Uncertainty-sensitive Activity Recognition: A Reliability Benchmark and the CARING Models

Alina Roitberg Monica Haurilet Manuel Martinez Rainer Stiefelhagen

Institute for Anthropomatics and Robotics
Karlsruhe Institute of Technology, Germany
{firstname.lastname}@kit.edu

***Abstract*—Beyond assigning the correct class, an activity recognition model should also to be able to determine, how certain it is in its predictions. We present the first study of how well the confidence values of modern action recognition architectures indeed reflect the probability of the correct outcome and propose a learning-based approach for improving it. First, we extend two popular action recognition datasets with a *reliability* benchmark in form of the expected calibration error and reliability diagrams. Since our evaluation highlights that confidence values of standard action recognition architectures do not represent the uncertainty well, we introduce a new approach which learns to transform the model output into realistic confidence estimates through an additional calibration network. The main idea of our Calibrated Action Recognition with Input Guidance (CARING) model is to learn an optimal scaling parameter *depending on the video representation*. We compare our model with the native action recognition networks and the temperature scaling approach - a wide spread calibration method utilized in image classification. While temperature scaling alone drastically improves the reliability of the confidence values, our CARING method consistently leads to the best uncertainty estimates in all benchmark settings.**

I. Introduction

Humans have a natural grasp of probabilities [5]: If we hear that a certain event is detected in a video by a neural network with 99% confidence, we automatically assume this to be the case. Such assumption however would be naive, as the inference merely gives us values of the last fully-connected layer which are usually optimized for a high top-1 accuracy on a fixed set of previously defined categories. As these values are usually normalized through the Softmax function to sum up to one, they *appear* to be class probabilities but they do not depict the true confidence of the model [6]. Besides, when engineers apply such deep learning models in practice, they will quickly discover the phenomenon of *model miscalibration* i.e. the resulting Softmax scores tend to be biased towards very high values [6], [8]. Unfortunately, such high confidence values are not only present in correct predictions but also in case of misclassifications. Despite impressive results in conventional classification, such overly self-confident models become a burden in applications, and might lead to tragic outcomes if assessing model uncertainty in its prediction plays an important role. Apart from the direct benefits of proper confidence values for decision-making systems, good assessment of uncertainty enhances model interpretability. For example, in the realistic scenario of open-world recognition, low-confidence input might be passed to human experts, which would provide the correct annotations (i.e.active learning) and therefore improve the decision boundary.

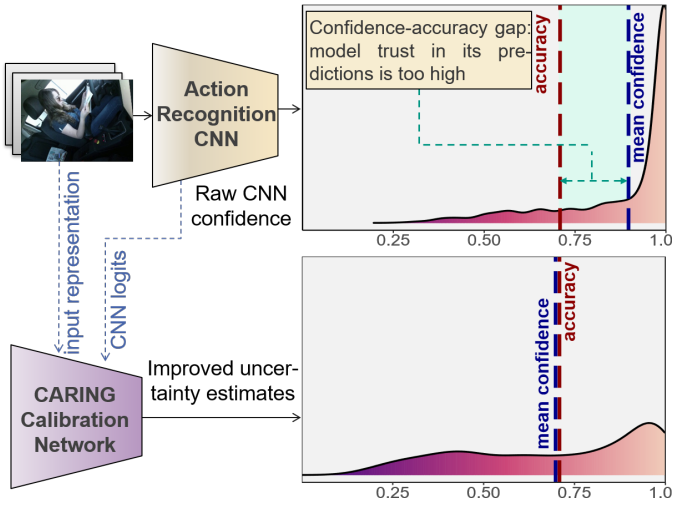


Fig. 1: Softmax confidence distribution of a popular video classification network (Pseudo 3D ResNet tested on a Drive&Act validation split) before and after the improvement through our **C**alibrated **A**ction **R**ecognition with **In**put **G**uidance model. Native confidence values underestimate model uncertainty (the majority of samples was rated with $> 90\%$ confidence, while the accuracy is significantly lower). We propose to incorporate the *reliability* of model confidence in the evaluation of activity recognition models and develop algorithms for improving it.

Uncertainty-aware models are vital for safety-critical applications of *activity recognition* approaches, which range from robotics and manufacturing [26] to autonomous driving [7] and surveillance. While obtaining well-calibrated probability estimates is a growing area in general image recognition [8], [10], this performance aspect did not yet receive any attention in the field of video classification. The impressive progress reported on the conventional action recognition benchmarks linked to the rise of deep learning [2], [9], [24] may therefore

draw a rather idealistic picture, as their validation is often limited to the top-1 accuracy on a static set of carefully designed actions [2], [13], [17]. While such neural networks are notably bad at detecting data ambiguities, examining how well the confidence values of activity recognition models indeed reflect the probability of a correct prediction has been overlooked in the past and is the main motivation of our work.

In this paper, we aim to elevate the role of uncertainty in the field of activity recognition and develop models, which do not only select the correct behavior class but are also able to *identify misclassifications*. In other words, the resulting probability value should indeed reflect the likelihood of the prediction to be correct. To this intent, we propose to incorporate the *reliability* of model confidence in the evaluation of activity recognition models and develop methods which transform oftentimes biased confidence outputs of the native action recognition models into reliable probability estimates.

