conference on artificial intelligence*, 2014.

[18] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams, "Convolutional networks on graphs for learning molecular fingerprints," in *Advances in neural information processing systems*, 2015, pp. 2224–2232.

[19] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. Van Den Berg, I. Titov, and M. Welling, "Modeling relational data with graph convolutional networks," in *European Semantic Web Conference*. Springer, 2018, pp. 593–607.

[20] B. Yang and T. Mitchell, "Leveraging knowledge bases in lstms for improving machine reading," 2017. [Online]. Available: http://arxiv.org/pdf/1902.09091v1

[21] J. Wang, Z. Wang, D. Zhang, and J. Yan, "Combining knowledge with deep convolutional neural networks for short text classification," in *IJCAI*, 2017, pp. 2915–2921.

[22] Z. Sun, J. Yang, J. Zhang, A. Bozzon, L.-K. Huang, and C. Xu, "Recurrent knowledge graph embedding for effective recommendation," in *Proceedings of the 12th ACM Conference on Recommender Systems - RecSys '18*, X. Amatriain, S. Pera, J. O'Donovan, and M. Ekstrand, Eds. New York, New York, USA: ACM Press, 2018, pp. 297–305.

[23] H. Wang, F. Zhang, X. Xie, and M. Guo, "Dkn: Deep knowledge-aware network for news recommendation," in *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, 2018, pp. 1835–1844.

[24] X. Wang, D. Wang, C. Xu, X. He, Y. Cao, and T.-S. Chua, "Explainable reasoning over knowledge graphs for recommendation." [Online]. Available: http://arxiv.org/pdf/1811.04540v1

[25] D. Yang, Z. Guo, Z. Wang, J. Jiang, Y. Xiao, and W. Wang, "A knowledge-enhanced deep recommendation framework incorporating gan-based models," in *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 17.11.2018 - 20.11.2018, pp. 1368–1373.

[26] N. Tandon, A. S. Varde, and G. de Melo, "Commonsense knowledge in machine intelligence," *ACM SIGMOD Record*, vol. 46, no. 4, pp. 49–52, 2018.

[27] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048–2057.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[29] A. Kosiorek, A. Bewley, and I. Posner, "Hierarchical attentive recurrent tracking," in *Advances in Neural Information Processing Systems*, 2017, pp. 3053–3061.

[30] Y. Huang, Y. Cheng, A. Bapna, O. Firat, D. Chen, M. Chen, H. Lee, J. Ngiam, Q. V. Le, Y. Wu, and z. Chen, "GPipe: Efficient Training of Giant Neural Networks using Pipeline Parallelism," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 103–112.

[31] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," 2019, arXiv:1905.11946v3.

[32] M. Abadi, A. Agarwal, P. Barham *et al.*, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems," 2015.