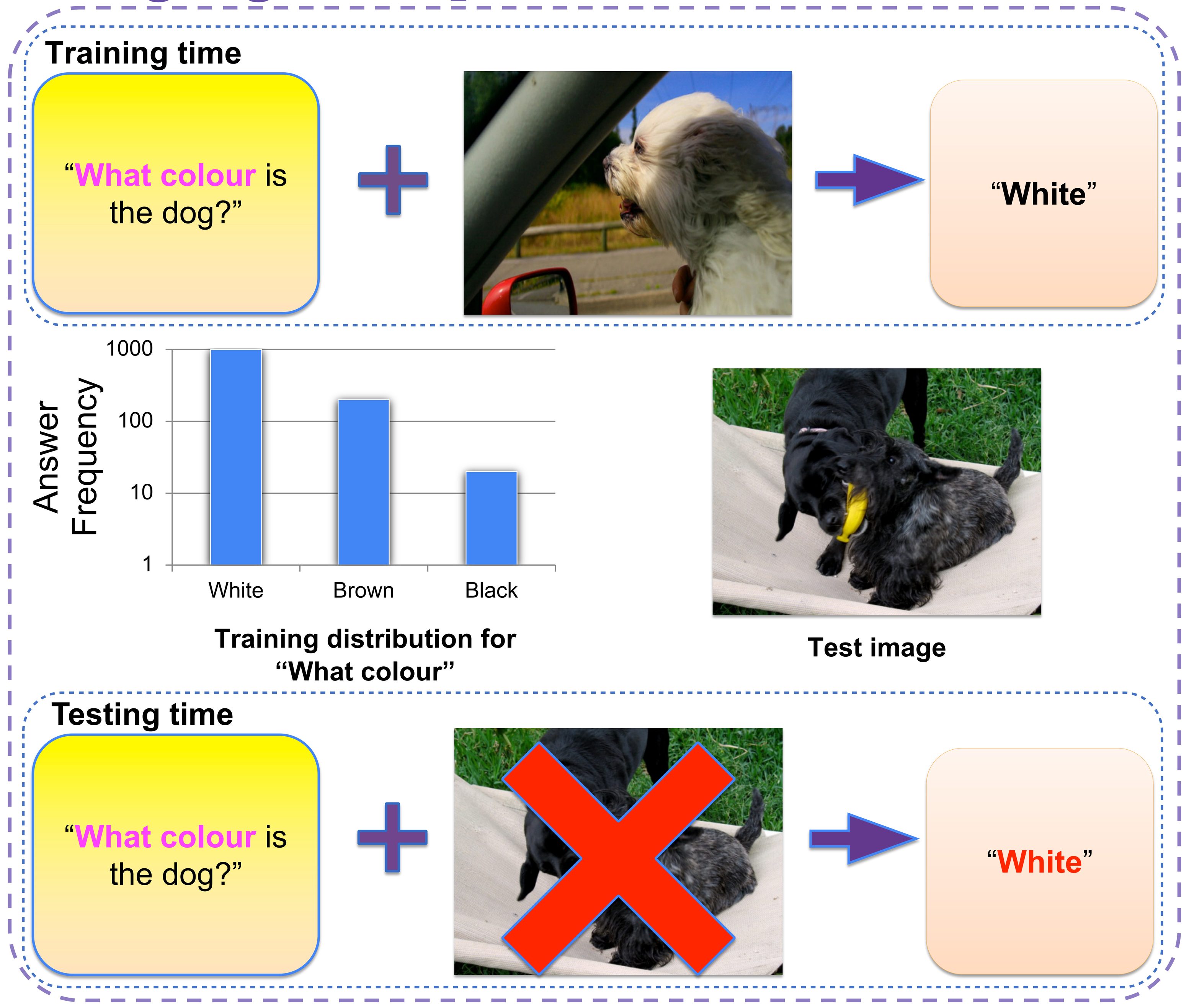
RMLVQA: A Margin Loss Approach for Visual Question Answering with Language Biases

Abhipsa Basu, Sravanti Addepalli, R.Venkatesh Babu
Vision & AI Lab, Indian Institute of Science, Bangalore, India

Contributions

- **Adaptive Angular Margin Loss**
  - A novel loss formulation manipulating the multimodal feature space, where the margins are estimated both from the training data and the model predictions.
  - Achieves state-of-the-art performance on the VQA-CP benchmark.
  - Model-agnostic.
- **Robust to Answer Distribution of test set**
  - Test-time ensembling makes the model generalisable to the in-domain VQA-v2 validation set.

