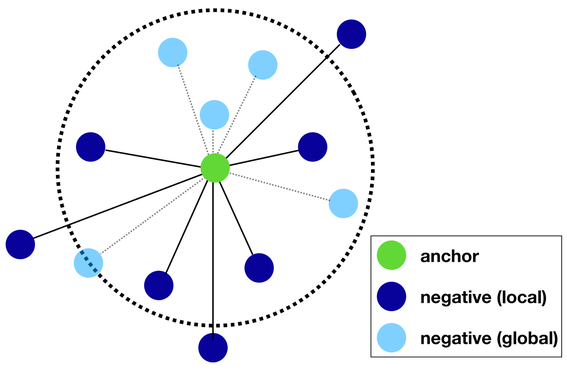
HAL: Improved Text-Image Matching by Mitigating Visual Semantic Hubs.

Fangyu Liu [1**], **Rongtian Ye** [2*], **Xun Wang** [3*], **Shuaipeng Li** [4]
[1]University of Cambridge, UK [2]Aalto University, Finland
[3]Malong Technologies, China [4]SenseTime Research, China
*equal contribution *correspondence to <fl399@cam.ac.uk>

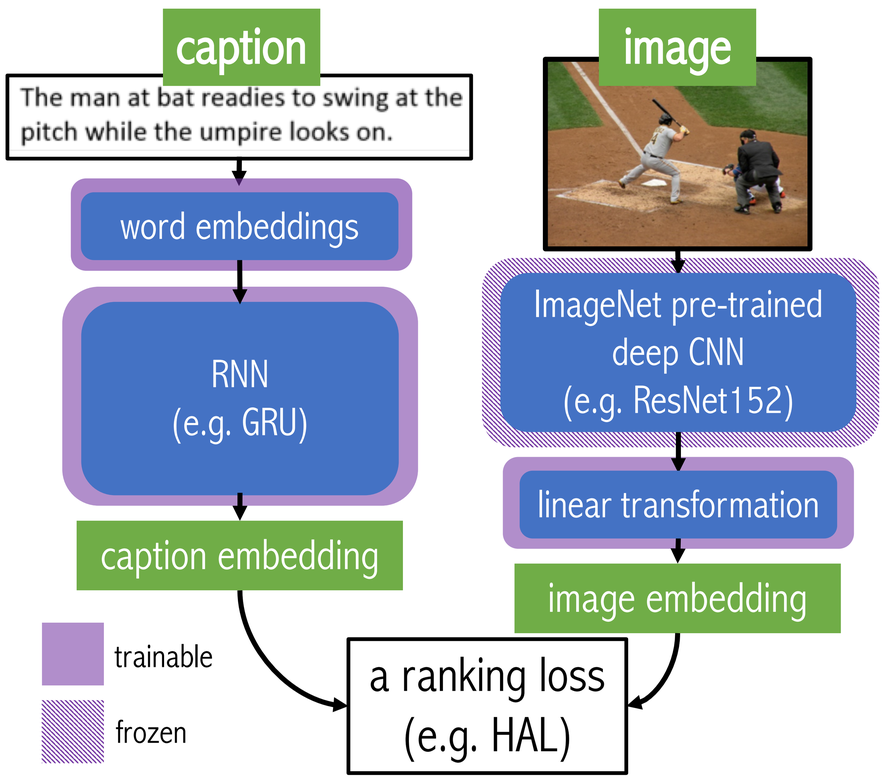
1 Overview



- **The hubness problem** is a general phenomenon in high-dimensional space where a small set of source vectors, dubbed **hubs**, appear too frequently in the neighborhood of target vectors.
- **Our contribution** is a novel training objective (**HAL**) that utilizes both local and global statistics to identify hubs in high-dimensional visual semantic embeddings (VSE). When evaluated on the task of text-image matching, HAL improves R@1 by a maximum of **7.4%** on MS-COCO and **8.3%** on Flickr30k when compared with strong baselines by Faghri et al. [2018] and Lee et al. [2018].
- Code release: `https://github.com/hardyqr/HAL`

2 HAL: Hubness-Aware Loss

The basic framework for learning VSE:



Revisit two (triplet-based) baseline losses:

1) **SUM:**

$$\mathcal{L}_{\text{SUM}} = \sum_{i\in I}\sum_{\bar{t}\in T\setminus\{t\}} [\,\alpha - S_{it} + S_{i\bar{t}}\,]_+ + \sum_{t\in T}\sum_{\bar{i}\in I\setminus\{i\}} [\,\alpha - S_{ti} + S_{t\bar{i}}\,]_+, \tag{1}$$

2) **MAX:**

$$\mathcal{L}_{\text{MAX}} = \sum_{i\in I}\max_{\bar{t}\in T\setminus\{t\}} [\,\alpha - S_{it} + S_{i\bar{t}}\,]_+ + \sum_{t\in T}\max_{\bar{i}\in I\setminus\{i\}} [\,\alpha - S_{ti} + S_{t\bar{i}}\,]_+. \tag{2}$$

Our proposed objective, HAL:

$$\mathcal{L}_{\text{HAL}} = \frac{1}{N}\sum_{i=1}^{N}\Big(\frac{1}{\gamma}\log(1+\sum_{m\neq i} e^{\gamma W_{mi}(S_{mi}-\epsilon)}) + \frac{1}{\gamma}\log(1+\sum_{n\neq i} e^{\gamma W_{in}(S_{in}-\epsilon)}) - \log(1+W_{ii}S_{ii})\Big) \tag{3}$$

