CAPE: CAM as a Probabilistic Ensemble for Enhanced DNN Interpretation - Supplementary Material -

Townim Faisal Chowdhury[1], Kewen Liao[2], Vu Minh Hieu Phan[1], Minh-Son To[3], Yutong Xie[1], Kevin Hung[4], David Ross[4], Anton van den Hengel[1], Johan W. Verjans[1], Zhibin Liao[1†]

[1]Australian Institute for Machine Learning, University of Adelaide, Australia, [2]Australian Catholic University, Australia, [3]Flinders University, Australia, [4]SA Pathology, Central Adelaide Local Health Network, Australia

1. Additional Information on ResNet-50

In this section, we show the tables mentioned by the two discussion points in the section "Discussion and Conclusion" of the main manuscript.

Prediction Confidence. In Table 1 we show the prediction confidence on training and validation sets, which demonstrates the need for softened classification prediction ($T = 2$) from the vanilla classification layer to guide the training of the CAPE model. In practice, we found that a larger T, *e.g.*, $T = 4$ will over-soften the vanilla model's prediction and reduce the performance of trained CAPE.

	Classifier Module	CMML	CUB	ImageNet
Train	Vanilla classification ($T = 1$)	98.4	85.8	76.3
	Vanilla classification ($T = 2$)	95.9	45.4	44.7
	Bootstrap (PF)	81.9	26.7	43.8
	Bootstrap (TS)	88.3	19.7	6.4
Val	Vanilla classification ($T = 1$)	96.9	75.7	79.7
	Vanilla classification ($T = 2$)	92.6	33.8	49.3
	Bootstrap (PF)	78.5	23.5	46.7
	Bootstrap (TS)	84.6	17.7	6.8

Table 1. The empirical mean of the prediction confidence over all three reported datasets on the ResNet-50 model.

Prediction Agreement. In Table 2, we show prediction agreement between the evaluated models. This demonstrates that each model's explanation is unique and cannot be used to explain each other even if they share the exact model parameters, *i.e.*, the Off-the-shelf model *vs*. Vanilla Classification model.

1.1. Additional Qualitative Figures

Due to the limited space in the main manuscript, we show the full qualitative examples and comparisons across all eight state-of-the-art CAM maps in this supplementary material. The additional samples are shown in Fig. 2, Fig. 3, and Fig. 4 for CUB, ImageNet, and CMML respectively, at the end of this document. The comparison between CAPE (PF) and (TS) suggests that (PF) CAPE generally gives a larger region of attention and accumulatively less softened class prediction, which is aligned with the observation in Table 1 (see the rows for Bootstrap (PF) and (TS)). In Fig. 4, the CAM, Grad-CAM, and Lift-CAM do not yield any attention for the CMML example. This is because these methods have produced all negative values for the respective CAM and the rectifier function clipped the values to zero, hence not showing any attention.

Compared models		CMML	CUB	ImageNet
Off-the-shelf	Vanilla Classification	90.5	89.0	87.4
Bootstrap (PF)	Off-the-shelf	91.9	94.9	93.8
Bootstrap (PF)	Vanilla Classification	96.6	89.3	88.1
Bootstrap (PF)	Softened Bootstrap (PF)	98.1	99.6	94.6
Bootstrap (PF)	Bootstrap (TS)	95.8	92.1	79.8
Bootstrap (TS)	Vanilla Classification	97.8	88.5	88.5

Table 2. Prediction agreement (%) between CAPE (TS/PF) and the Vanilla classification model evaluated on the ResNet-50 model. Softened Bootstrap (PF) denotes the prediction made by the CAPE layer with learned T' (see Eq. (10) in the main manuscript).