Contributions and Summary We argue, that for applications in industrial systems, activity recognition models must not only be accurate, but should also asses, how likely they are to be correct in their prediction through realistic confidence values. This paper makes the first step towards activity recognition models capable of *identifying their misclassifications* and has thee major contributions. (1) We present the first study of how well the confidence of the modern activity recognition architectures indeed reflects the likelihood of a prediction being correct. To this intent, we incorporate the Expected Calibration Error metric in the evaluation procedure of two action recognition CNNs: Pseudo 3D ResNet (P3D) [24] and Inflated 3D ConvNet (I3D) [2]. Our experiments on two action recognition datasets confirm, that the out-of-the-box probability values of such models do not reflect model uncertainty well (e.g. over 20% expected calibration error on HMDB-51 [13]). (2) We further aim for a framework which learns to transform the poorly calibrated confidence values of the native action recognition models into more realistic probability estimates. We enhance these architecture with the temperature scaling method [8], a prominent approach for model calibration in image recognition, which learns a single temperature parameter T used to scale the network logits. This method, however, learns *one global temperature value* for scaling, i.e. after calibration, the logit values are always divided by the same scalar. (3) We believe, that input representation gives us significant cues for quantifying network uncertainty, and present a new method for **C**alibrated **A**ction **R**ecognition with ***In**put **G**uidance* (CARING). In contrast to [8], CARING entails an additional calibration network, which takes as input intermediate representations of the observed activity and learns to produce *temperature values specific to this input*. While temperature scaling alone drastically improves the confidence values (e.g. the expected calibration error for the I3D model drops from 15.97% to 8.55%), our CARING method consistently leads to the best uncertainty estimates in all benchmarks, further reducing the error by 2.53% on Drive&Act.

II. Related Work

A. Activity Recognition

Activity recognition research is strongly influenced by progress in image recognition methods, where the core classification is applied on video frames and extended to deal with the video dimension on top of it. Similar to other computer vision fields, the methods have shifted from manually designed feature descriptors, such as Improved Dense Trajectories (IDT) [31] to Convolutional Neural Networks (CNNs) which learn intermediate representations end-to-end [31]. The first deep learning architecture to outperform IDTs was the two-stream network [27], [32], which comprises 2D CNNs operating on individual frames of color- and optical flow videos. The frame output is joined via late fusion [27], [32] or an additional recurrent neural network [4], [19]. The field further progressed through emergence of 3D CNNs, which leverage spatiotemporal kernels to deal with the time dimension [2], [9], [12], [29], [30]. This type of networks still holds state-of-the-art results in the field of action recognition, with Inflated 3D Network [2], 3D Residual Network [9] and Pseudo 3D ResNet [24] being the most prominent backbone architectures.

The above works develop algorithms with the incentive to improve the top-1 recognition accuracy on standard activity classification benchmarks *without taking the faithfulness of their confidence values into account* (as demonstrated in Figure 1 with an example of Pseudo3D ResNet). Our work focuses on *uncertainty-aware action recognition* and aims for models which confidence values indeed reflect the likelihood of a correct prediction. Note, that the developed methods drastically improve the ability of an action recognition network to assign proper confidence values, they do not affect the accuracy, as they are based on learned scaling of the logits without changing their order.

B. Identifying Model Misclassifications

While multiple authors expressed the need for better uncertainty estimates in order to safely integrate deep CNNs in real-life systems [10], [20], [28], the feasibility of predicted confidence scores has been missed out in the field of activity recognition. However, this problem has been addressed before in image classification [6], [8], person identification [1] and classical machine learning [3], [21], [23]. Some of the uncertainty estimation methods are handled from the Bayesian point of view, leveraging Monte Carlo Dropout sampling [6] or ensemble-based methods [15]. In such methods, the uncertainty is represented as a Gaussian distribution with output being the predictive mean and variance. In contrast, calibration-based approaches [8], [14], [18], [22], [23], [33], [34] have lower computational cost as they do not preform sampling and return a single confidence value. While these works approach the problem in a different way, they are all trained to obtain a proper confidence value on a held-out validation set following the initial training of the model and, thus, might be viewed as postprocessing methods. Recently, multiple calibration-based algorithms, such as isotonic regression, histogram binning,

and Bayesian quantile binning regression and were brought in the context of CNN-based image classification by Guo et al. [8]. The authors introduced temperature scaling, a simple variant of Platt Scaling [23], where a single parameter is learned on a validation set and to rescale the neural network logits. Despite its simplicity, the temperature scaling method has outperformed other approaches in the study by Guo et al. [8] and has since then been successfully applied in natural language processing [14], [22] and medical applications [11].

Several works have studied uncertainty estimation in the context of novelty detection [10], [16], [25]. A Bayesian approach has been used in a framework for recognizing activity classes which were not present during training [25]. Hendrycks and Gimpel have introduced a baseline for out-of-distribution detection using raw Softmax values [10], which was further improved by Liang et al. [16] through input corruptions and temperature scaling [8]. Our work, however, aims to study the confidence activity recognition models to *identify, whether the prediction is correct, or not* and is therefore more comparable to the model calibration benchmarks of [8], [21].

Our model builds on the approach of Guo et al. [8], extending it with *input-guided* scaling. In contrast to [8], which uses a static temperature parameter for all data points, we introduce an additional *calibration network* to estimate a proper scaling parameter depending on the input. Furthermore, our benchmark examines the reliability of model confidence values in context of action recognition for the first time.

