Google DeepMind

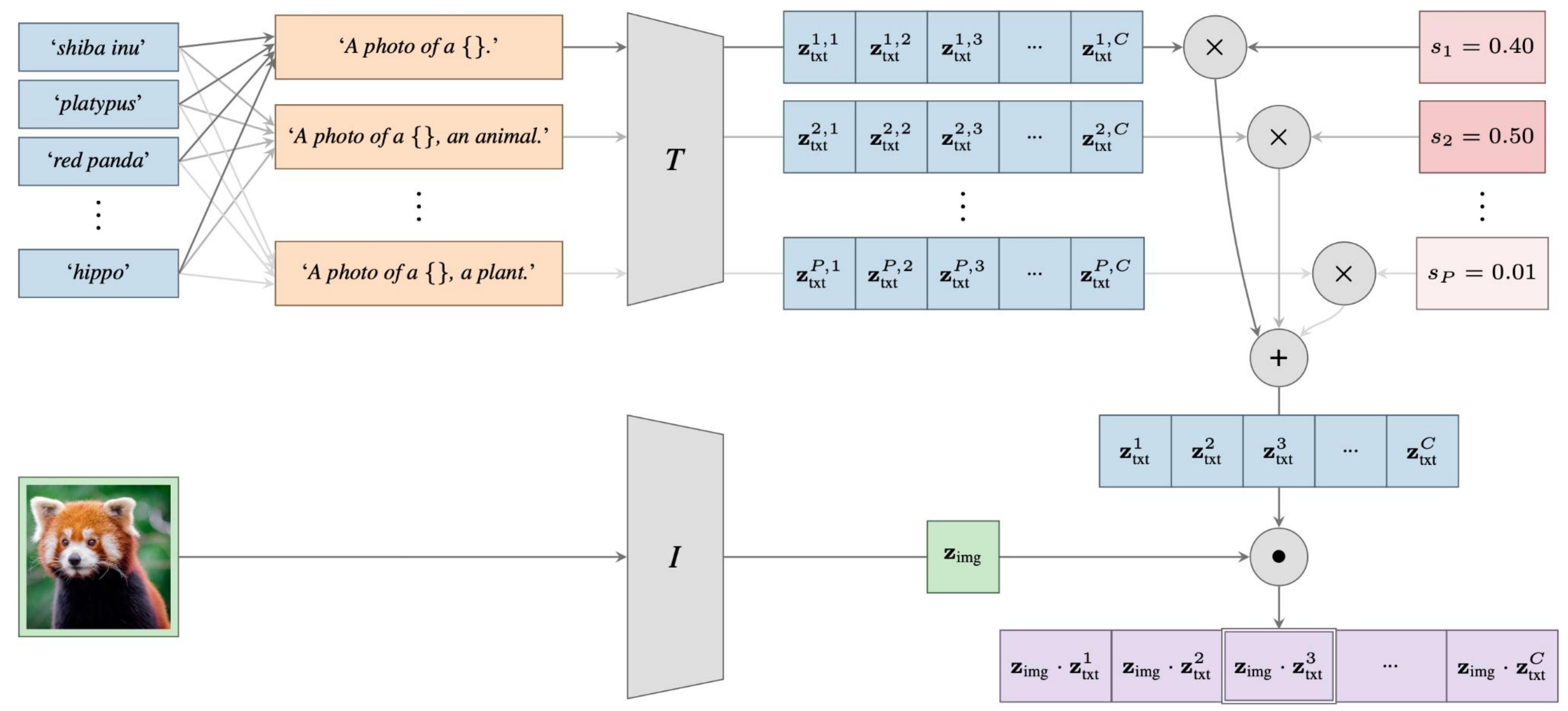
A Simple Zero-shot Prompt Weighting Technique to Improve Prompt Ensembling in Text-Image Models

James Urquhart Allingham *†[1] Jie Ren *[2] Michael W. Dusenberry [2] Xiuye Gu [3] Yin Cui †[4] Dustin Tran [2] Jeremiah Zhe Liu †[3,5] Balaji Lakshminarayanan [2]

Given a **zero-shot image classifier** and a large pool of prompts, we **automatically score the prompts** and ensemble those that are most suitable for a particular downstream dataset, **without access to labeled validation data.**

Scan to access the full paper.

Zero-shot classification with zero-shot prompt ensembling (ZPE)



Logits are calculated by combining text and image representations. The final text representation is a weighted ensemble of representations corresponding to different prompts. The ZPE scores for weighting each prompt are calculated without access to any labeled training data.

Max logit scoring

A simple but biased baseline:

1. $\texttt{logits} = \mathbf{Z}_{\mathrm{img}} \cdot \mathbf{Z}_{\mathrm{txt}}^{\top}$ `# Shape: N x C.`
2. $\texttt{max\_logits} = \max_c \texttt{logits}$ `# Shape: N.`
3. $$s_p = \frac{1}{N}\sum_{n=1}^{N} \texttt{max\_logits}_n$$

Top 10 ImageNet prompts:

*a example of a **person** practicing {} · a example of a **person** using {} · a cropped photo of a {} · a photo of the {} · a photo of the small {} · a cropped photo of the {} · a photo of the large {} · a example of the **person** {} · a example of a **person** {} · a example of {}*

Prompts are biased towards large scores due to:

1. **Word frequency bias** – for prompts containing words that appear frequently in the **pre-training data**.
2. **Spurious concept frequency bias** – for prompts that contain words mapping to common, but irrelevant, concepts in **test images**. E.g., images often contain pictures of people.

ZPE scoring – removing bias

1. $\texttt{logits} = \mathbf{Z}_{\mathrm{img}} \cdot \mathbf{Z}_{\mathrm{txt}}^{\top}$ `# Shape: N x C.`
2. $\texttt{logits}_{\mathrm{pretrain}} = \mathbf{Z}_{\mathrm{pretrain}} \cdot \mathbf{Z}_{\mathrm{txt}}^{\top}$ `# Shape: N' x C.`
3. $$E_{\mathrm{pretrain}} = \frac{1}{N'}\sum_{n=1}^{N'} \texttt{logits}_{\mathrm{pretrain},n}$$ `# Shape: 1 x C.`
4. $$E_{\mathrm{test}} = \frac{1}{N}\sum_{n=1}^{N} \texttt{logits}_n$$ `# Shape: 1 x C.`
5. $\texttt{logits}_{\mathrm{normalized}} = \texttt{logits} - (E_{\mathrm{pretrain}} + E_{\mathrm{test}})/2$
6. $\texttt{max\_logits} = \max_c \texttt{logits}_{\mathrm{normalized}}$ `# Shape: N.`
7. $$s_p = \frac{1}{N}\sum_{n=1}^{N} \texttt{max\_logits}_n$$

Top 10 ImageNet prompts:

itap of a {} · itap of the {} · itap of my {} · a black and white photo of a {} · a high contrast photo of a {} · a low contrast photo of a {} · a photo of a large {} · a photo of the large {} · a black and white photo of the {} · a high contrast photo of the {}

We address long tails by applying `softmax` to the scores, and do (optional) prompt selection using the Median Absolute Deviation to detect outliers.

Zero-shot accuracy results for CLIP-B/16

	ImageNet	ImageNet-A	ImageNet-R	ImageNet-Sketch	ImageNet-V2	**Avg**
class name	63.94	46.01	74.92	44.12	57.97	57.39
'A photo of {}.'	66.37	47.47	73.78	45.84	60.46	58.78
hand-crafted, equal average	68.31	49.13	77.31	47.65	61.83	60.85
pool set, equal average	67.59	49.35	77.33	46.92	61.37	60.51
max-logit scoring	67.63	49.37	77.38	46.95	61.39	60.55
ZPE (weighted average)	68.56	49.61	**77.69**	47.92	**62.23**	61.20
ZPE (prompt selection, ours)	**68.60**	**49.63**	77.62	**47.99**	62.21	**61.21**

	Caltech	Cars	C-10	C-100	DTD	Euro	Food	Flowers	Pets	Resisc	Sun	**Avg**
class name	77.84	61.60	87.30	58.59	44.04	46.90	86.68	63.57	81.38	53.74	60.70	65.67
'A photo of {}.'	82.73	63.45	88.36	65.49	42.93	47.85	88.19	66.84	87.74	55.96	59.95	68.13
hand-crafted, equal average	82.82	64.17	89.10	65.90	45.64	51.60	**88.66**	**71.23**	**88.91**	**65.44**	63.87	70.67
pool set, equal average	83.60	63.16	**89.56**	65.56	45.96	**54.63**	87.79	63.62	80.87	58.70	65.32	68.98
max-logit scoring	83.56	63.16	89.55	65.53	46.28	54.48	87.81	63.70	80.87	59.02	65.39	69.03
ZPE (weighted average)	84.68	64.13	89.34	66.40	**46.54**	53.42	88.50	67.64	86.81	64.18	**66.15**	70.71
ZPE (prompt selection, ours)	**85.54**	**64.62**	89.30	**66.63**	46.28	53.82	88.61	70.17	88.72	64.22	64.70	**71.15**

Other models

	INet	Variants	Fine	**All**
CLIP ResNet-50				
hand-crafted, equal average	59.48	42.52	59.36	55.15
pool set, equal average	58.24	42.17	56.04	52.71
ZPE (weighted average)	59.68	**42.97**	58.79	54.89
ZPE (prompt selection, ours)	**59.90**	42.87	**59.64**	**55.46**
LiT ViT-B/32				
hand-crafted, equal average	68.13	55.25	70.19	66.33
pool set, equal average	66.93	54.51	68.55	64.94
ZPE (weighted average)	68.60	55.67	70.81	66.89
ZPE (prompt selection, ours)	**68.88**	**55.72**	**71.78**	**67.58**
LiT ViT-L/16				
hand-crafted, equal average	78.55	72.65	77.73	76.51
pool set, equal average	77.49	71.74	75.58	74.74
ZPE (weighted average)	78.90	73.11	77.94	76.79
ZPE (prompt selection, ours)	**79.26**	**73.27**	**78.71**	**77.38**

* Equal contribution † Work done while at Google Research [1] University of Cambridge [2] Google DeepMind [3] Google Research [4] Nvidia [5] Harvard University