Towards Black-Box Explainability with Gaussian Discriminant Knowledge Distillation

Anselm Haselhoff[1], Jan Kronenberger[1], Fabian Küppers[1], Jonas Schneider[2]
{anselm.haselhoff, jan.kronenberger, fabian.kueppers}@hs-ruhrwest.de
jonas.schneider@efs-auto.com

[1]Ruhr West University of Applied Sciences
Bottrop, Germany

[2]Elektronische Fahrwerksysteme GmbH
Gaimersheim, Germany

Abstract

In this paper, we propose a method for post-hoc explainability of black-box models. The key component of the semantic and quantitative local explanation is a knowledge distillation (KD) process which is used to mimic the teacher's behavior by means of an explainable generative model. Therefore, we introduce a Concept Probability Density Encoder (CPDE) in conjunction with a Gaussian Discriminant Decoder (GDD) to describe the contribution of high-level concepts (e.g. object parts, color, shape). We argue that our objective function encourages both, an explanation given by a set of likelihood ratios and a measure to describe how far the explainer deviates from the training data distribution of the concepts. The method can leverage any pre-trained concept classifier that admits concept scores (e.g. logits) or probabilities. We demonstrate the effectiveness of the proposed method in the context of object detection utilizing the DensePose dataset.

1. Introduction

It is a well known issue that modern neural networks exhibit a lack of explainability and interpretability in their predictions. A neural network serves as a mapping from high dimensional feature space (e.g. an image) to a commonly low dimensional output describing its decisions. However, it is still not possible to get meaningful insights into such a mapping allowing an interpretation of the network predictions. Nevertheless, the ability to explain neural network predictions is of particular interest as it is mandatory to understand decisions of autonomous systems such as driver assistance systems and to gain trust. In recent work, several approaches like gradient-based methods [2, 20, 15] or perturbation-based methods [12, 5] aim to gain insights into the neural network's decision-making. Most of these approaches visualize the relevant image regions as saliency maps. However, recent work has claimed that these maps only indicate omitted input regions and do not provide relevant semantic explanations for the decision [13].

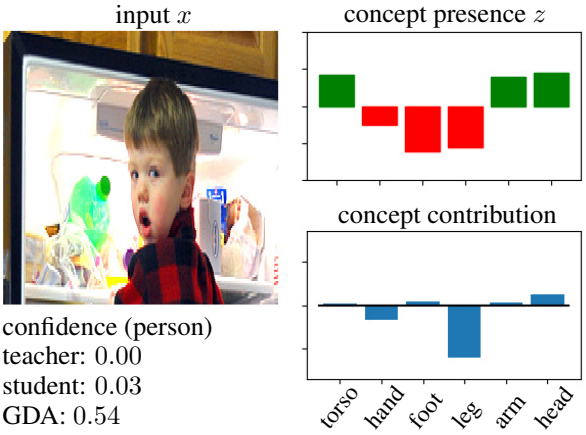

Figure 1: Semantic and quantitative explanation for the teacher's decision [18]: The input image (left) contains a partially occluded person which leads to a low confidence score ($p_t(\mathrm{y} = 1|x) \approx 0.0$). Our generative student model ($p_s(\mathrm{y} = 1|z) \approx 0.03$) can closely reconstruct the teacher's confidence based on the presence of concepts (torso, hand, etc.). The presence or absence of concepts is determined by a concept model and the corresponding logits (top right, semantic explanation). Finally, the quantitative explanation (bottom right) is obtained by the contribution of each concept (log likelihood ratio) to the student's confidence. A parallel GDA model confirms the successful knowledge distillation.

Another branch of research focuses on methods that try to explain and verify neural network predictions by the presence or absence of semantic concepts, i.e. body parts like arms or legs for persons [9, 4, 17]. These approaches provide meaningful explanations but also require modifications

