CHiQPM: Calibrated Hierarchical Interpretable Image Classification

Thomas Norrenbrock[1], Timo Kaiser[1], Sovan Biswas[2], Neslihan Kose[2], Bodo Rosenhahn[1]
[1]Leibniz University Hannover, [2] Intel Labs

CVPR JUNE 3-7, 2026 DENVER COLORADO

Abstract

CHiQPM follows QPM[1], representing classes with very few general broadly shared concepts, unlike Prototypical models. CHiQPM improves the global interpretability, introduces hierarchical local explanations and the first form of interpretable conformal prediction by predicting along them.

Related Work

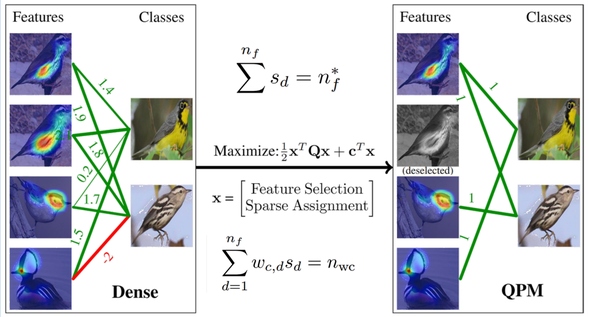
Conformal Prediction (CP)

- Using Split Conformal Prediction[2], one can guarantee an error rate of less than α by predicting all classes which are less non-conform than the α most nonconform ground truth samples were on equally distributed calibration data. Minimal requirements apply to the Nonconformity Score s

$$\mathbb{Y}(x_{test}) = \{c \in \mathbb{C} : s(x_{test}, c) \leq \text{Quantile}(1-\alpha, \mathbb{S})\}$$

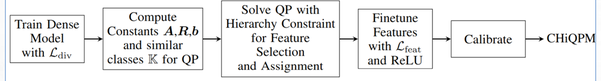
QPM

- After training a dense black-box model, QPM formulates a QP to select features and assigns them to classes. This binary QP is constrained for compactness via the number of features it selects and the number it assigns to each class. During the Finetuning with fixed assignment, features become concept detectors for concepts their assigned classes share.



Methods

- CHiQPM improves QPM via a constraint in the QP for interpretability and improved finetuning using Feature Grounding Loss $\mathcal{L}_{\text{feat}}$
- Its structure enables Hierarchical Local Explanations
- Interpretable Conformal Prediction creates prediction sets from within the hierarchical explanation that maintain guarantees of CP



CHiQPM

- Additional constraint in QP that enforces more contrastive class pairs

$$(\boldsymbol{w}_c \circ \boldsymbol{w}_{c'})^T \boldsymbol{s} = n_{\text{wc}} - 1 \quad \forall (c, c') \in \mathbb{K}$$

- K are $(\mathbf{n_c}\ \rho)$ most similar classes in Dense model
- Finetuning with ReLU and Feature Grounding Loss $\mathcal{L}_{\text{feat}}$

$$\mathcal{L}_{\text{feat}} = -\frac{\sum_{i\in\mathbb{F}} \frac{f_i^*}{|\mathbb{F}|} - \sum_{i\in\overline{\mathbb{F}}} \frac{f_i^*}{|\overline{\mathbb{F}}|}}{\max(f^*)}$$

Hierarchical Local Explanation

The hierarchical local explanations enable extended local explanations

- Features found in image as inner nodes sorted by activation
- Classes are leaves
- Green Path highlights prediction set determined via CP

CHiQPM's local explanations answer these questions simultaneously:

- What meaningful features of which classes are found in this image?
- How does each feature narrow down the set of potential predictions into increasingly similar classes?
- Which set shall be predicted to guarantee a configurable average accuracy?
- Which features would have needed to activate more strongly in order to predict a smaller set with sufficient certainty?

Interpretable Conformal Prediction

- Nonconformity Score s restricts predictions to hierarchical explanation, limits maximum level n^{limit} and considers subtrees. $\delta_n^{(c)}$ encodes if class c shares all the same top n features with predicted class $\hat{c}$.

$$s(c) = \underbrace{\delta^c_{n^{\text{limit}}}}_{\text{Limitation}} \cdot \underbrace{\left(-f^*_{i^{\text{div}}} - \sum_{j=1+n^{\text{limit}}}^{n_{\text{wc}}-1} \delta^c_j f^*_{\hat{\mathbb{F}}^c_j}\right)}_{\text{Limited } s_{\text{sel}}}$$