Method	CUB					
	AD ↓	IC ↑	ADD ↑	ADCC ↑	mIoU ↓	BC ↑
CAM [7]	3.5	49.7	27.7	54.9	96.79	3
Grad-CAM [5]	3.4	50.1	29.2	56.3	96.79	4
Grad-CAM++ [1]	4.2	47.5	26.3	58.4	95.69	0
Layer-CAM [3]	3.5	48.4	28.6	56.4	97.06	2
Score-CAM [6]	6.5	46.1	46.5	78.1	43.95	7
CAPE (PF)	14.3	33.7	22.7	71.4	17.19	4
CAPE (TS)	21.9	22.6	19.6	73.4	21.34	4
μ-CAPE (PF)	4.1	52.7	40.3	55.0	94.25	6
μ-CAPE (TS)	4.2	47.6	37.6	54.0	96.96	1

Table 3. Comparison of different CAM interpretation methods for CUB using Swin Transformer V2-B as the DNN architecture. ↓ and ↑ indicate lower or higher is better. The top-3 scores are marked from darker to lighter green colors.

†Corresponding author.

Method	CUB
Naive AVG	86.75
Off-the-shelf	87.15
Bootstrap (TS)	86.83
Bootstrap (PF)	87.14
Vanilla Classifier	87.12

Table 4. Accuracy comparison for Swin Transformer V2-B model on CUB.

2. Experiments on Swin Transformer Model

We trained the Swin V2-B transformer on CUB. The training configuration for the CUB dataset is the same as the ResNet-50 model for the CUB dataset, except the batch size is set to 16 to cope with the larger GPU memory usage of the Swin Transformer model. We report the accuracy comparison of the Swin Transformer [4], vanilla classification model, and CAPE methods in Table 4.

2.1. Quantitative Analysis

We compare our method with five state-of-the-art CAM methods (CAM [7], Layer-CAM [3], Score-CAM [6], Grad-CAM [5], Grad-CAM++ [1]) on CUB and ImageNet dataset. Table 3 presents the quantitative analysis using the same evaluation metrics in the main manuscript.

On CUB, the classification performance of the CAPE models is very close to the vanilla classifier even without training, shown in Table 4. In particular, the Off-the-shelf CAPE and Bootstrap (PF) models (87.15% and 87.12%) marginally surpass the performance of the vanilla classifier (87.12%). In contrast, the performance gap between the Vanilla Classifier and Bootstrap (PF) on ResNet-50 was 1.22%, the Vanilla Classifier was better. Finally, the Bootstrap (TS) resulted in a lower performance of 86.83%, suggesting the full training course is unnecessary.

2.2. Qualitative Analysis

Fig. 1 shows the visualization of different CAM methods for the Swin V2-B. It is clear that the model attention examples in Fig. 1 are all widespread. We suspect that this is due to the fact that the Transformer model tokenizes the image into non-overlapping patches and processes at the patch level, the spatial correspondence between the original input and the output CAM becomes weak. This means that all patches in a transformer layer can access information of all patches in the layer below. With the large model parameters encapsulated in Swin V2-B, all patch tokens likely learn similar attention pathways, therefore all visualized methods appear to have widespread attention placed on the input image.

3. CMML Dataset Details

3.1. Data Collection

The investigated Chronic myelomonocytic leukemia (CMML) dataset (data statistics shown in Table 5 (a)) was collected from the South Australian Pathology (SA Pathology) laboratory using a Cellavision DI-60 scanner from the period November 2021 to February 2023, in 4 batches. The blood film staining protocol used a dual Wright's/Giemsa 0.26% stain solution and Sorensen's buffer pH 6.8 from Kinetik. The scanner detected blood cells on individual images where the cell of interest is centered. We used the identified monocytes by the scanner as the raw input images. The produced monocyte images were then squared or nearly squared in height and width of 352 or 356 pixels, corresponding to a spatial resolution of $36 \times 36 \mu m$. The collected dataset of images was also manually examined to filter out non-monocytes classified incorrectly by the scanner and images with multiple monocytes. The process resulted in 4,067 monocyte images from 171 individuals. The labels are assigned at the individual level with two classes: Normal and CMML, determined by individual medical records. For each individual included in this study, the number of monocyte images varies from 5 to 171. CMML individuals have on average 46 images vs 17 for normal individuals. The causes of the variations include: 1) when the WBC count is very low (typically $< 0.5 \times 10^9/\mathrm{L}$), the scanner may have difficulties scanning sufficient WBCs in a study; 2) normal individuals have fewer monocytes ($< 10\%$ of total WBCs) than CMML individuals ($> 10\%$); and 3) suspected CMML individuals were repeatedly scanned. We capped the number of images per individual to 80 which further reduces the samples used for training and testing to 3,899.

