Learning to Calibrate and Rerank Multi-label Predictions

Cheng Li, Virgil Pavlu, Javed Aslam, Bingyu Wang, and Kechen Qin
{chengli, vip, jaa, rainicy}@ ccs.neu.edu, qin.ke@husky.neu.edu
Northeastern University

Multi-label Classification Problem

Assign a subset of candidate labels to an object (image, document, video, etc.)

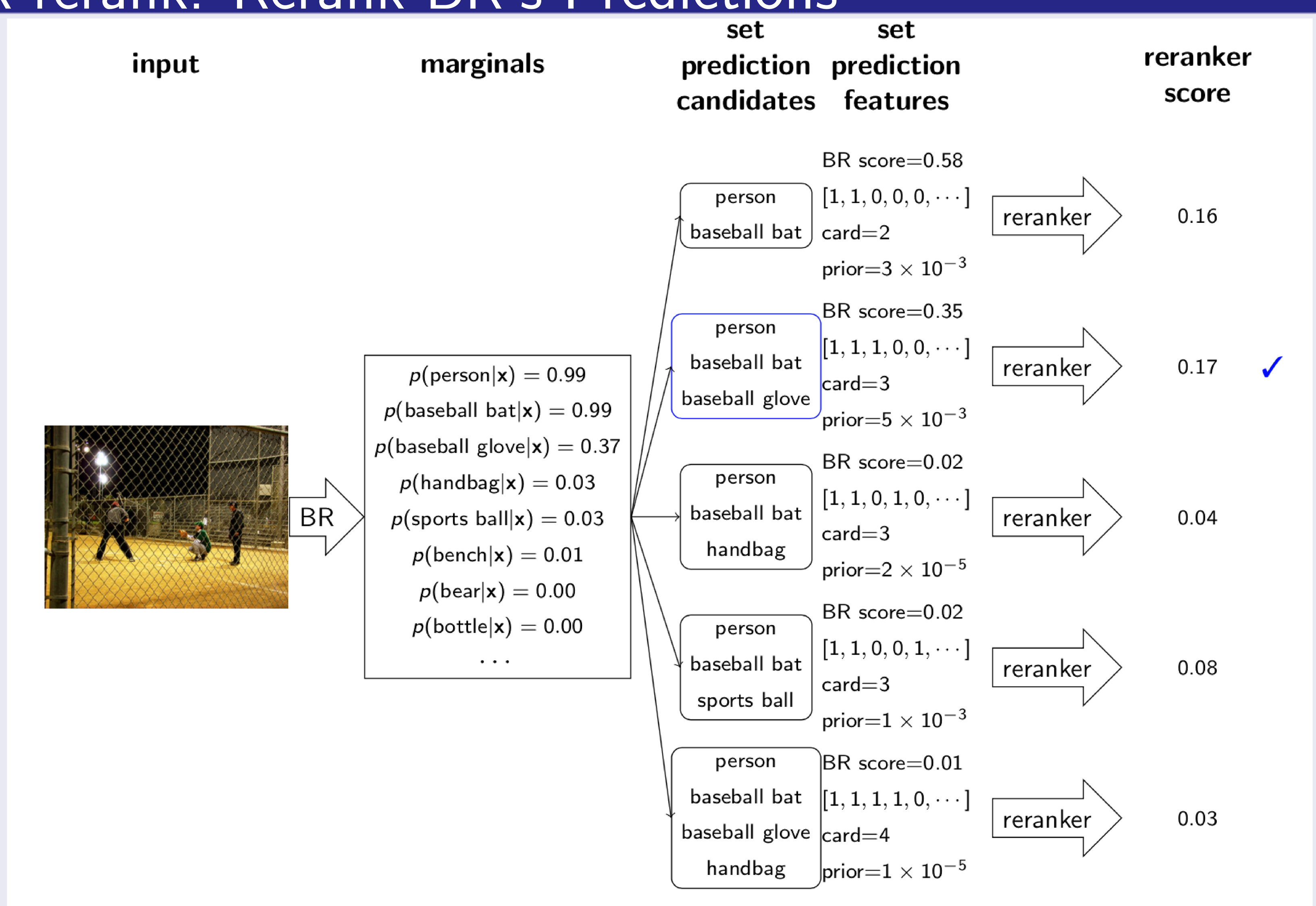
$$\mathbf{x} \xrightarrow{h} \mathbf{Y} = [Y_1, Y_2, ..., Y_L] = \overbrace{[1, 0, 0, 1, 0, ..., 1]}^{\text{length } L}$$

☑ person, ☐ handbag ☐ bear, ☑ baseball bat, ☑ baseball glove, ☐ bottle, ☐ car ...