III. Uncertainty-sensitive Action Recognition

A. Problem Definition: Reliable Confidence Measures

We introduce the *reliability of model confidence benchmark* to supervised multi-class activity recognition, where the models are usually validated via top-1 accuracy only [13], [17]. Given an input video clip x with a ground-truth label a_{true} and the set of all possible target classes $a \in \mathcal{A}\{1, ..., m\}$, let m be our activity recognition model predicting an activity label a_{pred} and the corresponding model confidence value $conf(a_{pred})$: $m(x) = [a_{pred}, conf(a_{pred})]$. A *reliable* model ought to not only learn to predict the correct activity (i.e. $a_{pred} = a_{true}$), but also give us well-calibrated confidence estimates $conf(a_{pred})$, which indeed reflect the probability of a successful outcome $\mathcal{P}(a_{pred} = a_{true})$. A perfectly calibrated i.e. reliable model is often formalized as $\mathcal{P}(a_{pred} = a_{true}|conf(a_{pred}) = p) = p, \quad \forall p \in [0, 1]$ [8]. In other words, the inadequacy of model confidence values is directly linked to the gap between the average model confidence and model accuracy. To quantify the calibration quality of the models' confidence scores, we use Expected Calibration Error (ECE) metric [8]. To compute ECE, we divide the space $[0, 1]$ of possible probabilities into K segments (in our case, $K = 10$). We then compute the model accuracy and average model confidence for samples belonging to each individual segment. In a perfectly calibrated model, the difference between accuracy and average confidence of the individual segments would be zero. To quantify how well we can rely on the confidence scores produced by the model,

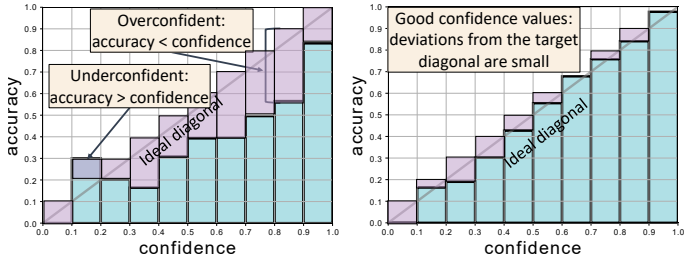


Fig. 2: Reliability diagrams of a model with poor confidence estimates (left) and a well-calibrated model (right). The illustrated data are the confidence values of the Pseudo 3D ResNet a the Drive&Act validation split before and after the improvement with the CARING calibration network.

we compute the distance between the mean confidence and accuracy in each bin and then calculate the average over all such segments, weighted by the number of samples in each bin. Formally, the expected calibration error is defined as:

$$ECE = \sum_{i=1}^{K} \frac{N_{bin_i}}{N_{total}} |acc(bin_i) - conf(bin_i)|, \qquad (1)$$

where N_{bin_i} is the number of samples with probability values inside the bounds of bin_i, $acc(bin_i)$ and $conf(bin_i)$ are the accuracy and average confidence of such examples respectively and N_{total} is the total number of data points (in all bins).

The expected calibration error can be visualized intuitively using *reliability diagrams* (example provided in Figure 2). First, the space of possible probabilities (X-axis) is discretized into K equally sized bins (we choose $K = 10$), as previously described for the ECE calculation. Samples with predicted confidence between 0 and 0.1 fall into the first bin, between 0.1 and 0.2 into the second bin and so on. For each segment, we plot a bar with height corresponding to the accuracy in the current segment. In an ideal case, the accuracy should be equal to the average confidence score inside this bin, meaning, that the bars should have the height of the diagonal. As we see in Figure 2, these are often beyond the diagonal if the Pseudo 3D ResNet model probabilities are used out of the box. This means that the model tends to be overly confident, as the accuracy in the individual bins tends to be *lower* than the probability produced by the model.

B. Backbone Neural Architectures

First, we describe the backbone architectures examined in our study. We consider two prominent spatiotemporal CNNs for activity recognition: Inflated 3D ConvNet [2] and Pseudo3D ResNet [24]. Both architectures directly operate on the video data and learn the intermediate embeddings together with the classifier layers in an end-to-end fashion. Inflated 3D ConvNet deals with the spatial and temporal dimensions of our input by leveraging hierarchically stacked 3D-convolution and -pooling kernels with the size of $3 \times 3 \times 3$. P3D ResNet, on the other hand, mimics 3D convolutions by applying a filter on the spatial domain ($3 \times 3 \times 1$) followed by one in the temporal dimension ($1 \times 1 \times 3$). Furthermore, P3D ResNet

leverages residual connections to improve the gradient flow, which allows a remarkable depth of 152 layers, while Inflated 3D ConvNet is 27 layers deep.

As in other CNNs, the neurons of the last fully-connected layer are referred to as a *logit vector* $\vec{y}$ with its activations y_a representing *not normalized* scores of activity a being the current class. A straight-forward way to obtain the model confidence which mimics a probability function, is to normalize the scores using Softmax: $conf(a_{pred}) = \max_{a \in \mathcal{A}} \frac{exp(y_a)}{\sum_{\hat{a} \in \mathcal{A}} exp(y_{\hat{a}})}$. During training, the cross-entropy loss is computed using the Softmax-normalized output, optimizing the network for high top-1 accuracy. Both architectures have demonstrated impressive results in activity recognition [2], [17], [24], but an evaluation of how well their Softmax-values indeed reflect the model uncertainty remains an open question and is therefore addressed in this work.