3.2. Motivation

The CMML dataset depicts a clinically important but difficult diagnostic problem. In Table 5 (b), we show the result of a human study on 153 monocyte images (53% are from CMML individuals) rated by 3 hematologists, where the performances are largely inconsistent with the recorded diagnosis from bone marrow biopsy. This suggests that individual image-level recognition cannot be done reliably. We first make the assumption that the majority of CMML individuals will predominantly have abnormal monocytes, though some monocytes could be normal. Then, all the captured image instances of an individual inherit the same label from the individual level, for the purpose of training and testing. Finally, in the testing phase, the individual's diagnosis is aggregated by averaging the predictions from the image instances.

From Table 2 in the main manuscript, we show that fitting a vanilla ResNet-50 on this task achieves 90.5% mean

	Normal	CMML	Total
Training set	57 (928)	14 (616)	71 (1544)
Validation set	40 (748)	10 (472)	40 (1220)
Test set	40 (648)	10 (487)	50 (1,135)
Total	137 (2,324)	34 (1,575)	171 (3,899)

(a) Data statistics

	BM Diagnosis	Observer 1	Observer 2	Observer 3
BM Diagnosis	100.0	59.5	49.7	48.4
Observer 1	59.5	100.0	52.3	65.4
Observer 2	49.7	52.3	100.0	55.6
Observer 3	48.4	65.4	55.6	100.0

(b) Human performance & variability

Table 5. (a) Data statistics and (b) Human observer accuracy (%) against the bone marrow (BM) diagnosis and inter-observer agreement.

Semantic Class →		Nucleus	Nuc/Cyto Boundary	Cytoplasm	Cyto-Ext Boundary	Cell Exterior	Nuc/Cyto/Ext Boundary
Simplex Definition →		$(100, 0, 0)$	$(50, 50, 0)$	$(0, 100, 0)$	$(0, 0, 100)$	$(0, 50, 50)$	$(33, 33, 33)$
Method	Class						
CAPE (TS)	Normal	6.5±8.8	4.6±6.0	5.5±7.0	5.1±5.2	36.8±23.8	1.4±2.4
	CMML	11.8±16.3	4.0±5.5	2.6±3.9	2.1±2.9	18.7±17.4	0.9±1.8
	CMML-Normal	5.3±22.0	-0.6±10.0	-2.9±9.3	-2.9±7.0	-18.1±37.1	-0.4±3.4
CAPE (PF)	Normal	6.5±7.8	4.0±4.7	4.2±4.3	3.6±3.1	38.4±19.1	1.0±1.6
	CMML	5.9±5.9	3.2±3.5	2.8±3.5	2.5±3.0	27.0±19.1	0.9±1.4
	CMML-Normal	-0.6±12.7	-0.8±7.4	-1.4±6.9	-1.1±5.4	-11.4±36.9	-0.1±2.4

Table 6. The image region contributions (%) to 6 pre-defined semantic classes using the CAPE ResNet-50 model on all test images. For each method, the 12 (mean) contributions from the Normal/CMML class and the semantic class combinations sum to 100%.

accuracy. With the distribution of approximately 20% individuals belonging to the CMML category and sampled proportionally in the training, validation, and test sets, this accuracy indicates that the DNN may have found some image cues that correlate to the CMML diagnosis. The reason for evaluating the CMML dataset is to visualize what image regions have been used to make the model decisions in order to provide insights for the hematologists to understand any morphological/appearance change of CMML in monocytes.