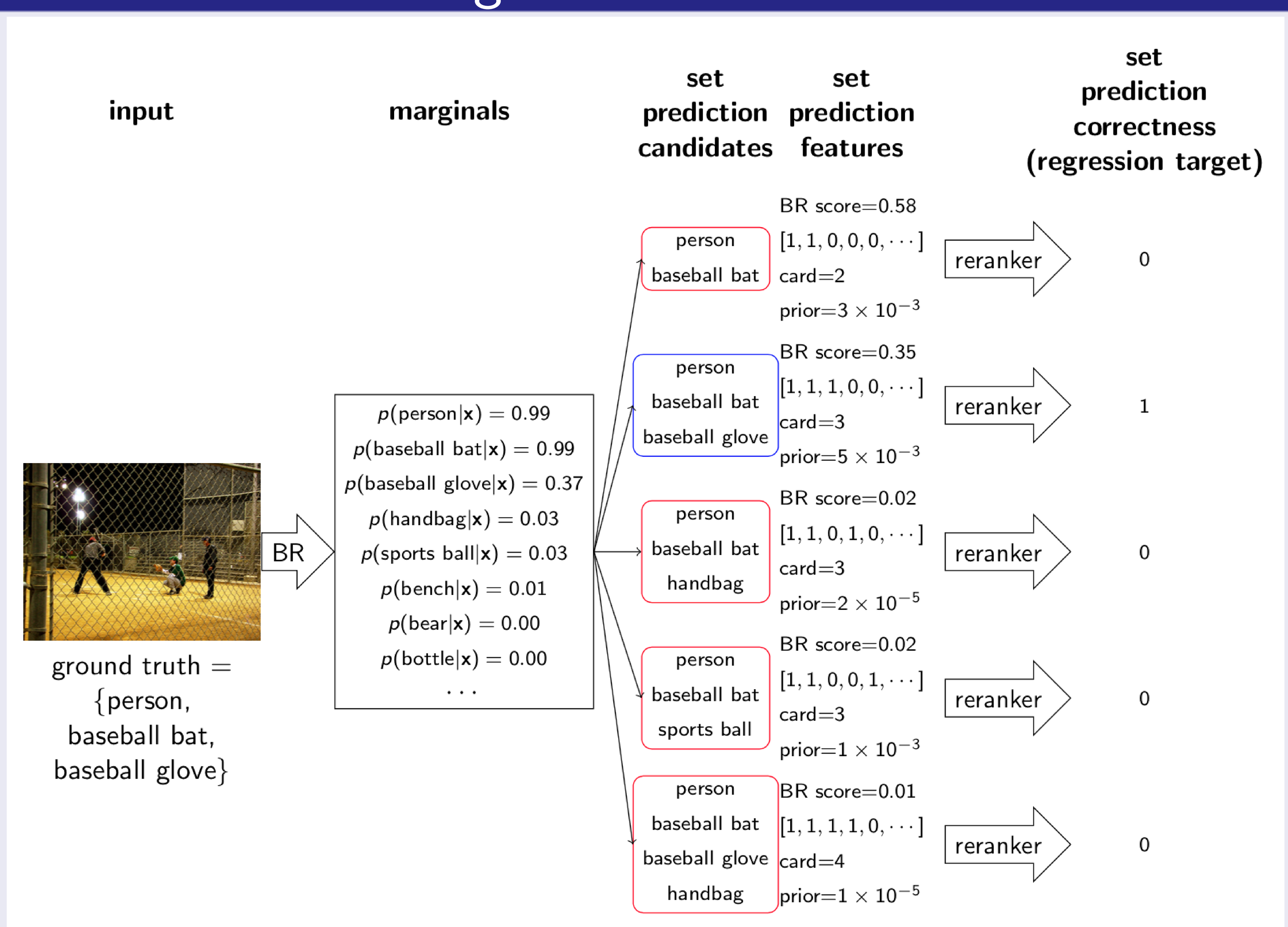
Commonly Used Method: Binary Relevance (BR)