C. Calibration via Temperature Scaling

A popular way for obtaining better confidence estimates from CNN logits in image recognition is *temperature scaling* [8]. Temperature scaling simplifies Platt scaling [23], and is based on learning a single parameter τ which is further used to "soften" the model logits. The logits are therefore divided by τ before applying the Softmax function $\vec{y}_{scaled} = \vec{y}/\tau$. With $\tau > 1$ the resulting probabilities become smoother, moving towards $\frac{1}{m}$, where m is the number of classes. Contrary, scaled probability would approach 1 as τ becomes closer to 0. After the neural network is trained for supervised classification in a normal way, we fix the model weights and optimize τ on a held-out validation set using Negative-Log-Likelihood. Despite method simplicity, temperature scaling has been highly effective for obtaining well-calibrated image recognition CNNs, surpassing heavier methods such as Histogram binning and Isotonic Regression [8].

As this method has not been explored for spatiotemporal video classification CNNs yet, we augment the Inflated 3D ConvNet and Pseudo 3D ResNet models with a post-processing temperature scaling module. We optimize τ using Gradient Descent with a learning rate of 0.01 for 50 epochs.

Note, that as the networks are fully trained and their weights remain fixed while learning the scaling parameter τ, transformation of the logits does not influence their order and therefore the *model accuracy stays the same*. In other words, while temperature scaling gives us better uncertainty estimates, the predicted activity class does not change as all logits are divided by the same scalar.

D. Calibrated Action Recognition with Input Guidance

In this section, we introduce a new model for obtaining proper confidence estimates by learning how to scale the logits *depending on the input*. While our evaluation described in the next section reveals, that previous method clearly improves model confidence calibration, it does not take into account representation of the current example, i.e., the logits are always divided by the *same* global scalar τ.

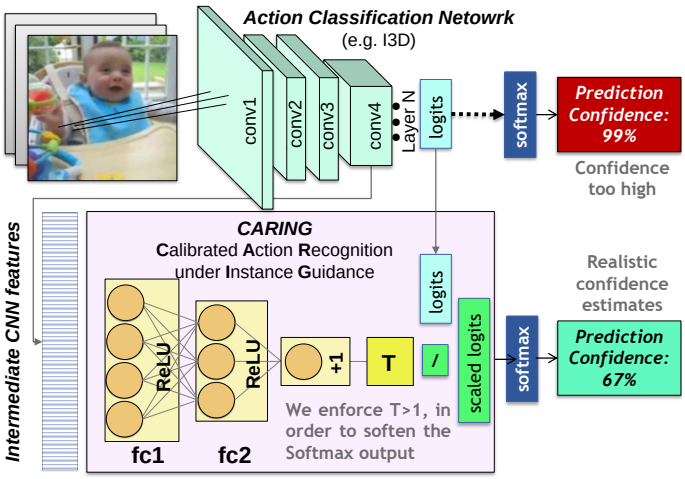


Fig. 3: Overview of the Calibrated Action Recognition under Instance Guidance Model (CARING). CARING is an additional neural network which learns to infer the scaling factor $\mathcal{T}$ depending on the instance representation. The logits of the original activity recognition network are then divided by T, giving better estimates of the model uncertainty.

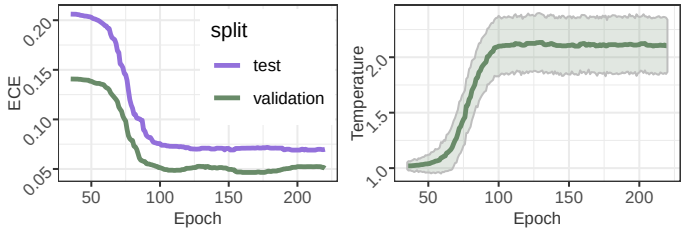
We believe, that the input itself carries useful signal for inferring model confidence and build on the temperature scaling approach [8] with one crucial difference: the scaling factor is not global but different for varying input. Our main idea is therefore to learn acquiring the scaling parameter $\mathcal{T}(\vec{z})$ on-the-fly at test-time depending on the input representation $\vec{z}$, so that the scaled logits become $\vec{y}_{scaled} = \vec{y}/\mathcal{T}(\vec{z})$. To learn the input-dependent temperature value $\mathcal{T}(\vec{z})$, we introduce an additional *calibration neural network*, which we refer to as the CARING model (**C**alibrated **A**ction **R**ecognition under **In**put **G**uidance), as the it guides the scaling of the logits depending on the current instance. An overview of our model is provided in Figure 3. CARING network comprises two fully-connected layers, with the output of the second layer being a single neuron used to infer the input-dependent temperature scalar. Note, that we extend the last ReLU activation with an addition of 1 to enforce $\mathcal{T}(\vec{z}) \geq 1$, required to soften the probability scores. Input-dependent temperature $\mathcal{T}(\vec{z})$ is therefore obtained as:

$$\mathcal{T}(\vec{z}) = 1 + relu(W_2\ relu(W_1\vec{z} + \vec{b_1}) + \vec{b_2}), \qquad (2)$$

where W_1,W_2, b_1 and b_2 are the network weight matrices and bias vectors and $\vec{z}$ is the input representation, for which we use the intermediate features of the original activity recognition network ($\vec{z}$ has a size of 1024 for Infalted 3D ConvNet and 2048 for Pseudo 3D ResNet). We then scale the logits by the inferred instance-dependent temperature $\mathcal{T}(\vec{z})$ and our prediction probability becomes:

$$conf(a_{pred}) = \max_{a \in \mathcal{A}} \frac{exp(\frac{y_a}{\mathcal{T}(\vec{z})})}{\sum_{\hat{a} \in \mathcal{A}} exp(\frac{y_{\hat{a}}}{\mathcal{T}(\vec{z})})}. \qquad (3)$$

We train the CARING model on a held-out validation set with Negative Log Likelihood loss for 300 epochs (learning



(a) Expected Calibration Error improvement during the training procedure for validation and test data.

(b) Average temperature and its standard deviation estimated by our model during training.

Fig. 4: CARING model evolution during training for one Drive&Act split. Both average value and standard deviation of the learned input-dependent scaling parameter $\mathcal{T}(\vec{z})$ rise as the training proceeds (right figure). Jointly with the decrease of the calibration error (left figure), this indicates the usefulness of learning different scaling parameters for different inputs.

rate of 0.005, weight decay of $1e^{-6}$). Similarly to the approach described in Section III-D, CARING can be viewed as a post-processing step for obtaining better uncertainty confidence and *does not affect the predicted activity class and model accuracy*, as the order of the output neurons does not change.

We validate, that learning input-dependent temperature value is indeed better than using a single global scaling parameter by examining the evolution of different model metrics during training. Figure 4 illustrates changes of the expected calibration error (defined in Section III-A) and the average and standard deviation of the inferred scaling parameter $\mathcal{T}(\vec{z})$ measured over the validation data. Figure 4b reveals, that both, the mean and standard deviation of temperature rises during training, leading to a lower calibration error (Figure 4a). The observed increase in the standard deviation of the scaling parameter confirms that handling the logits differently dependent on the input is beneficial in our task.

IV. Experiments

A. Benchmark settings

Since there is no established evaluation procedure targeting the reliability of confidence values in context of activity recognition, we adapt existing evaluation protocols for two conventional action classification datasets, Drive&Act [17] and HMDB-51 [13], for our task. We choose the Drive&Act [17] testbed for driver activity recognition as our main benchmark, as it is application-driven and encompasses multiple challenges typical for real-life systems (e.g. fine-grained categories and unbalanced data distribution). Drive&Act comprises 34 fine-grained activity classes, which, however are highly unbalanced as the number of examples ranged from only 19 examples of *taking laptop from backpack* to 2797 instances of *sitting still*. As CNNs have a lower performance when learning from few examples, we sort the behaviors by their frequency in the dataset and divide them into *common* (top half of the classes) and rare (the bottom half). We subsequently evaluate the models in three modes: considering *all activities*, as it is usually done, using only the *overrepresented*- or only the *rare* classes.

Model	ECE		NLL	
	val	test	val	test
Drive&Act - Common Classes				
P3D [24] Ⓢ	16.9	19.39	1.63	1.85
I3D [2] Ⓢ	10.22	13.38	0.90	1.27
P3D + Temperature Scaling [8]Ⓤ	5.65	5.7	1.28	1.48
I3D + Temperature Scaling [8]Ⓤ	5.31	6.99	0.57	0.83
CARING - P3D (ours) Ⓤ	4.81	**4.27**	1.19	1.42
CARING - I3D (ours) Ⓤ	**2.57**	5.26	**0.50**	**0.78**
Drive&Act - Rare Classes				
P3D [24] Ⓢ	31.49	37.25	3.43	4.68
I3D [2] Ⓢ	31.48	43.32	3.41	4.54
P3D + Temperature Scaling [8] Ⓤ	17.83	21.09	2.26	2.99
I3D + Temperature Scaling [8] Ⓤ	24.97	32.38	1.96	2.62
CARING - P3D (ours) Ⓤ	**13.73**	**19.92**	2.12	2.93
CARING - I3D (ours) Ⓤ	18.34	23.6	**1.55**	**2.17**
Drive&Act - All Classes				
P3D [24] Ⓢ	17.89	21.09	1.77	2.12
I3D [2] Ⓢ	11.72	15.97	1.10	1.56
P3D + Temperature Scaling [8] Ⓤ	5.89	6.41	1.35	1.63
I3D + Temperature Scaling [8] Ⓤ	6.59	8.55	0.68	0.99
CARING - P3D (ours) Ⓤ	4.58	**5.26**	1.26	1.57
CARING - I3D (ours) Ⓤ	**3.03**	6.02	**0.58**	**0.9**
HMDB-51				
I3D [2] Ⓢ	10.29	20.11	0.98	1.97
I3D + Temperature Scaling [8] Ⓤ	4.00	7.75	**0.81**	1.57
CARING - I3D (ours) Ⓤ	**3.38**	**5.98**	**0.81**	**1.54**

Ⓢ Standard activity recognition models Ⓤ Uncertainty-aware models

TABLE I: Reliability of confidence values on the Drive&Act [17] and HMDB-51 [13] datasets for standard activity recognition models and their extensions with uncertainty-aware calibration algorithms.

We further validate the models on HMDB-51 [13], a more general activity recognition dataset comprising of YouTube videos. The benchmark covers 51 activity classes, which are more discriminative in their nature (e.g. laughing and playing football) and are perfectly balanced (three splits with 70 training and 30 test examples for every category).

Input to the P3D- and I3D models are snippets of 64 consecutive frames. If the original video segment is longer, the snippet is chosen randomly during training and at the video center at test-time. If the video segment is shorter, we repeat the last frame until the 64 frame snippet is filled.

Following the problem definition of Section III-A, we extend the standard accuracy-driven evaluation protocols [13], [17] with the expected calibration error (ECE), depicting the deviation of model confidence score from the true misclassification probability. In addition, we report the Negative Log Likelihood (NLL), as high NLL values are linked to model miscalibration [8]. Since HMDB-51 does not contain a validation split, we randomly separate 10% of the training data for this purpose. As done in the original works [13], [17], we report the average results over the three splits for both testbeds.