3.3. Analysis and Discussion

CAM methods indicate image regions that matter to the model outcome but the region is not meaningful unless we know what is inside the region. To illustrate, in conventional image classification, we can instantly tell whether a CAM-highlighted region is part of a dog or other objects, so we can judge whether CAMs make sense. However, for CMML, we don't have that prior knowledge, therefore we first annotated randomly selected 220 images with nucleus and cytoplasm segmentation by a hematologist. These images were used to train a Mask R-CNN [2] model to produce predictions for the entire CMML dataset. With this information and CAPE-produced probabilistic image region contribution, we show that we can summarize the CAPE output of the entire test set to produce a statistical analysis of attention placement on different region types with semantic meanings: nucleus, cytoplasm, and cell exterior region.

The statistical analysis is shown in Table 6 by summarizing all image predictions (made from the five-fold cross-validation) for the entire dataset. For each image, we aggregate the image region predictions by horizontal, vertical flipping, and +/- 90% rotations. Since an image region can be a square that sits on the boundary of two or more semantic classes, we define six semantic classes shown as the column titles of Table 6. The definition of *simplex* for any semantic class is determined by predefined triplet percentages of (Mask R-CNN) segmentation pixels (Nucleus%, Cytoplasm%, Exterior%) compositing an image region. An image region's probabilistic contribution to the overall model decision is assigned to the bag of the closest semantic class determined by the $\mathcal{L}_2$ distance between the image region and the defined semantic class position on the simplex surface. Finally, for each combination of semantic class and diagnostic class, we compute the mean and standard deviation contribution value of the bag and show that in the corresponding cell in Table 6. We further include the statistics of the CAPE difference between the CMML and Normal classes. From Table 6, we derive several observations such as the following.

1. The nucleus region favors the CMML diagnosis more in the (TS) model but stays mutual in the (PT) model.
2. The cytoplasm region favors the Normal decision more.
3. The Cyto-Ext boundary shows a bias towards the Normal class.
4. The cell exterior region constitutes the largest decision-making and is more biased towards the Normal class. Note that the cell exterior region has the largest image area and hence potentially can host more attention placement.
5. The rest of the boundaries have relatively less area and do not show a significant bias to either class and hence contribute insignificantly to the overall decision.