- train one binary classifier to estimate each label probability $p(Y_\ell|\mathbf{x})$
- predict each label independently: predict label ℓ if $p(Y_\ell = 1|\mathbf{x}) > 0.5$
- prediction confidence $p(\mathbf{Y}|\mathbf{x}) = p(Y_1|\mathbf{x}) \times p(Y_2|\mathbf{x}) \times \cdots \times p(Y_L|\mathbf{x})$
- ☹ make prediction mistakes due to ignoring label dependencies. For example, BR fails to predict "baseball glove" for the image above.
- ☹ confidence score does not align with actual accuracy

BR-rerank: Rerank BR's Predictions



Train Reranker to Judge BR's Predictions



BR-rerank: Classification Accuracy

Dataset	BR	*BR-rerank*	2BR	DBR	CBM	CRF	SPEN	PDS	DVN	PC	PCC	Rakel	MLKNN
BIBTEX	16.6	21.5	16.1	20.2	22.9	**23.3**	14.8	16.1	16.2	20.3	21.4	18.3	8.4
OHSUMED	36.6	**42.0**	37.5	37.6	40.5	40.4	29.1	34.8	18.6	29.5	38.0	39.3	25.4
RCV1	44.5	53.2	42.3	45.8	**55.3**	53.8	27.5	40.8	13.7	39.7	48.7	46.0	46.2
TMC	30.4	**33.3**	32.1	31.7	30.8	28.2	26.7	23.4	20.3	23.0	31.3	27.6	18.9
WISE	52.9	60.5	51.8	55.8	**61.0**	46.4	-	52.4	28.3	-	55.9	3.5	2.4
MSCOCO	34.7	**35.9**	33.7	32.0	31.1	35.1	34.1	25.0	29.9	31.1	32.1	32.6	29.1
ranking	6.3	1.8	6.7	5.7	3.3	3.8	10.0	9.8	11.2	10.0	4.5	6.8	11.0

BR-rerank: Running Time (in seconds)

Dataset	BIBT	OHSUM	RCV1	TMC	WISE	MSCO
BR	4	3	7	8	80	1380
BR-rerank	9	6	10	11	88	1393
CBM	64	210	70	224	1320	8520
CRF	353	268	1223	771	16363	14760

Reranker Score: Calibrated Confidence

Confidence score is called calibrated if it aligns with accuracy.

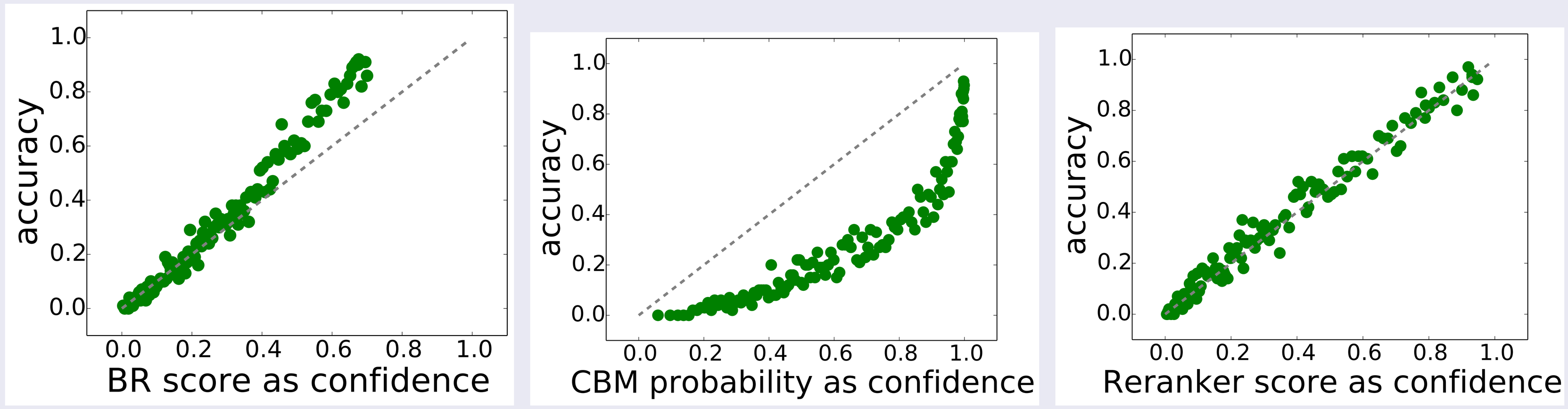
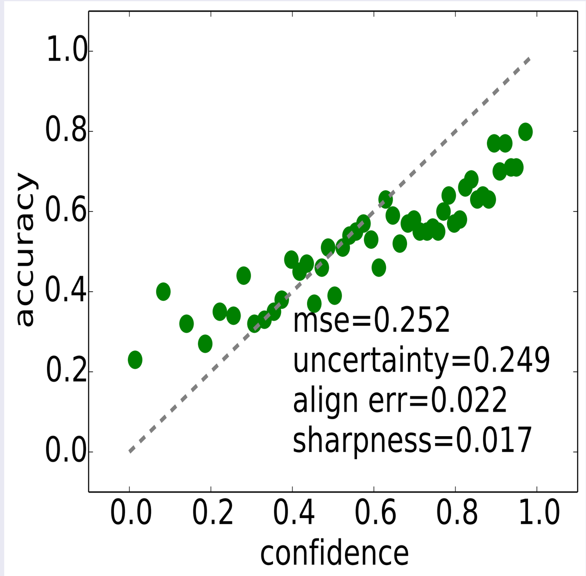
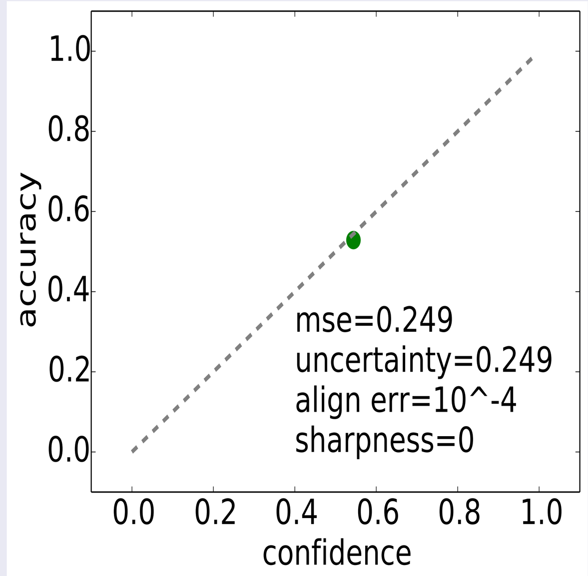


Figure: BR vs CBM vs Reranker confidence scores on MSCOCO data

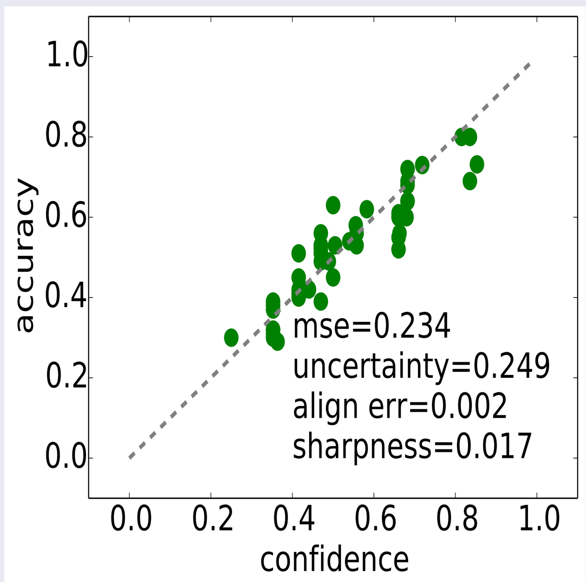
Reranker vs Other Post-Calibrators



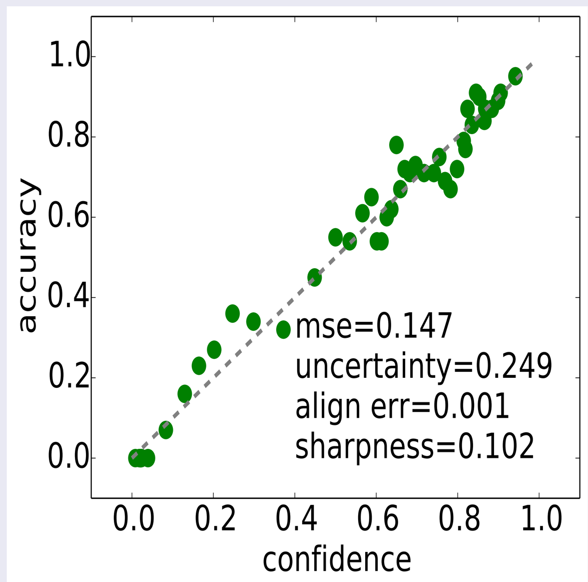
(a) uncalibrated score



(b) trivial calibrator



(c) isotonic regression



(d) reranker calibrator

Figure: Compare different post-calibrations for BR predictions on WISE data

Dataset	uncentainty	uncalib		isotonic		card isotonic		tree		reranker	
		MSE	sharp	MSE	sharp	MSE	sharp	MSE	sharp	MSE	sharp
BIBTEX	0.133	0.193	0.007	0.140	0.002	0.109	0.038	0.086	0.065	**0.068**	**0.072**
OHSUMED	0.232	0.226	0.015	0.221	0.013	**0.182**	**0.051**	0.211	0.039	0.189	0.047
RCV1	0.247	0.175	0.077	0.175	0.075	0.159	0.093	0.134	**0.129**	**0.123**	0.126
TMC	0.212	0.192	0.019	0.192	0.020	0.192	0.022	0.194	0.029	**0.180**	**0.032**
WISE	0.249	0.252	0.017	0.234	0.017	0.151	0.098	0.166	0.093	**0.147**	**0.102**
MSCOCO	0.227	0.158	0.075	0.151	0.075	0.150	0.076	0.163	0.070	**0.143**	**0.083**

Table: BR prediction calibration performance in terms of MSE (the smaller the better) and sharpness (the bigger the better).

Definitions of evaluation metrics:

- confidence score: $c(\mathbf{Y}) \in [0, 1]$
- 0/1 correctness: $v(\mathbf{Y}) \in \{0, 1\}$
- average accuracy of predictions with confidence c: $e(\mathrm{c}) = p[v(\mathbf{Y}) = 1|c(\mathbf{Y}) = \mathrm{c}]$
- alignment error: $\mathbb{E}[e(c(\mathbf{Y})) - c(\mathbf{Y})]^2$; the discrepancy between the claimed confidence and the actual accuracy; the smaller the better.
- sharpness: $\mathrm{Var}[e(c(\mathbf{Y}))]$; how widely spread the confidence scores are; the bigger the better.
- mean squared error (MSE): $\mathbb{E}[(v(\mathbf{Y}) - c(\mathbf{Y}))^2]$; the difference between the confidence and the actual 0/1 correctness; the smaller the better.

$$\underbrace{\mathbb{E}[(v(\mathbf{Y}) - c(\mathbf{Y}))^2]}_{MSE} = \underbrace{\mathbb{E}[(e(c(\mathbf{Y})) - c(\mathbf{Y}))^2]}_{alignment\ \ error} - \underbrace{\mathrm{Var}[e(c(\mathbf{Y}))]}_{sharpness} + \underbrace{\mathrm{Var}[v(\mathbf{Y})]}_{uncertainty}$$