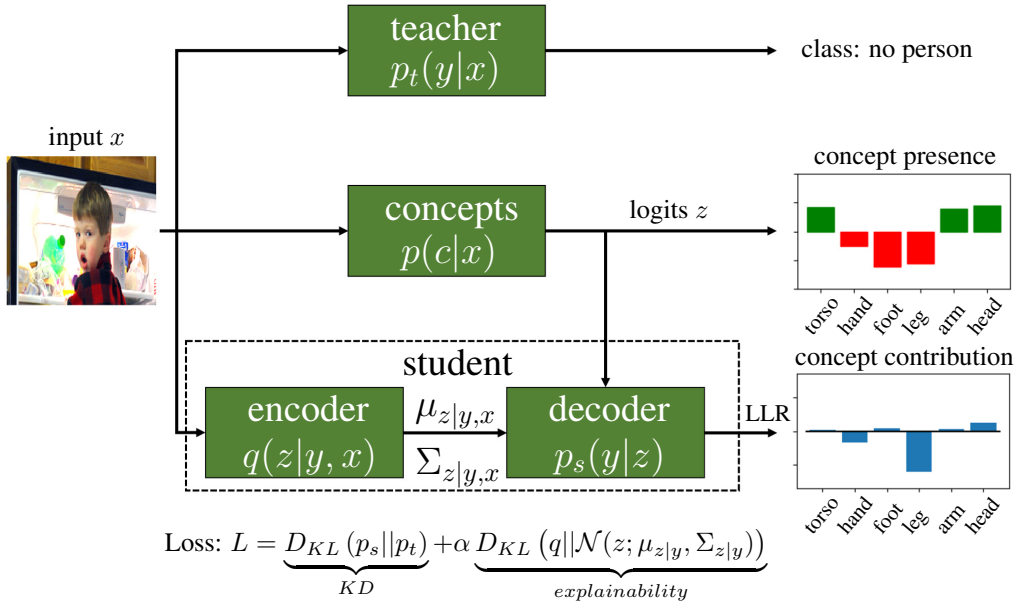

Figure 2: Overview of our approach. The *teacher* represents the network that should be explained, in this case a DNN predicting the class *no person* given the input image x. The student network consists of an encoder and a decoder block and learns to reproduce the predictions of the teacher while internally using the concept presence to explain its prediction. On the one hand, the student is trained to mimic the behavior of the teacher (KD). On the other hand, the student is also trained to match the moments of a fixed concept logit distribution (explainability). Finally, the output of the student is used to calculate a log likelihood ratio that in turn is used to evaluate the contribution of each concept to the decision of the teacher.

to the base network which in turn might also affect its performance. Therefore, the focus of this work is to develop a method that doesn't require a modification of the base network and delivers semantic and quantitative explanations based on a concept model (cp. Figure 1). We follow the strategy presented in [3] and use a student-teacher approach to add post-hoc explainability to a DNN. In contrast to [3], we introduce a new generative encoder-decoder model to determine the contribution of each visual concept to the student's and teacher's decision, respectively. Therefore, we exchange the crucial weighting schema of [3] by a probabilistic model with a built-in interpretation of the concept contribution.

Contributions: To summarize, (i) we present to our knowledge the first generative knowledge distillation (KD) approach to explainability that can be applied to black-box models and thus do not affect their performance; (ii) we propose an encoder-decoder structure, which utilizes a concept model in the "latent" space to obtain semantic and quantitative explanations. The corresponding loss controls the trade-off between KD and explainability; (iii) we define a measure to evaluate how far the explainer has to deviate from the training data distribution to perform a prediction; (iv) we investigate the viability of this measure and analyze the explanations on the DensePose [6] dataset.

2. Related Work

In the scope of image processing, gradient-based visualization methods like LRP [2], Deconvolution [20] or DeepLIFT [15] attempt to approximate explanations by using backpropagation. Therefore, the gradient of each input pixel is evaluated with regard to the output signal. The gradient reflects the importance of the according pixel to the output and thus allows for measuring the relevance of image regions. In contrast, GradCAM [14] only uses the layer activations of the last layer to visualize relevant feature regions. However, these approaches require accessing the intermediate layer activations and are not agnostic to the model architecture. Another way of obtaining neural network explanations is to perturbate image regions and to measure the influence to the output prediction [5]. Using this approach, it is possible to highlight the most relevant input regions for the network output while being independent to the underlying network architecture. A similar approach is given by LIME [12] where an input image is perturbed around its neighborhood to highlight the most relevant regions. An explanation can also be given by generating a text-output based on an inner feature layer [16]. Combining both approaches yields a text-based description of images in conjunction with a visualization of important image regions and is demonstrated by [19].

Although all of these approaches already achieve good performance, it is important to not only use quantitative local explanations, but also to obtain semantic explanations in terms of simple and human understandable concepts like shape, color or parts that constitute an object. The authors in [8] use a vector representation of the logit layer as concepts and compare them to random counter examples. The importance of each concept, regarding a tested class, is measured by its change when evaluated for another class. Higher changes lead to a greater importance of this concept. In contrast, the authors of [9] attach concept classifiers to the inner layers of the network to predict the presence or absence of the concepts, whereas [4] use invertable layers to map the features to a latent space that in turn represents the concepts. Due to the invertability, the latent space can be mapped back to the feature space. Further, [17] propose feature attribution layers in conjunction with input occlusions to determine relevant concepts as well as their importance to the network decision. However, all of these approaches require modifications of the base network. In some cases, this is not feasible or might also affect performance. In this work, we use a student-teacher framework that does not need any modification of the base network and is also independent of its architecture.

Our approach is based on the work of [3] where a student-teacher approach is used to predict the contribution of concepts to a teacher's decision. The contributions are obtained by an additional explainer network. We also utilize a pre-trained concept model and perform knowledge-distillation. The explainer network of [3] consists of a weighted linear combination of the concepts denoting the contribution of each concept to the final decision. The weights are in turn determined by an additional neural network. However, we argue that these weights are only restricted by a weight prior at the beginning of the training process and thus can be freely chosen by the network. In addition, a method to choose either the magnitude of the weights or their ratio for different concepts is missing. Therefore, it is a challenging task to choose and interpret the a-priori weights. In contrast, our probabilistic approach is characterized by a built-in method to determine "a-priori weights" based on the training data. These weights are given as distribution and are used during the whole training process to trade-off knowledge distillation and explainability. Finally, our approach uses a Bayes Classifier and could benefit from the current research in the field of symbolic explanations, the so-called PI-explanations of Naive Bayes Classifiers (NBCs) [10].

3. Knowledge Distillation

Let $p_t(y|x)$ denote the teacher model that is the subject to be explained, y the target class, and x the input data. $p_s(y|z)$ is the decoder of an explainable generative

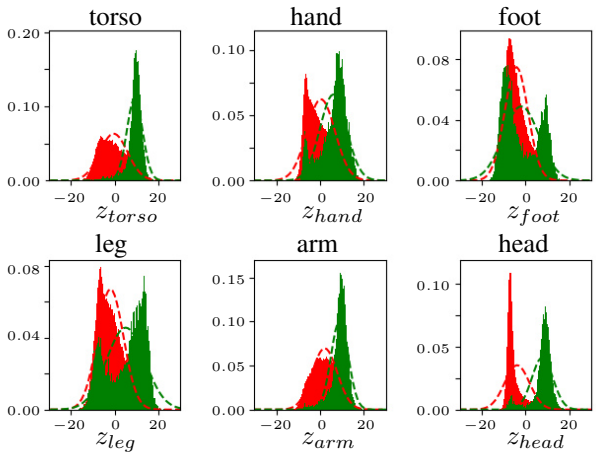

Figure 3: Distributions of concept model logits $p(z|y)$ when predicting the proposal dataset $\mathcal{P}$. Green denotes the distributions for positive samples (person), while red denotes the negative ones.

student model. Our proposed method aims to train the student model, that is, learn the parameters of the student to mimic the behavoir of the teacher model utilizing visual concepts c. We follow the strategy from [3] and use the concept model $p(c|x)$ to predict n visual concepts as semantic explanations. The pre-trained concept model encodes e.g. object part, color, and shape information. The n-dimensional output of the concept model serves as the input feature vector for the generative explainer; more specifically we use the logits z as features.

The task of the generative model is to reconstruct the teacher's output by only combining the visual concepts z in terms of a special Gaussian Discriminat Analysis (GDA) model. The benefit of using the GDA is that we are able to sample from the learned Gaussians and we can calculate the contribution of each visual concept to the class prediction. This approach is closely related to a supervised VAE. In contrast to a VAE, the objective is not to approach a Gaussian $\mathcal{N}(0, I)$ but to match the conditional training data distribution of the concept logits $p(z|y) = \mathcal{N}(z; \mu_{z|y}, \Sigma_{z|y})$. In addition, the decoder doesn't reconstruct the quantitative input variable but the categorical output of the teacher.

Our student model consists of two building blocks, a Gaussian Discriminant Decoder (GDD) and a Concept Probability Density Encoder (CPDE). The details are presented in section 3.2 and 3.3. The overall structure of the approach is visualized in Figure 2.

3.1. Concept Model

The concept model $p(c|x)$ predicts the presence of each concept for a given input x. We start with a pre-trained ResNet50 [7] as base network and add a set of n classifica-

tion heads. For the categorical outputs we use a sigmoid activation to predict the presence or absence of the concepts. The training process of the concept model is independent of the knowledge distillation process and the model can be exchanged by any pre-trained model that predicts reasonable concepts for the given application. Once the concept model is trained, it is used to create the conditional training data distribution of the concept logits $p(z|y)$ based on the student-teacher training data (cp. Figure 3). An example for the logits of a single input sample is shown in Figure 2 (concept presence).