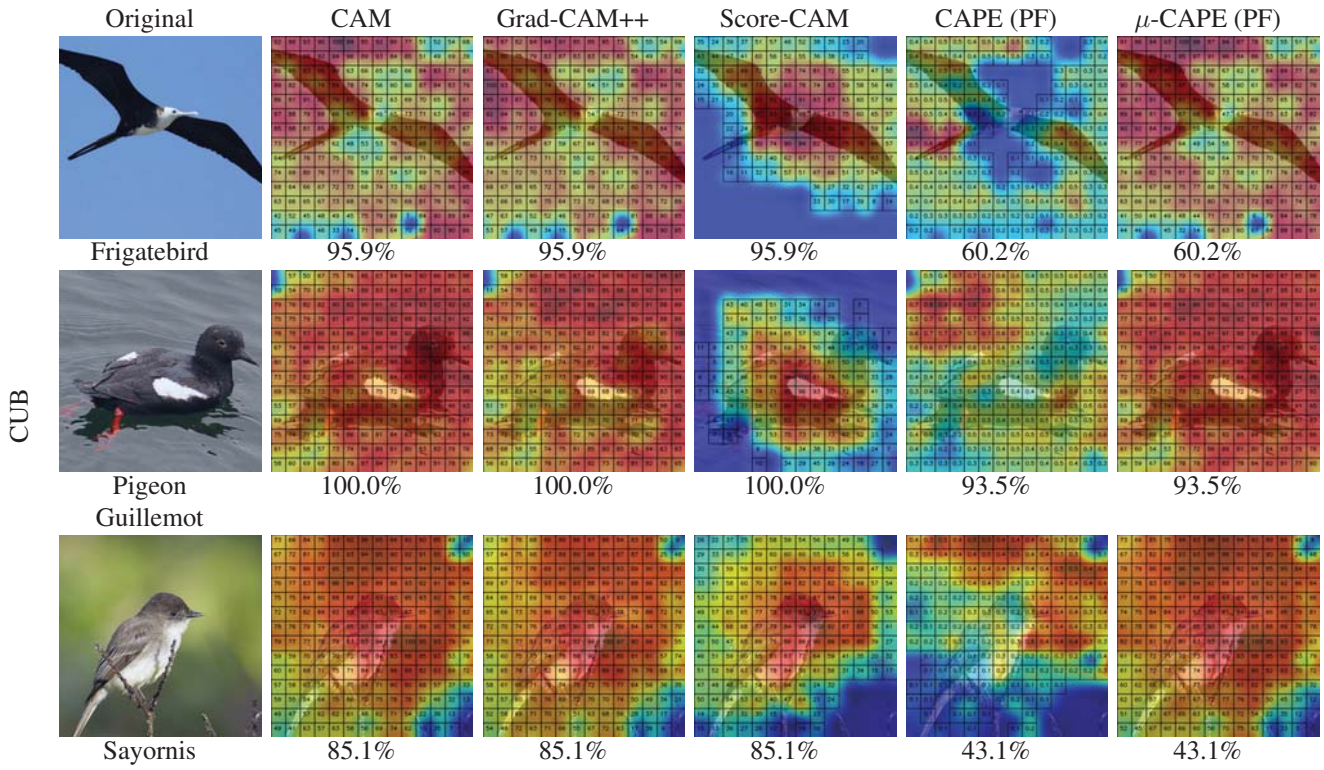

Figure 1. Qualitative analysis for the CUB dataset using the Swin V2-B model. The class confidence scores are shown under the respective explanation maps, where CAM, Grad-CAM++, and Score-CAM visualize for the original classification model. CAPE and μ-CAPE visualize for the post-fitted (PF) CAPE classification layer. Note that the shown values pre-upsampling values where we omit values $< 0.5\%$ for CAPE and μ-CAPE and $< 5\%$ for the other CAMs.

6. The nucleus and cell exterior are the two semantic regions that have the largest standard deviations, meaning they are frequently used to decide the CMML diagnosis.

Therefore, through these observations, one potential research direction is to look into the more fine-grained nucleus morphology analysis and another to examine the potential of red blood cell morphology analysis for CMML.

3.4. Dataset Availability

The ethical approval and data sharing agreement of the CMML research does not cover the public release of the image dataset. Hence the dataset will not be made publicly available.

References

[1] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018. 1, 2

[2] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *IEEE Intl. Conf. on Computer Vision*, pages 2961–2969, 2017. 3

[3] Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30:5875–5888, 2021. 1, 2

[4] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022. 2

[5] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE Intl. Conf. on Computer Vision*, pages 618–626, 2017. 1, 2