B. Confidence Estimates for Action Recognition

In Table I we compare CNN-based activity recognition approaches and their uncertainty-aware versions in terms of the

Activity	Number of Samples	Recall	I3D Ⓢ Mean Conf.	I3D Ⓢ ΔAcc	I3D Ⓢ ECE	CARING-I3D Ⓤ Mean Conf.	CARING-I3D Ⓤ ΔAcc	CARING-I3D Ⓤ ECE
Five most common activities								
sitting_still	2797	95.1	97.96	2.86	2.86	93.84	-1.26	**1.84**
eating	877	86.42	93.26	6.84	9.33	80.99	-5.43	**5.75**
fetching_an_object	756	76.03	93.77	17.74	18.28	79.42	3.4	**5.32**
placing_an_object	688	66.77	93.03	26.25	26.25	75.9	9.13	**9.25**
reading_magazine	661	92.93	98.58	5.65	6.09	93.35	0.42	**2.87**
Five most underrepresented activities								
closing_door_inside	30	92.31	98.51	6.21	**8.22**	86.00	-6.31	8.30
closing_door_outside	22	81.82	93.55	11.73	20.97	86.86	5.04	**19.81**
opening_backpack	27	0	98.82	98.82	98.82	82.69	82.69	**82.69**
putting_laptop_into_backpack	26	16.67	92.67	76.00	76.00	76.46	59.8	**59.80**
taking_laptop_from_backpack	19	0.00	85.25	85.25	85.25	70.08	70.08	**70.08**
Ⓢ Standard activity recognition models						Ⓤ Uncertainty-aware models		

TABLE II: Analysis of the resulting confidence estimates of the initial I3D model and its CARING version for individual common and rare Drive&Act activities. *Recall* denotes the recognition accuracy of the current class, while *Mean Conf.* denotes the average confidence estimate produced by the model. Supplemental to the Expected Calibration Error (*ECE*), we report the difference between the mean confidence value and model accuracy (denoted ΔAcc). While in a perfectly calibrated model ΔAcc is 0, ECE is a better evaluation metric, as e.g. if a lot of samples have too high and too low confidence values, their average might lead to a misconception of good calibration. While there is room for improvement for underrepresented and poorly recognized activity classes, the CARING model consistently leads to better uncertainty estimates.

expected calibration error and NLL for *rare*, *overrepresented* and *all* Drive&Act classes as well as in the HMDB-51 setup. First, we verify our suspicion that native activity recognition architectures provide unreliable confidence estimates: confidence scores produced by I3D score have a misalignment of 15.97% for Drive&Act and 20.11% for HMDB-51. Similar issues are present in P3D: 21.2% ECE on Drive&Act, an error far too high for safety-critical applications.

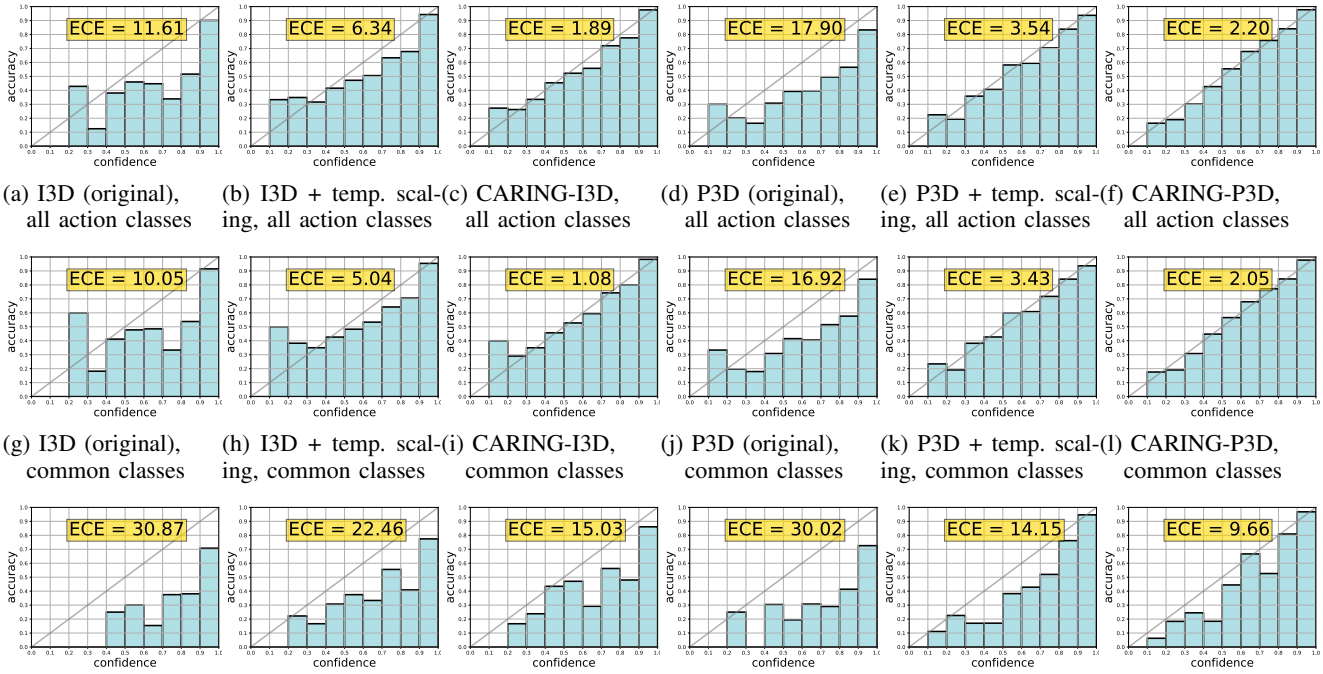
Model reliability is clearly improved by learning to obtain proper probability estimates, as all uncertainty-aware variants surpass the raw Softmax values. Interestingly, although I3D has better initial uncertainty estimates than P3D (ECE of 21.09% for P3D, 15.97% for I3D), P3D seems to have a stronger response to both, temperature scaling and CARING approaches then I3D (ECE of 5.26% for CARING-P3D, 6.02% for CARING-I3D). However, as this difference is very small ($< 1\%$), we would rather recommend using I3D, as it mostly gives higher accuracy [2], [17], [24]. While we consider the expected calibration error to be of vital importance for applications, we realize that this metric is complementary to model accuracy and encourage taking both measures into account when selecting the right model. We want to remind, that both temperature scaling and the CARING method *do not influence the model accuracy* (see Sections III-C and III-D). For Pseudo 3D ResNet we achieve an overall accuracy of 54.86% (validation) and 46.62% (test) on Drive&Act, which does not change through our uncertainty-based modifications. Consistently with [17] I3D achieves a higher accuracy of 68.71% for validation and 63.09% for test set [1].