Language Bias problem in VQA



Normalised Cross-Entropy Loss

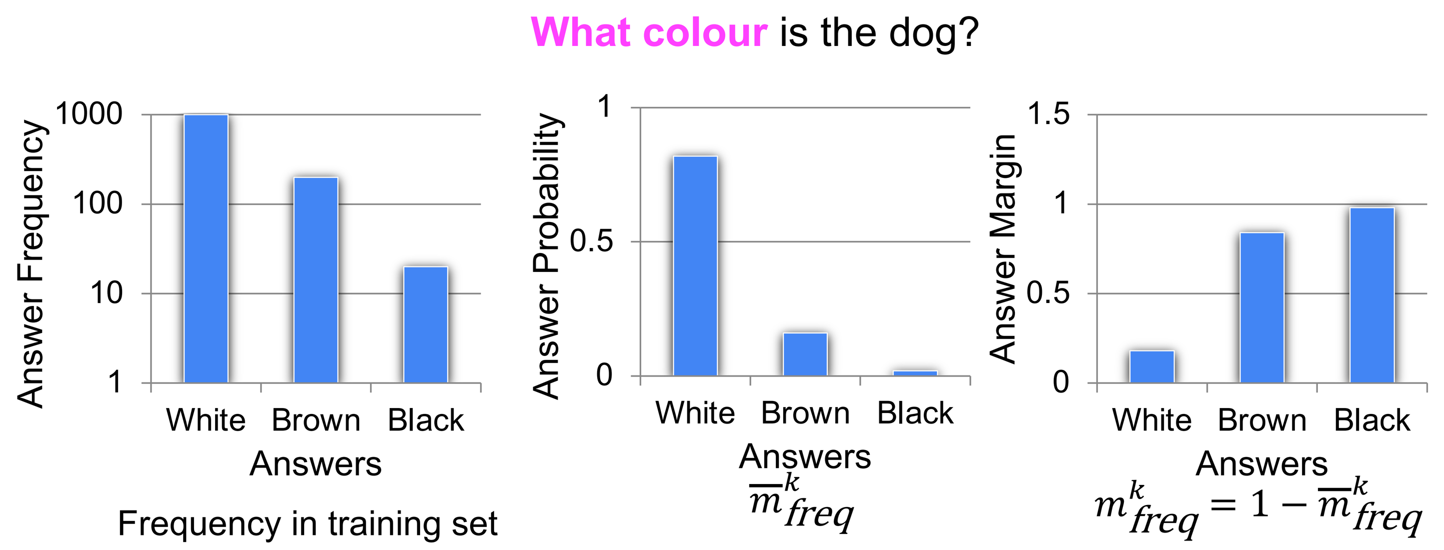
Let x be the final feature vector, and W be the weight matrix for the classifier.

(a) $f_i = W_i^T x$ (b) $\hat{W}_i = \frac{W_i}{\|W_i\|}$ (c) $\hat{x} = s\frac{x}{\|x\|}$