[6] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 24–25, 2020. 1, 2

[7] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016. 1, 2

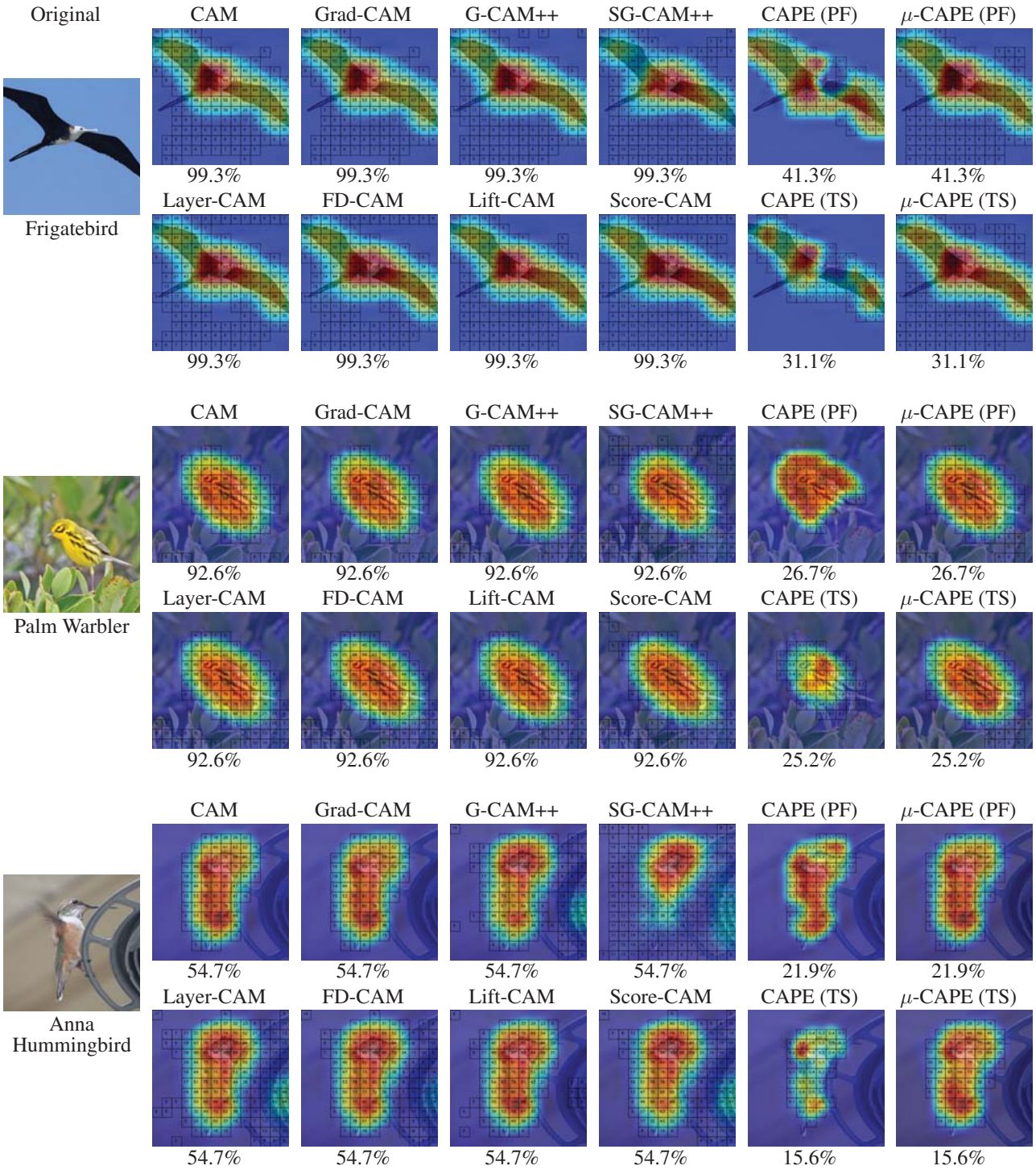

Figure 2. Qualitative visualization using the ResNet-50 backbone model for CUB dataset. The class confidence scores are shown under the respective explanation maps. "G-CAM++" and "SG-CAM++" denote Grad-CAM++ and Smooth Grad-CAM++ respectively.