[1]The slight deviation from the accuracy reported in the original work [17] (between 0.18% and 1.3%) is due to random factors in the training process.

As expected, the model confidence reliability correlates with the amount of training data (see distinguished areas for *common*, *rare* and *all* classes of Drive&Act in Table I). For example, the *common classes* setting encounters the lowest expected calibration error for both original and uncertainty-aware architectures (13.38% for I3D, 5.26% for CARING-I3D). Leveraging intermediate input representation via our CARING calibration network leads to the best probability estimates on both datasets and in all evaluation settings. Thereby, the CARING strategy surpasses the raw neural network confidence by 9.95% and the temperature scaling method by 2.53% on Drive&Act, highlighting the usefulness of learning to obtain probability scores *depending on the input*.

We further examine model performance for the individual classes, considering the five most frequent and the five most uncommon Drive&Act activities separately in Table II. In addition to ECE, we report the accuracy for samples belonging to the individual class, the average confidence value they obtained with the corresponding model and the difference between them (denoted ΔAcc). While a such global confidence-accuracy disagreement is interesting to consider (and is 0 for a perfectly calibrated model) it should be viewed with caution, as it might lead to an incorrect illusion of good confidence calibration, as e.g.a lot of samples with too high and too low confidence values might cancel each other out through averaging.

Reliability of the confidence scores is significantly improved through the CARING method and is connected to the amount of training data and the accuracy. Models have significant issues with learning from few examples (e.g. 76% I3D and 59.80% CARING-I3D ECE for *putting laptop into backpack*). For both, over- and underrepresented classes, the ECE of



(a) I3D (original), all action classes (b) I3D + temp. scaling, all action classes (c) CARING-I3D, all action classes (d) P3D (original), all action classes (e) P3D + temp. scaling, all action classes (f) CARING-P3D, all action classes

(g) I3D (original), common classes (h) I3D + temp. scaling, common classes (i) CARING-I3D, common classes (j) P3D (original), common classes (k) P3D + temp. scaling, common classes (l) CARING-P3D, common classes

(m) I3D (original), rare classes (n) I3D + temp. scaling, rare classes (o) CARING-I3D, rare classes (p) P3D (original), rare classes (q) P3D + temp. scaling, rare classes (r) CARING-P3D, rare classes

Fig. 5: Reliability diagrams of different models reflect the agreement between the confidence values and the empirically measured probability of correct prediction (results of one Drive&Act validation split). A model with *perfectly calibrated uncertainty scores would match the diagonal* (a detailed explanation in Section IV-C). Note, that the ECE values deviate from Table I, as they visualize a single split, while the final reported results are averaged over all splits. While the temperature scaling consistently improves the confidence estimates, our CARING model leads to the lowest calibration error in all settings.

easy-to-recognize activities (i.e.the ones with high accuracy) is lower. Before calibration, the average confidence value is always higher than the accuracy (positive ΔAcc) disclosing that the models are too optimistic in their predictions. Interestingly, after the CARING transformation is applied, the average model confidence is lower than the accuracy for some classes, such as *eating*. CARING models therefore tend to be more conservative in their assessment of certainty.

C. Calibration Diagrams

In Figure 5, we visualize the agreement between the predicted model confidence and the empirically measured probability of the correct outcome via reliability diagrams (explained in Section III-A) . In case of good estimates, the result will be close to the diagonal line. Values above the diagonal are linked to models being overly confident in their prediction, while values below indicate that the model doubts the outcome too much and the accurate prediction probability is higher than assumed.

First, we discuss the reliability diagrams of the original action recognition networks. Both P3D and I3D confidence values deviate from the target, with a clear bias towards too optimistic scores (i.e.values are oftentimes below the diagonal in Figures 5a, 5d, 5g, 5j, 5m, 5p). One exception is an above-diagonal peak in the low probability segment for *all* and *common* classes, meaning that in "easier" settings, low confidence examples often turn out to be correct (5a, 5d, 5g, 5j). In the "harder" setting of *rare* activities (Figure 5m, 5p), the bias towards too high probabilities is present for all values.