(d) $\hat{f}_i = \hat{W}_i^T \hat{x} = \|\hat{W}_i\| \|\hat{x}\| \cos\theta_i$

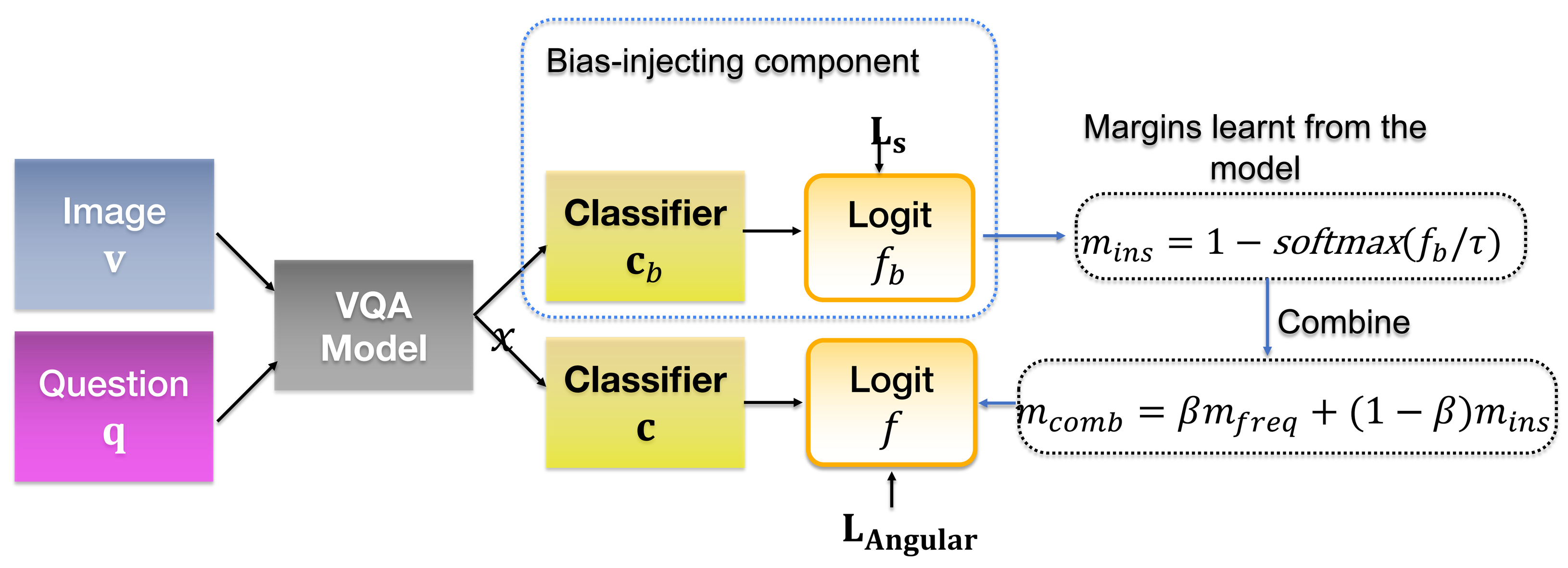
(e) $L_{ns} = \sum_{i=1}^{|\mathcal{A}|} - a_i \log \frac{\exp(s\cos\theta_i)}{\sum_{j=1}^{|\mathcal{A}|} \exp(s\cos\theta_j)}$ $a_i \in \{0,1\}$ - one hot encoding

Adaptive Margin Calculation



- Avoid overfitting of the calculated frequency-based margins to the sparse answers (like Black) by by passing them through a Gaussian [2], i.e. $\overline{m}^k_{ran}[i] = \mathcal{N}(\overline{m}^k_{freq}, \sigma)$, where $i = 1,2,\ldots,|\mathcal{A}|$. σ is a hyper-parameter
- Finally, the randomised margins are calculated by inverting the above, i.e.$m^k_{ran}[i] = 1 - \overline{m}^k_{ran}[i]$

Overview of RMLVQA and the learnt margins



The final angular margin loss becomes:

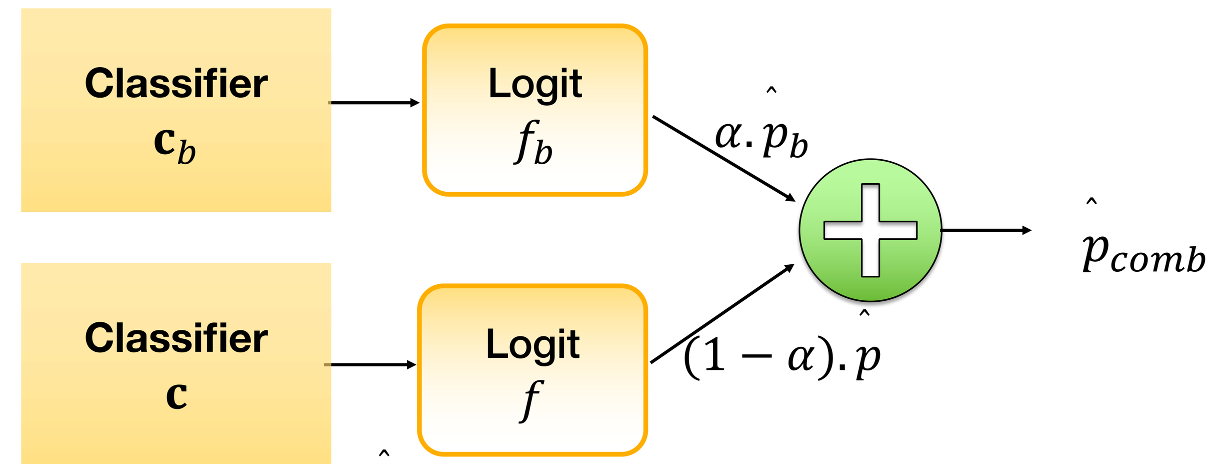
$$L^k_{Angular} = \sum_{i=1}^{|A|} - a_i \log \frac{\exp(s\cos(\theta_i + m^k_{comb}[i]))}{\sum_{j=1}^{|A|} \exp(s\cos(\theta_j + m^k_{comb}[j]))}$$

- The bias-injecting component clusters the feature space based on the bias - the *question type*.
- We use a supervised contrastive loss[3] based on the answers - keeps each answer within a question type distinct in the feature space

$$L_{sup-con} = \sum_{j\in B} - \frac{1}{P_j} \sum_{p \in P_j} \log \frac{\exp(x_j^T x_p/\tau)}{\sum_{n\in N_j} \exp(x_j^T x_n/\tau)}$$

Finally, the total loss becomes: $\mathcal{L} = L_{Angular}(m_{comb}) + L_s + L_{sup-con}$

Test-time Ensembling

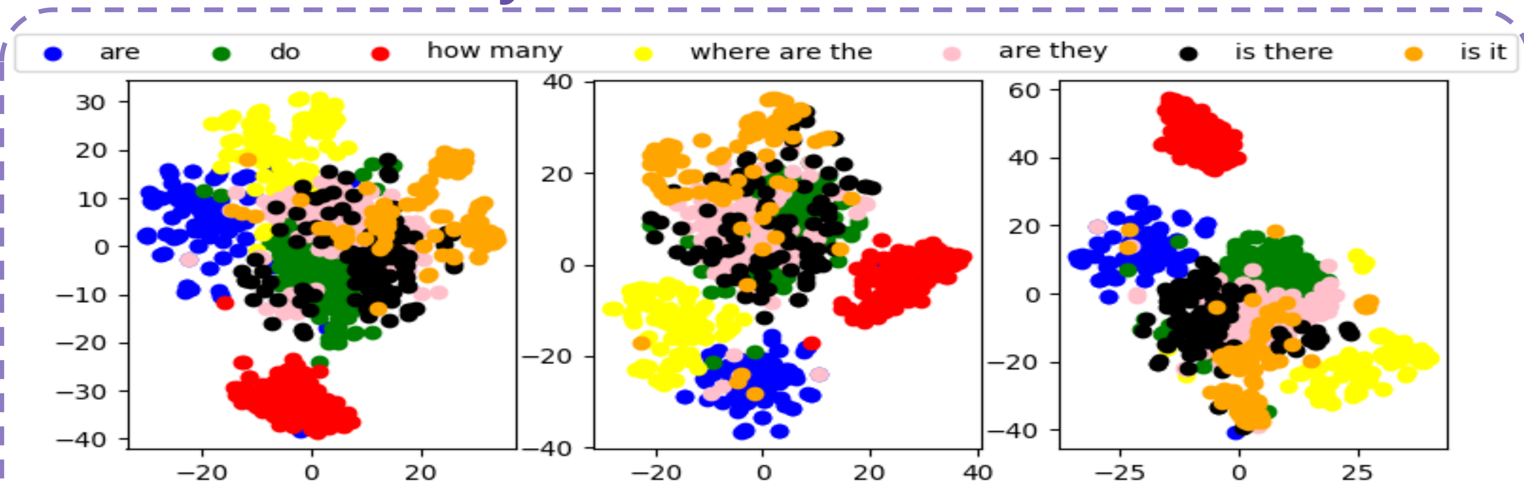


- $\hat{p}_b = softmax(f_b/\tau), \hat{p} = softmax(f)$
- Final prediction: $\hat{p}_{comb} = \alpha.\hat{p}_b + (1-\alpha).\hat{p}.\ \alpha = 0.5$

Performance on VQA-CP v2 and VQA-v2

Method	VQA-CP (OOD)	VQA-V2 (ID)	Diff
UpDn (ERM)	39.74%	63.48%	23.74%
RUBi	47.11%	-	-
LMH	52.15%	56.35%	4.2%
CF-VQA	55.05%	**60.94%**	5.89%
AdaVQA (Margin Loss)	54.02%	46.98%	7.04%
RMLVQA (Ours)	**60.41%**	59.99%	**0.42%**

Further Analysis of Model Performance



Feature space, when trained by (a) the vanilla margin loss, (b) the randomised margin loss, (c) randomised margin loss + bias-injecting component

What color shirt is the man with the ball wearing?

GT: Red Baseline: Blue RMLVQA: Red

References: [1] Deng, et al. "ArcFace: Additive Angular Margin Loss for Deep Face Recognition" International Conference on Machine Learning. CVPR, 2019. [2] Boutros et al, ElasticFace: Elastic Margin Loss for Deep Face Recognition. In CVPR Workshop, 2022. [3] Khosla, Prannay, et al. "Supervised contrastive learning." *Advances in neural information processing systems* 33 (2020): 18661-18673.

Acknowledgements: This work was supported by a research grant (CRG/2021/005925) from SERB, DST, Govt. of India. Abhipsa Basu is supported by the PMRF fellowship