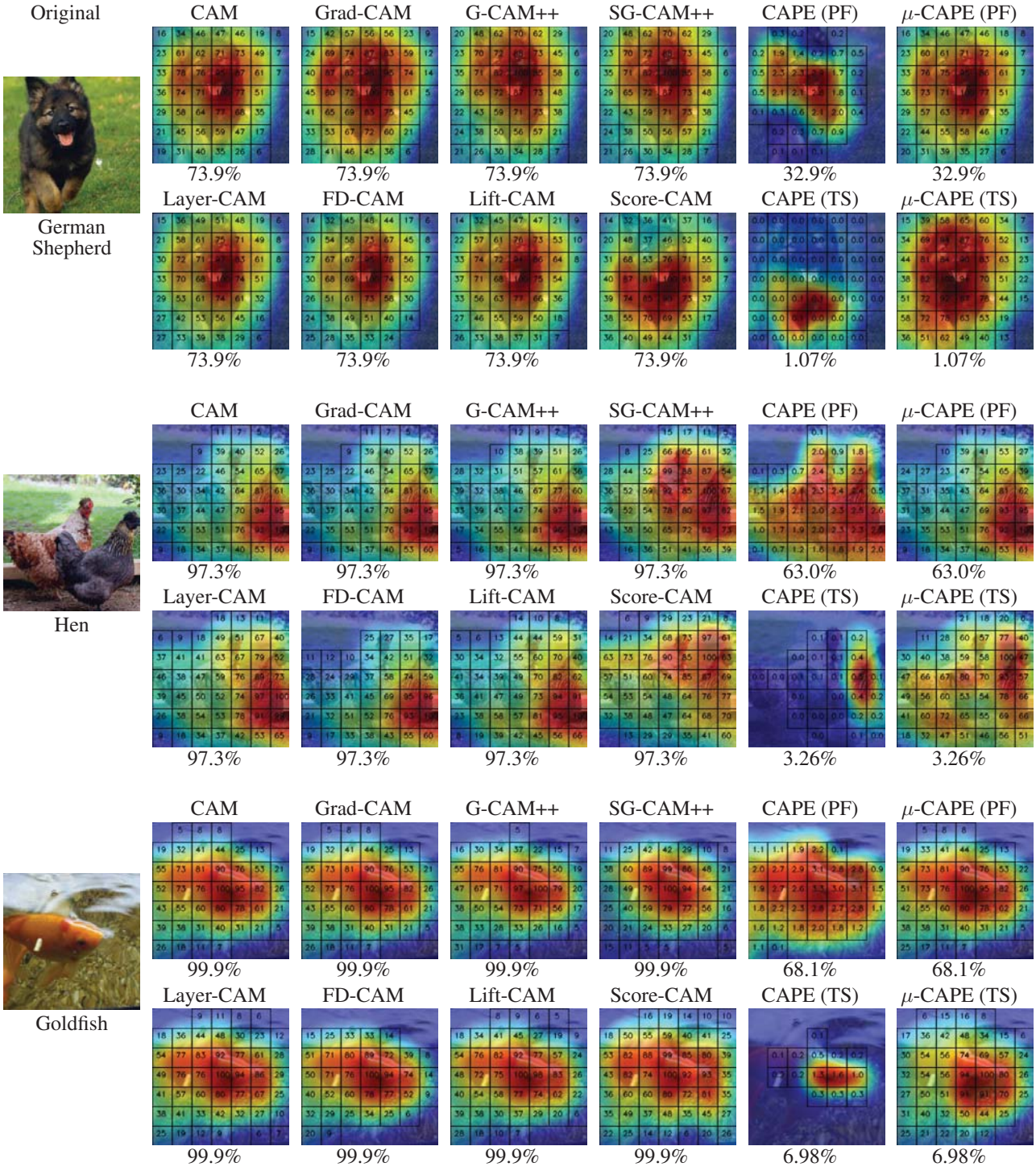

Figure 3. Qualitative visualization using the ResNet-50 backbone model for Imagenet. "G-CAM++" and "SG-CAM++" denote Grad-CAM++ and Smooth Grad-CAM++ respectively.