3.2. Gaussian Discriminant Decoder (GDD)

The structure of the decoder is based on a GDA model and has no directly trainable parameters in the context of student-teacher learning. It serves as a proxy to convert the encoder prediction into an interpretable classification output. Based on the concept model and the training data, we can determine the class conditional mean $\mu_{z|y}$ and covariance $\Sigma_{z|y}$ of the corresponding concept logits. The likelihood $p(z|y)$ and prior $p(y)$ alone are sufficient to deploy an auxiliary GDA classifier that can be used in parallel to the teacher model. For inferring the class in terms of a maximum a posteriori probability (MAP) estimate the Bayes' theorem can be applied

$$p(\mathrm{y}=i|z) = \frac{p(z|\mathrm{y}=i)p(\mathrm{y}=i)}{\sum_k p(z|\mathrm{y}=k)p(\mathrm{y}=k)}. \tag{1}$$

To conveniently extract the concept contributions, we assume conditional independence and factorize the concept likelihood $p(z|y) = \prod_{j=1}^{n} p(z_j|y)$. In practice, this leads to a diagonal covariance matrix of the form $\Sigma_{z|y} = \mathrm{diag}\left(\sigma^2_{z_1|y}, \ldots, \sigma^2_{z_n|y}\right)$. Instead of Equation 1, we use the corresponding quadratic decision function $\log p(\mathrm{y}=i|z)$, neglect the normalizer and utilize a softmax to define our generative student model $p_s(y|z)$. Expanding the decision function for each class, we get

$$f_y(z) = z^T W_y z + w_y^T z + b_y, \tag{2}$$

with

$$\begin{aligned} W_y &= -\frac{1}{2}\Sigma^{-1}_{z|y}, \\ w_y &= \Sigma^{-1}_{z|y}\mu_{z|y}, \\ b_y &= -\frac{1}{2}\mu^T_{z|y}\Sigma^{-1}_{z|y}\mu_{z|y} - \frac{1}{2}\log\det\Sigma_{z|y} + \log p(\mathrm{y}=y). \end{aligned}$$

Equation 2 constitutes the decoder of our student model and is a function of the concept logits while it is parameterized by the mean, covariance and prior (cp. Figure 2). This part of our student model is called Gaussian Discriminant Decoder (GDD). However, how do we obtain an explanation? For simplicity, we now assume a 2-class problem and the log-odds ratio of Equation 1 leads to

$$\mathrm{logit}\left(p(\mathrm{y}=1|z)\right) = \sum_{j=1}^{n} \log\frac{p(z_j|\mathrm{y}=1)}{p(z_j|\mathrm{y}=0)} + \log\frac{p(\mathrm{y}=1)}{p(\mathrm{y}=0)}.$$

Due to the factorized likelihood, the contribution of each concept is given by the log likelihood ratio for the individual concepts $LLR_j = \log\frac{p(z_j|\mathrm{y}=1)}{p(z_j|\mathrm{y}=0)}$. Positive values basically confirm the decision for the target class $\mathrm{y}=1$ and negative values contradict the decision. An example is shown in Figure 2 (concept contribution). At this point, we have an explainable model that is independent of the teacher and hence, the explanation for the teacher's decision is still missing. To get around this hurdle, we incorporate an encoder model into the student to facilitate KD by adjusting the parameters of the concept logit distribution.

3.3. Concept Probability Density Encoder (CPDE)

The encoder is the only part of the student that has learnable parameters and therefore the encoder is responsible for adjusting the output of our student's GDD $p_s(y|z)$ to be aligned with the teacher model $p_t(y|x)$. More importantly, this adjustment forces our decoder to explain the decisions of the teacher since it has to reconstruct the decision by a set of interpretable concepts. The goal of the concept probability density encoder (CPDE) is to predict the class conditional mean and covariance used by the GDD. In order to equip the student with the necessary degree of freedom, the student's distribution $q(z|y,x)$ now depends on the input data x, that is, we predict the mean $\mu_{z|y,x}$ and covariance $\Sigma_{z|y,x}$. The encoder itself is again based on a ResNet50 [7] with softplus activations to guarantee positive variance predictions. We only train the last fully connected layer.