A simple gradient analysis on HAL:

$$w^+ = \left|\frac{\partial \mathcal{L}_{\text{HAL}}}{\partial {S_{ij}}^+}\right| = \frac{W_{ij}}{1+W_{ij}S_{ij}} \text{ if } i=j,$$

$$w^- = \left|\frac{\partial \mathcal{L}_{\text{HAL}}}{\partial {S_{ij}}^-}\right| = \underbrace{\frac{W_{ij}e^{\gamma W_{ij}(S_{ij}-\epsilon)}}{1+\sum_{m\neq j} e^{\gamma W_{mj}(S_{mj}-\epsilon)}}}_{\text{weighted by image modality}} + \underbrace{\frac{W_{ij}e^{\gamma W_{ij}(S_{ij}-\epsilon)}}{1+\sum_{n\neq i} e^{\gamma W_{in}(S_{in}-\epsilon)}}}_{\text{weighted by text modality}}. \tag{4}$$

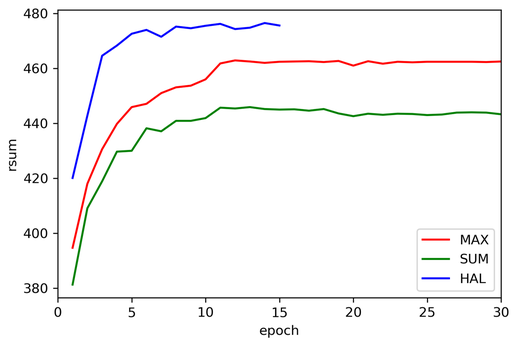
$*$ Gradient analysis suggests that HAL, by its formulation, automatically utilizes local statistics (within a mini-batch) for sample weighting. W_{ij}s contain information obtained from global statistics (across the whole training set). Please refer paper for details.

Why HAL?

- a soft weighting scheme utilizing information from hubs, taking all samples into account; both local and global sample distributions are considered → effectively locate hubs/hard samples
- compared to MAX: avoid *pseudo* hard negatives → robust to noisy labels; consider all samples → more information utilized
- compared to SUM: SUM treats all samples equivalently (no weighting at all), neglecting rich relationships among samples

3 Results (on MS-COCO)

HAL vs. SUM/MAX:



#	architecture	loss	image→text					text→image					
			R@1	R@5	R@10	Med r	Mean r	R@1	R@5	R@10	Med r	Mean r	rsum
1.1	GRU+ResNet152	SUM	53.2	85.0	93.0	**1.0**	3.9	41.9	77.2	88.0	**2.0**	8.7	438.3
1.2		MAX	58.7	88.2	94.8	**1.0**	3.2	45.0	78.9	88.6	**2.0**	8.6	454.2
1.3		HAL	**64.4**	89.2	94.9	**1.0**	3.0	46.3	78.8	88.3	**2.0**	7.9	462.0
1.4		HAL+MB	64.0	**89.9**	**95.7**	**1.0**	**2.8**	**46.9**	**80.4**	**89.9**	**2.0**	**6.1**	**466.7**

HAL vs. State-of-the-art:

#	architecture	image→text					text→image					
		R@1	R@5	R@10	Med r	Mean r	R@1	R@5	R@10	Med r	Mean r	rsum
2.1	[Kiros et al., 2015] (ours)	49.9	79.4	90.1	2.0	5.2	37.3	74.3	85.9	**2.0**	10.8	416.8
2.2	[Vendrov et al., 2016]	46.7	-	88.9	2.0	5.7	37.9	-	85.9	**2.0**	8.1	-
2.3	[Huang et al., 2017a]	53.2	83.1	91.5	**1.0**	-	40.7	75.8	87.4	**2.0**	-	431.8
2.4	[Liu et al., 2017]	56.4	85.3	91.5	-	-	43.9	78.1	88.6	-	-	443.8
2.5	[You et al., 2018]	56.3	84.4	92.2	**1.0**	-	45.7	81.2	90.6	**2.0**	-	450.4
2.6	[Wehrmann, 2018]	57.8	87.9	95.6	**1.0**	3.3	44.2	80.4	90.7	**2.0**	**5.4**	456.6
2.7	[Faghri et al., 2018]	58.3	86.1	93.3	**1.0**	-	43.6	77.6	87.8	**2.0**	-	446.7
2.8	[Faghri et al., 2018] (ours)	60.5	89.6	94.9	**1.0**	3.1	46.1	79.5	88.7	**2.0**	8.5	459.3
2.9	[Liu and Ye, 2019]	58.3	89.2	95.4	**1.0**	3.1	45.0	80.4	89.6	**2.0**	7.2	457.9
2.10	[Wu et al., 2019]	64.3	89.2	94.8	**1.0**	-	48.3	81.7	**91.2**	**2.0**	-	469.5
2.11	GRU+RN152 + HAL	65.4	90.4	96.4	**1.0**	2.5	47.4	80.6	89.0	**2.0**	7.3	469.2
2.12	GRU+RN152 + HAL + MB	**66.3**	**91.7**	**97.0**	**1.0**	**2.4**	**48.7**	**82.1**	90.8	**2.0**	5.6	**476.6**
2.13	[Lee et al., 2018]	70.9	94.5	97.8	-	-	56.4	**87.0**	**93.9**	-	-	500.5
2.14	[Lee et al., 2018] + HAL	**78.3**	**96.3**	**98.5**	1.0	2.6	**60.1**	86.7	92.8	1.0	5.8	**512.7**