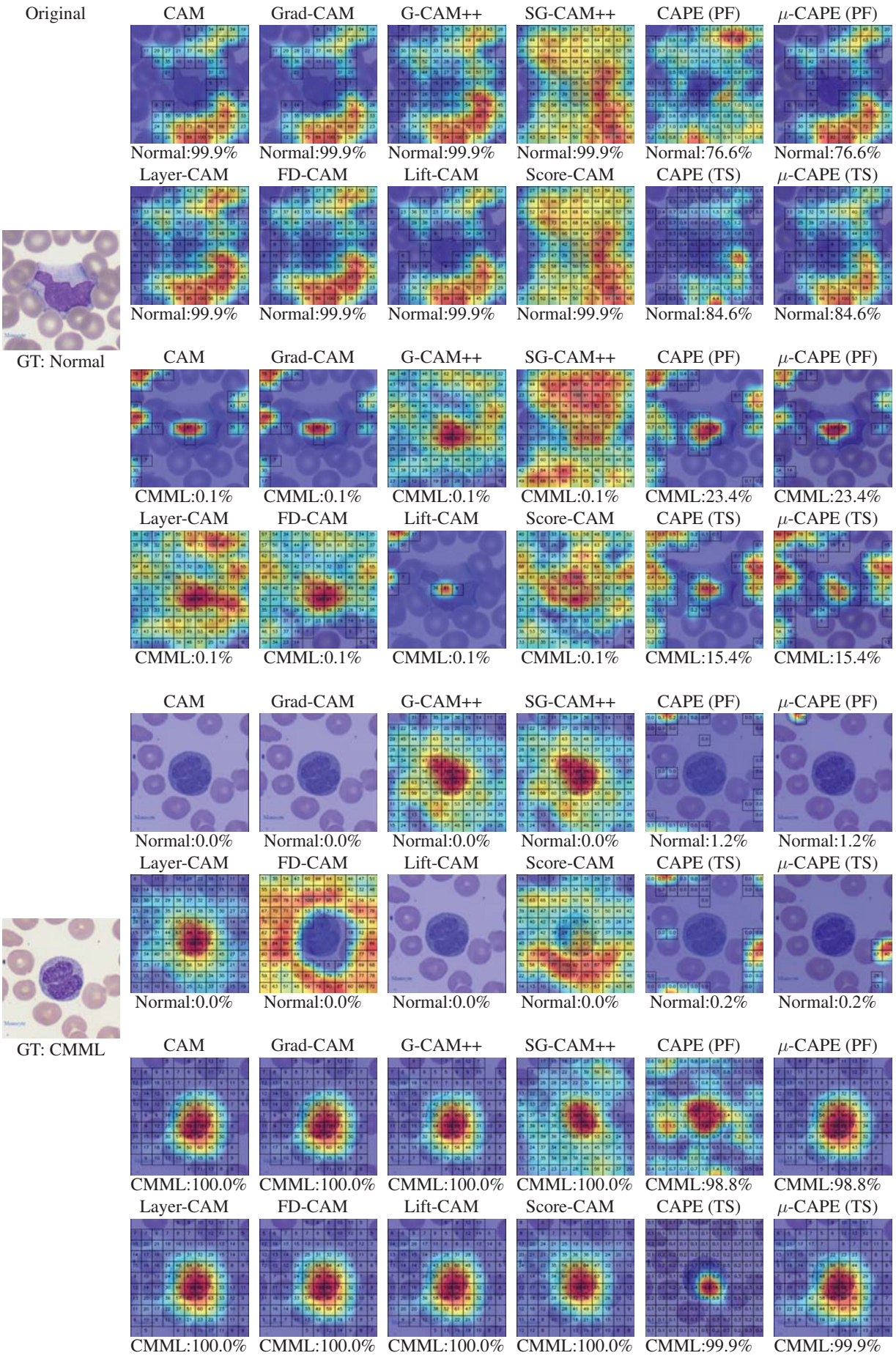

Figure 4. Qualitative visualization using the ResNet-50 backbone for one Normal example (top) and one CMML example (bottom).