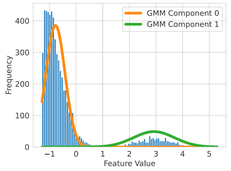
Results

Point Predictor

- State-of-the-art Performance as Compact Interpretable Point Predictor

Method	Accuracy ↑ CUB	CAR	IN	Com-pact	Contrastiveness ↑ CUB	CAR	IN	SG ↑ CUB
Dense Resnet50	86.6	92.1	76.1	-	74.4	75.1	71.6	34.0
glm-saga$_5$	78.0	86.8	58.0	○	74.0	74.5	71.7	2.5
PIP-Net	82.0	86.5	-	○	99.5	99.5	-	6.7
ProtoPool	79.4	87.5	-	○	76.7	78.9	-	13.9
SLDD-Model	84.5	91.1	72.7	+	87.2	89.7	93.4	29.2
Q-SENN	84.7	91.5	74.3	+	93.0	94.2	92.6	23.4
QPM	85.1	91.8	74.2	+	96.0	97.7	89.3	47.9
CHiQPM (**Ours**)	**85.3**	**91.9**	**75.3**	+	**99.9**	**100**	**99.9**	**75.0**

- CHiQPM reaches near perfect Contrastiveness



Method	Acc ↑	Feature↑ Alignment	Feature ↓ Sparsity
CHiQPM (**Ours**)	**85.3**	**3.8**	**22.3**
w/o $\mathcal{L}_{\text{feat}}$	**85.3**	3.0	39.9
and L1-Regularization	**85.3**	3.4	31.3
and L1 for $\overline{\mathbb{F}}$	85.1	3.6	26.4

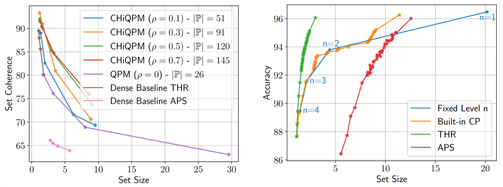
- Feature Grounding Loss $\mathcal{L}_{\text{feat}}$ improves Grounding and Sparsity

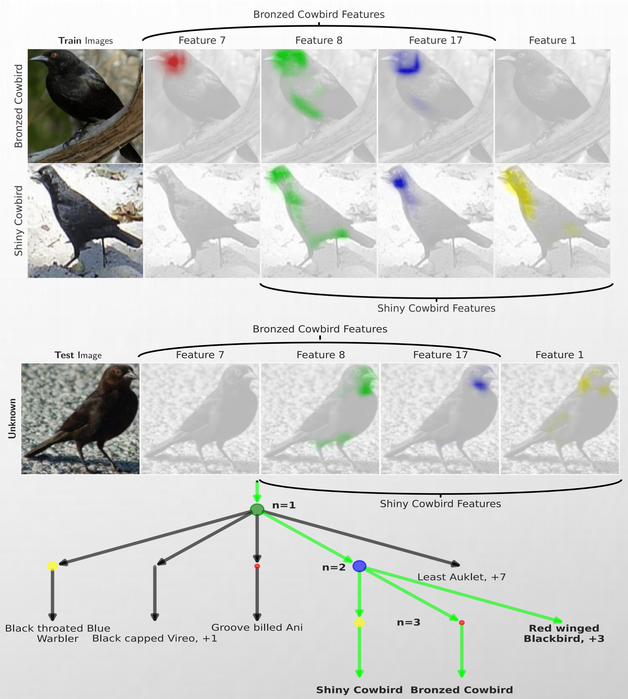
Conformal Prediction

- Built-in interpretable Conformal Prediction competitively efficient

$\|\mathbb{Y}\|\downarrow$ Method	Inter-pretable	CUB α=0.12	α=0.1	α=0.075	α=0.05	CARS α=0.075	α=0.05	α=0.0025	INET α=0.22	α=0.2	α=0.175	α=0.15
Ours	✓	1.22	1.73	2.94	9.05	1.05	1.25	8.25	1.10	1.42	3.25	4.58
$s = s_{\text{sel}}$	✓	4.62	6.15	9.53	29.4	3.62	5.95	28.4	8.14	11.3	17.9	30.5
$s = s_{\text{up}}$	✓	3.03	3.91	8.87	18.7	2.32	3.27	17.9	4.36	6.23	11.8	31.4
THR	✗	1.16	1.32	1.67	2.41	1.02	1.15	2.09	1.05	1.16	1.40	1.87
APS	✗	6.30	7.20	8.54	11.3	5.64	6.83	9.61	16.7	18.9	22.1	26.8

- Sets with increased Set Coherence





Global Explanation comparing Shiny and Bronzed Cowbird above
Local Explanation for test Image with no visible red eye

Conclusions

- CHiQPM as SOTA Compact Interpretable Point Predictor with Global and Local Interpretability
- Hierarchical Local Explanations that offer several additional insights into the reasoning for a single test image
- Interpretable Conformal Prediction that predicts strictly upwards in the Hierarchical Explanation while being competitively efficient

Contact

Thomas Norrenbrock
Leibniz University Hannover
Email: norrenbrock@tnt.uni-hannover.de

References

1. Norrenbrock, Thomas, et al. "QPM: Discrete Optimization for Globally Interpretable Image Classification." In: The Thirteenth International Conference on Learning Representations. 2025
2. Papadopoulos, Harris, et al. "Inductive confidence machines for regression." In: European conference on machine learning. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002

Code with Demo: