

# Im2Text and Text2Im: Associating Images and Texts for Cross-Modal Retrieval (Supplementary)

Yashaswi Verma

<http://researchweb.iiit.ac.in/~yashaswi.verma/>

C. V. Jawahar

<http://www.iiit.ac.in/~jawahar/>

CVIT

IIIT-Hyderabad, India

<http://cvit.iiit.ac.in>

Here, first we briefly discuss the WSABIE algorithm [1], and then present the proposed extension of WSABIE to adapt it for captions.

## WSABIE

WSABIE (Web Scale Annotation by Image Embedding) learns a mapping space where both images and annotations (e.g. labels) are represented. The mapping functions for both the modalities are learnt jointly by minimizing the WARP (Weighted Approximate-Rank Pairwise) loss, that is based on optimizing precision at  $k$ . Each image is represented by  $x \in \mathbb{R}^P$ , and each annotation  $i \in \mathcal{Y} = \{1, \dots, Y\}$ , where  $Y$  is the (fixed) vocabulary size. Then, a mapping is learnt from image feature space to the joint space  $\mathbb{R}^P$ :

$$\Phi_I(x) : \mathbb{R}^P \rightarrow \mathbb{R}^P. \quad (1)$$

while jointly learning a mapping function for annotations:

$$\Phi_W(i) : \{1, \dots, Y\} \rightarrow \mathbb{R}^P. \quad (2)$$

Both these mappings are chosen to be linear; i.e.,  $\Phi_I(x) = Vx$ , and  $\Phi_W(i) = W_i$  where  $W_i$  indices the  $i^{\text{th}}$  column of a  $P \times Y$  matrix. The goal is to learn the possible annotations of a given image such that the highest ranked ones best describe the semantic content of the image. For this, the following model is considered:

$$f_i(x) = \Phi_W(i)^T \Phi_I(x) = W_i^T Vx, \quad (3)$$

where the possible annotations  $i$  are ranked according to the magnitude of  $f_i(x)$  in descending order. This family of models have constrained norm:

$$\begin{aligned} \|V_i\|_2 &\leq \lambda, i = 1, \dots, p, \\ \|W_i\|_2 &\leq \lambda, i = 1, \dots, Y. \end{aligned} \quad (4)$$

which acts as a regularizer. Algorithm 1 shows the pseudo-code for learning model variables using a stochastic gradient descent algorithm that minimizes WARP loss (where  $L(k) = \sum_{j=1}^k \alpha_j$ , with  $\alpha_j = \frac{1}{j}$ ).

**Algorithm 1** WSABIE Algorithm

---

**Require:** labeled data  $(x_i, y_i), y_i \in \{1, \dots, Y\}$

**repeat**

  Pick a random labeled example  $(x_i, y_i)$

  Let  $f_{y_i}(x_i) = \Phi_W(y_i)^T \Phi_I(x_i)$

  Set  $N = 0$

**repeat**

    Pick a random annotation  $\bar{y} \in \{1, \dots, Y\} \setminus y_i$ .

    Let  $f_{\bar{y}}(x_i) = \Phi_W(\bar{y})^T \Phi_I(x_i)$

$N = N + 1$

**until**  $f_{\bar{y}}(x_i) > f_{y_i}(x_i) - 1$  or  $N \geq Y - 1$

**if**  $f_{\bar{y}} > f_{y_i}(x_i) - 1$  **then**

    Make a gradient step to minimize:

$L(\lfloor \frac{Y-1}{N} \rfloor) |1 - f_{y_i}(x_i) + f_{\bar{y}}(x_i)|_+$

    Project weights to enforce constraints in Eq. 4.

**end if**

**until** validation error does not improve.

---

**Adapting WSABIE for Captions**

In case of captions, we have a (training) set of captions  $\mathcal{C} = \{c_i\}$  rather than a fixed annotation vocabulary. In order to adapt WSABIE for captions, we modify the feature mapping given in Eq. 2 such that instead of learning a separate parameter vector for each annotation, we learn a single parameter matrix for all the captions. Given a caption  $c \in \mathcal{C}$  represented by  $y \in \mathbb{R}^q$ , a mapping is learnt from caption feature space to the joint space  $\mathbb{R}^P$ :

$$\Phi_Z(y) : \mathbb{R}^q \rightarrow \mathbb{R}^P, \quad (5)$$

where  $Z$  is a  $P \times q$  matrix. Now, given a set of captions, the goal is to learn the possible caption(s) of a given image such that the highest ranked ones best describe the semantic content of the image. For this, the following model is considered:

$$g_y(x) = \Phi_Z(y)^T \Phi_I(x) = y^T Z^T V x. \quad (6)$$

Similar to Eq. 4, this family of models have constrained norm:

$$\begin{aligned} \|V_i\|_2 &\leq \lambda, i = 1, \dots, p, \\ \|Z_i\|_2 &\leq \lambda, i = 1, \dots, q. \end{aligned} \quad (7)$$

which acts as a regularizer. Algorithm 2 shows the pseudo-code for learning the model variables using a stochastic gradient descent algorithm. It is similar to Algorithm 1 except that instead of randomly picking an annotation from vocabulary, now we randomly pick a caption from the training set consisting of image-caption pairs.

**Algorithm 2** Adapted WSABIE Algorithm for Captions

---

**Require:** labeled data  $(x_i, c_i)$ ,  $y$  is a feature vector representing caption  $c \in \mathcal{C}$

---

**repeat**Pick a random labeled example  $(x_i, c_i)$ Let  $g_{y_i}(x_i) = \Phi_Z(y_i)^T \Phi_I(x_i)$ Set  $N = 0$ **repeat**Pick a random caption  $\bar{c} \in \mathcal{C} \setminus c_i$ .Let  $g_{\bar{y}}(x_i) = \Phi_Z(\bar{y})^T \Phi_I(x_i)$  $N = N + 1$ **until**  $g_{\bar{y}}(x_i) > g_{y_i}(x_i) - 1$  or  $N \geq |\mathcal{C}| - 1$ **if**  $g_{\bar{y}} > g_{y_i}(x_i) - 1$  **then**

Make a gradient step to minimize:

$$L(\lfloor \frac{|\mathcal{C}|-1}{N} \rfloor) |1 - g_{y_i}(x_i) + g_{\bar{y}}(x_i)|_+$$

Project weights to enforce constraints in Eq. 7.

**end if****until** validation error does not improve.

## References

- [1] Jason Weston, Samy Bengio, and Nicolas Usunier. WSABIE: Scaling up to large vocabulary image annotation. In *IJCAI*, 2011.