The decoder, as described in subsection 3.2, is an auxiliary classifier that explains its decision with reference to the training data distribution $p(z|y)$. If we exchange $p(z|y)$ with the input dependent distribution $q(z|y,x)$ of the encoder, we obtain a regular student-teacher setup. The additional degree of freedom, that is incorporated into the encoder, allows the student to completely "neglect" the presence or absence of concepts to follow his goal and approximate the teacher's output as close as possible. It is worth mentioning that this is a worst case scenario where the concepts are incomplete or insufficient to mimic the teacher's behavior. Thus, we have the two extremes: 1.) an explainable GDD model which is independent of the teacher and 2.) a student-teacher framework which can freely choose the mean and covariance to describe the concept contributions. The inevitable trade-off between the explainability and the knowledge distillation is obtained by combining both objectives into the loss function.

3.4. Loss Function

Our objective function encourages both, an explanation given by a set of concept likelihood ratios (explanation) and a knowledge distillation (KD). The KD forces the output of the student to mimic the teacher whereas the regularizer encourages the student's distribution to mimic those of an explainable GDA by minimizing

$$L = \underbrace{D_{KL}\left(p_s||p_t\right)}_{KD} + \alpha \underbrace{D_{KL}\left(q||\mathcal{N}(z;\mu_{z|y},\Sigma_{z|y})\right)}_{explainability}, \quad (3)$$

with the Kullback-Leibler divergence $D_{KL}\left(p_s||p_t\right)$ between decoder p_s and teacher distribution p_t. $D_{KL}\left(q||\mathcal{N}(z;\mu_{z|y},\Sigma_{z|y})\right)$ denotes the divergence between the encoder distribution $q(z|y,x)$ and training data distribution $p(z|y)$. The hyper parameter α controls the trade-off between knowledge distillation and explainability.

4. Experiments

We use our approach to monitor the decisions made by a person detection model. To train the components of our proposed method, we use different subsets that have been generated by the DensePose dataset [6]. Furthermore, we use the six concepts torso, hand, foot, leg, arm and head to explain the network predictions of a pretrained Faster RCNN R-50 [11] provided by Detectron2 [18]. We further introduce the concept dataset $\mathcal{C}$ consisting of crops from DensePose dataset with corresponding concept labels. To not confuse the concepts due to concurrent presence (e.g. arm and torso are often visible at the same time), we especially cropped single concepts. A concept is counted as present if the ground-truth part-segmentation provided by DensePose covers at least 5 % of the image. In total, we generated 107.864 concept samples. The proposal dataset $\mathcal{P}$ ($>$ 247.000 samples) represents the predictions made by the teacher model. This dataset can either be created by a classification network or, in our case, by the predictions/proposals obtained by an object detection network. To convert the detections into a classification dataset, each predicted bounding box is cropped and saved with the corresponding confidence score. This information defines our teacher model. In Figure 4, we provide some cropped samples of $\mathcal{P}$ with the corresponding confidence scores. Note that the samples are cropped after the non-maximum suppression as we treat the network as black-box.

4.1. Concept Model

In a first step, we trained a ResNet50 on the concept dataset $\mathcal{C}$ that is independent of the person detection network to predict the presence of concepts in an image. The results are shown in Table 1. Concepts close to the torso often appear inside a cropped image of the training data, whereas the concept hand is absent in many cases. This leads to a slightly inferior performance compared to the other concepts.

Figure 4: Samples of the proposal dataset $\mathcal{P}$ with the corresponding ground-truth body-part segmentation of the concepts as well as the predicted person confidence.

Table 1: Results (%) of the concept model on $\mathcal{C}$.

Concept	Accuracy	Precision	Recall	F_1
torso	91.1	94.2	93.0	93.6
hand	88.3	91.5	83.6	87.4
foot	94.2	82.4	81.1	81.7
leg	89.0	85.8	85.5	85.6
arm	88.4	91.1	90.6	90.8
head	94.6	96.8	92.1	94.4

4.2. Knowledge Distillation Performance

In this section, we evaluate the performance of our knowledge distillation process. The ability of the student to mimic the behavior of the teacher is measured by the similarity of the predicted probability scores. Therefore, we utilize the Pearson correlation coefficient to assess a linear relationship between student's and teacher's confidence whereas the Spearman's rank correlation coefficient captures a monotonic relationship. A baseline is obtained by using a standard GDA model which in turn is exactly our GDD model with the distribution parameters of the regularizer $p(z|y)$. The distribution $p(z|y)$ is extracted from the proposal dataset $\mathcal{P}$ (Figure 3) and new samples can be classified by the GDA model using the logit predictions from the concept model. It is obvious that the baseline GDA model shouldn't be highly correlated with the teacher since the input features are different, the GDA is of limited capacity and the GDA is a parallel model without insights into the teacher. Even if the decoder (GDD model) defines a quadratic decision function, the added encoder (CPDE) is a neural network and therefore induces a highly nonlinear decision function. This degree of freedom is nec-

essary to enable knowledge distillation. The high correlation values of our student (CPDE & GDD) compared to the standard GDA shown in Table 2 confirm the successful knowledge distillation process. We performed a grid search ($\alpha \in \{0.1, \ldots, 1.0\}$) on the training data to determine an appropriate α. As a measure to choose an α we analyzed the trade-off between student-teacher and student-GDA correlation, but all analyzed values induce similar performance in terms of explainability. Therefore, we have chosen $\alpha = 0.1$ with highest student-teacher correlation.