We see a clear positive impact of temperature scaling (Figures 5b, 5e, 5h, 5k, 5n, 5q) and our CARING model (Figures 5c, 5f, 5i, 5l, 5o, 5r). CARING models outperform other approaches in all settings and lead to almost perfect reliability diagrams for *all* and *common* classes. Still, both temperature scaling and CARING methods have issues with rare classes, with model confidence still being too high, marking an important direction for future research.

Note, that ECE might be in a slight disarray with the visual reliability diagram representation, as the metric weighs the misalignment in each bin by the amount of data-points in it, while the reliability diagrams do not reflect such frequency distribution. For example, while the CARING-I3D model in Figure 5i slightly exceeds the target diagonal, it has lower expected calibration error than CARING-P3D which seems to produce nearly perfect results in Figure 5l. As there are only very few examples in the low-confidence bin, they are overshadowed by smaller differences in the high-confidence bins, which contribute much more as they have more samples.

V. Conclusion

Automated activity understanding opens doors for new ways of human-machine interaction but requires models that can identify uncertain situations. This paper goes beyond the traditional goal of high top-1 accuracy and makes the first step towards activity recognition models capable of *identifying their misclassifications*. To this intent, we measure the *reliability of model confidence* and evaluate it for two prominent action recognition architectures, revealing, that the raw Softmax values of such networks do not reflect the probability of correct prediction well. We further implement two strategies for learning to convert poorly calibrated confidence values into realistic uncertainty estimates. First, we combine the native action recognition models with the off-the-shelf temperature scaling [8] approach which divides the network logits by a single learned scalar. We then introduce a new approach which learns to produce individual input-guided temperature values dependent on the input representation through an additional calibration network. We show in a thorough evaluation, that our model consistently outperforms the temperature scaling method and native activity recognition networks in producing realistic confidence estimates. The experiment results hold great promise for uncertainty-aware models, a crucial step towards real-life applications of activity recognition algorithms.

References

[1] Aayush Bansal, Ali Farhadi, and Devi Parikh. Towards transparent systems: Semantic characterization of failure modes. In *European Conference on Computer Vision*, pages 366–381. Springer, 2014.

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[3] Morris H DeGroot and Stephen E Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983.

[4] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

[5] Laura Fontanari, Michel Gonzalez, Giorgio Vallortigara, and Vittorio Girotto. Probabilistic cognition in two indigenous mayan groups. *Proceedings of the National Academy of Sciences*, 111(48), 2014.

[6] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059, 2016.

[7] Patrick Gebert, Alina Roitberg, Monica Haurilet, and Rainer Stiefelhagen. End-to-end prediction of driver intention using 3d convolutional neural networks. In *IEEE Intelligent Vehicles Symposium (IV)*, 2019.

[8] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, 2017.

[9] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA*, pages 18–22, 2018.

[10] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *Proceedings of International Conference on Learning Representations*, 2017.

[11] Yingxiang Huang, Wentao Li, Fima Macheret, Rodney A Gabriel, and Lucila Ohno-Machado. A tutorial on calibration measurements and calibration models for clinical prediction models. *Journal of the American Medical Informatics Association*, 27(4):621–633, 2020.

[12] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2013.

[13] Hilde Kuehne, Hueihan Jhuang, Rainer Stiefelhagen, and Thomas Serre. Hmdb51: A large video database for human motion recognition. In *High Performance Computing in Science and Engineering '12*, pages 571–582. Springer, 2013.

[14] Aviral Kumar and Sunita Sarawagi. Calibration of encoder decoder models for neural machine translation. *arXiv preprint arXiv:1903.00802*, 2019.

[15] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in neural information processing systems*, 2017.

[16] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

[17] Manuel Martin, Alina Roitberg, Monica Haurilet, Matthias Horne, Simon Reiß, Michael Voit, and Rainer Stiefelhagen. Drive&act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2801–2810, 2019.

[18] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[19] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[20] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 427–436, 2015.

[21] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632, 2005.

[22] Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. Analyzing uncertainty in neural machine translation. *arXiv preprint arXiv:1803.00047*, 2018.

[23] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[24] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[25] Alina Roitberg, Ziad Al-Halah, and Rainer Stiefelhagen. Informed Democracy: Voting-based Novelty Detection for Action Recognition. In *British Machine Vision Conference (BMVC)*, September 2018.

[26] Alina Roitberg, Nikhil Somani, Alexander Perzylo, Markus Rickert, and Alois Knoll. Multimodal human activity recognition for industrial manufacturing processes in robotic workcells. In *ACM International Conference on Multimodal Interaction*, 2015.

[27] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[28] Niko Sünderhauf, Oliver Brock, Walter Scheirer, Raia Hadsell, Dieter Fox, Jürgen Leitner, Ben Upcroft, Pieter Abbeel, Wolfram Burgard, Michael Milford, et al. The limits and potentials of deep learning for robotics. *The International Journal of Robotics Research*, 37(4-5), 2018.

[29] Du Tran and Alexander Sorokin. Human activity recognition with metric learning. *Computer Vision–ECCV 2008*, pages 548–561, 2008.

[30] Gul Varol, Ivan Laptev, and Cordelia Schmid. Long-term temporal convolutions for action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2017.

[31] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 3551–3558. IEEE, 2013.

[32] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European Conference on Computer Vision*, pages 20–36. Springer, 2016.

[33] Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Icml*, volume 1, pages 609–616. Citeseer, 2001.

[34] Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699, 2002.