Table 2: Pearson and Spearman correlation coefficients between teacher confidence and the output of our student (CPDE & GDD) trained with $\alpha = 0.1$. A GDA model is used as a baseline. Correlation relevance: *p-value* ≤ 0.001.

	Pearson r	Spearman ρ
GDA baseline	.69	.51
CPDE & GDD	**.81**	**.64**

In addition, we use a standard evaluation protocol (accuracy, precision, recall, F_1 score) to demonstrate the effectiveness of the student-teacher learning. The models are evaluated using intersection over union (IoU) scores of .5 and .75 to define a matching ground-truth bounding box. Although the GDA has a lower correlation w.r.t. the teacher, around 92 % of all samples can be classified correctly (cp. Table 3). Therefore, the GDA could be helpful to discover corner cases where the GDA and the teacher disagree. On the other hand, the student is characterized by a higher correlation and a classification performance that almost matches the teacher's performance.

Table 3: Classification results for IoU .5 (top) and .75 (bottom) compared to the ground-truth data on the test dataset.

	Accuracy	Precision	Recall	F_1
Teacher	95.83	85.74	69.56	76.81
GDA baseline	92.01	57.68	**73.02**	64.45
CDPE & GDD	**94.48**	**78.54**	61.08	**68.72**

	Accuracy	Precision	Recall	F_1
Teacher	97.78	75.31	96.32	84.53
GDA baseline	91.87	42.67	**85.15**	56.86
CDPE & GDD	**96.63**	**67.04**	82.18	**73.84**

4.3. Explanation

Insights into the Concept Knowledge: Compared to the standard GDA, the student has a higher performance and correlation (cp. subsection 4.2). This is achieved by the student's ability to adjust the distribution $q(z|y, x)$ based on the input data x. An interesting insight into the student's knowledge is provided by comparing the GDA distribution $p(z|y)$ and $q(z|y, x)$. In Figure 5, the deviation of the student's mean and covariance w.r.t. to the GDA is visualized. Surprisingly, the deviation of the student model from the GDA is marginal for most concepts, except for the concepts foot and leg. By comparing the deviation-magnitude of $\mu_{z|y,x}$ with the training data distribution from Figure 3, we observe that the student model has learned to cope with the bimodal nature of the concept distribution of foot and leg, i.e., the distribution $p(z|y = 1)$ (green) of the concept foot has two modes (approx. $z_{foot} \in \{-9, 10\}$) and the GDA utilizes the mean $\mu_{z_{foot}|y=1} \approx -2$. Figure 5 reveals that the student has learned to shift this mean in the range

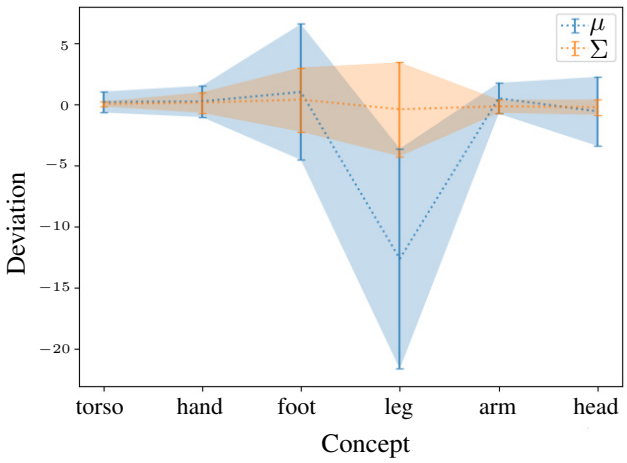

Figure 5: Comparison of the GDA baseline (regularizer) and the CPDE. For each concept, the deviation of the predicted mean $\mu_{z|y,x}$ and covariance $\Sigma_{z|y,x}$ of our encoder is evaluated w.r.t. the GDA distribution $p(z|y)$. Thus, we get an insight into the student's behavior and how far he has to shift the mean and covariance of each concept to mimic the teacher's behavior. It is obvious that the student's strong deviation for the concepts leg and foot are necessary to capture the bimodal nature of these concepts (cp. Figure 3).

Table 4: Spearman correlation coefficients ρ between teacher confidence and concept LLR. Correlation relevance: *p-value* ≤ 0.001. (except concept foot of GDA baseline).

	GDA baseline	CDPE & GDD
torso	.48	**.54**
hand	.18	**.28**
foot	−.02	**.17**
leg	.16	**.32**
arm	.37	**.40**
head	.35	**.53**

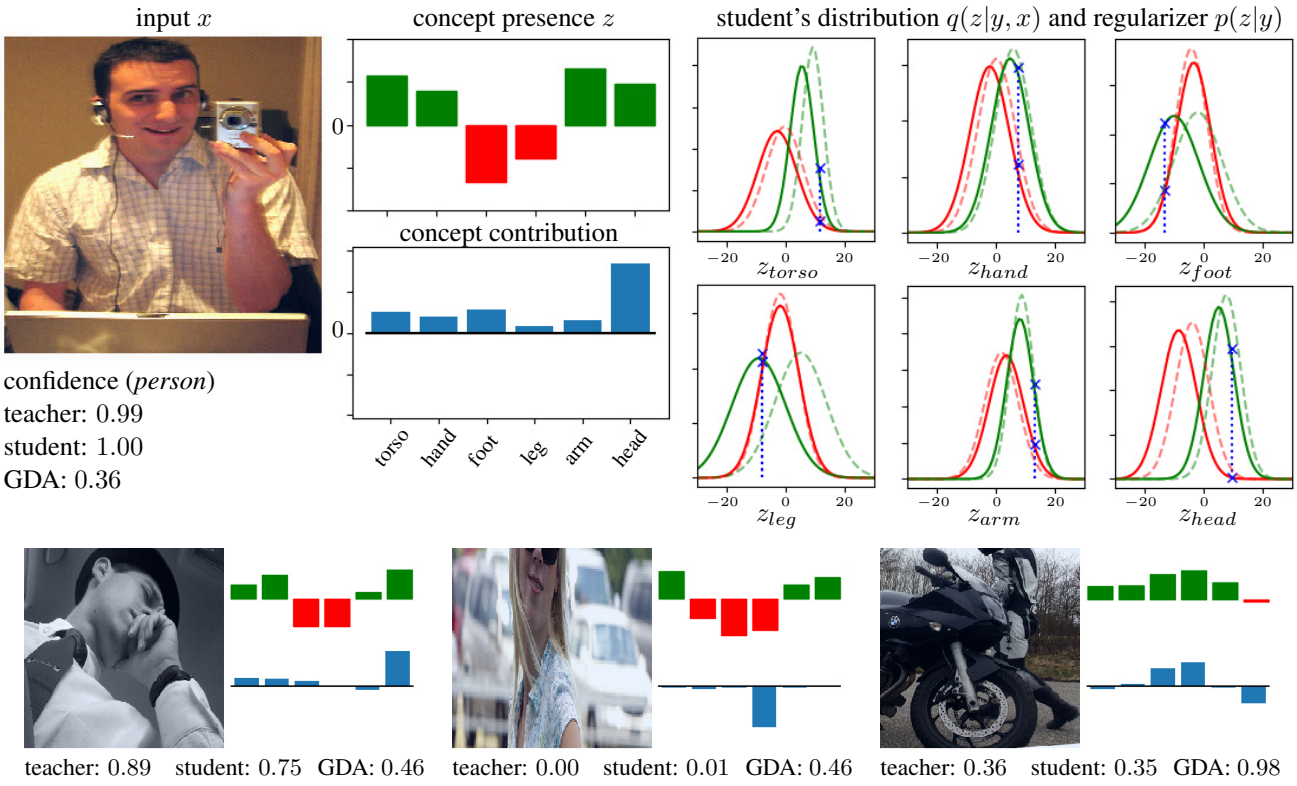

Figure 6: Top: An example of explanations with the input image, the logits z from the concept model (concept presence) and the log likelihood ratios (concept contribution) for the positive class (*person*). The confidence of the student and teacher is very similar, while the GDA model delivers a contradictorily confidence value. On the right the predicted encoder distribution $q(z|y,x)$ (solid line) and the regularizer $p(z|y)$ (dashed line) are visualized. The distribution for the positive class $y = 1$ is shown in green and in red for the negative class. Bottom: Additional example images with corresponding concept explanations obtained by our student model.

$[-5, 6]$. Therefore, the student can dynamically take advantage of the two modes to improve its performance and mimic the teacher using the approximation of the modes $\hat{z}_{foot} \in \{-7, 4\}$. A similar behavior can be observed for the concept leg. A visual example of the shifted mean for the concepts foot and leg is shown in Figure 6 (top right). To summarize, due to the regularizer (explainability), most of the concept explanations are very similar to the GDA baseline and small deviations are sufficient to mimic the teacher's behavior. In addition, the student can leverage the bimodal nature of concept distributions to improve KD even further.

Quantitative Evaluation: To evaluate the explanation, we follow the strategy presented in [1] and use the Spearman rank correlation metrics to measure the "similarity" between concepts and the teacher's decision in terms of a monotonic relation. The correlation is calculated between the log likelihood ratio LLR of each concept and the confidence of the teacher. The LLR represents the contribution of each concept to the prediction of the teacher. The correlation coefficients are again compared to the standard GDA model without any KD. Table 4 reveals that our encode-decoder (CPDE & GDD) framework has improved the correlation by a large margin.

Qualitative Examples: Qualitative examples of semantic and quantitative local explanations are visualized in Figure 6. We observe that our encoder changes the distributions slightly while keeping them in the same range as the distributions of the GDA baseline model. Furthermore, the confidence scores of the teacher, student and GDA model confirm an appropriate knowledge distillation. The GDA model delivers contradictorily confidence values, whereas the student and teacher show a similar behavior.

5. Conclusion

In this paper, we present a method to add post-hoc explainability to black-box models. We use visual concepts to obtain semantic and quantitative explanations and propose a new generative student-teacher framework that learns to mimic the base network. The trade-off between knowledge distillation and explainability is ensured by our objective

function which utilizes a regularizer capturing the information stored in a standard GDA model. In order to quantify the contribution of different semantic visual concepts, the concept distribution is factorized to provide separate log likelihood ratios. These ratios reveal the inner-workings of our teacher model. The viability of our approach is evaluated on the DensePose dataset. We show in our experiments that our student model is able to mimic the teacher model with high correlation while providing meaningful concept explanations for each prediction. Therefore, we conclude that our framework implements a reasonable knowledge distillation while enhancing explainability without sacrificing the teacher's performance. This framework is a good contribution towards explainable and interpretable neural networks, allowing for a potential usage even in safety critical applications like autonomous driving.

Acknowledgement

The authors gratefully acknowledge support of this work by Elektronische Fahrwerksysteme GmbH, Gaimersheim, Germany. The research leading to the results presented above are funded by the German Federal Ministry for Economic Affairs and Energy within the project "KI Absicherung – Safe AI for automated driving".

References

[1] J. Adebayo, M. Muelly, I. Liccardi, and B. Kim. Debugging tests for model explanations. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 700–712. Curran Associates, Inc., 2020. 7

[2] S. Bach, A. Binder, G. Montavon, F. Klauschen, K. Müller, and W. Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015. 1, 2

[3] Runjin Chen, Hao Chen, Jie Ren, Ge Huang, and Quanshi Zhang. Explaining neural networks semantically and quantitatively. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 2, 3

[4] P. Esser, R. Rombach, and B. Ommer. A disentangling invertible interpretation network for explaining latent representations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 3

[5] R. C Fong and A. Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3429–3437, 2017. 1, 2

[6] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018. 2, 5

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 3, 4

[8] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018. 3

[9] Jan Kronenberger and Anselm Haselhoff. Dependency Decomposition and a Reject Option for Explainable Models. In *2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019. 1, 3

[10] J. Marques-Silva, T. Gerspacher, M. Cooper, A. Ignatiev, and N. Narodytska. Explaining naive bayes and other linear classifiers with polynomial time and delay. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 20590–20600. Curran Associates, Inc., 2020. 3

[11] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 91–99, 2015. 5

[12] M. Ribeiro, S. Singh, and C. Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016. 1, 2

[13] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215, 05 2019. 1

[14] R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 2

[15] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje. Not just a black box: Learning important features through propagating activation differences. *CoRR*, 2016. 1, 2

[16] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015. 2

[17] Weibin Wu, Yuxin Su, Xixian Chen, Shenglin Zhao, Irwin King, Michael R. Lyu, and Yu-Wing Tai. Towards global explanations of convolutional neural networks with concept attribution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 3

[18] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 1, 5

[19] 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*, pages 2048–2057. PMLR, 2015. 2

[20] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 1, 2