Supplementary materials: Linking Image and Text with 2-Way Nets

April 12, 2017

Contents

List of Figures

•A white dog is sleeping on a street and a bicycle
•The white dog lays next to the bicycle on the sidewalk.
•Dog snoozing by a bike on the edge of a cobblestone street.
•A picture of a dog laying on the ground.
•A puppy rests on the street next to a bicycle.

•A Subway sandwich, potato chips, raisins, and a mug of coffee.
•A sandwich on a sesame seed bun next to a pile of french fries and a cup of ketchup.
•A half eaten sandwich sitting on a wrapper.
•A half a subway sandwich sitting on top of a paper.
•A Subway Sandwich with chips, raisins and a coffee cup.

•A man riding a skateboard down a street.
•The man is riding his skateboard down the street.
•A young boy riding a twisty skateboard down the street.
•A young boy is riding on a two wheeled skateboard.
•A man with a hat on riding his skateboard.

•There's a stuffed bear sitting on a table.
•A stuffed bear on a clear glass table.
•A brown teddy bear sitting on top of an answering machine.
•Two brown stuffed teddy bears sitting on a dresser.
•A brown teddy bear sitting on top of a phone.

•Many cars drive and are parked on a road.
•A red car driving down a mountain road next to a herd of sheep.
•A car driving down a quiet road during a sunny day.
•The car is driving down the road and the sheep is in the road.
•A van with words on it is driving down the street.

Figure 1: Examples of results from the image description task on the COCO dataset. Five sentences are produced for each image. The sentences are chosen by matching their 2-Way Net representation with the produced image's representations.

1 Appendix A - Image-Sentence matching examples

This section contains results from the image-sentence experiments. Examples of both the image query and image describe tasks are presented. For the image description task, each image is shown alongside five sentences with the highest matching score. Sentences which are correct, according to the specific dataset, are marked in green. For the image query task, for each sentence, five images with the best matching score are shown. Examples are shown for Flickr8k [2], Flickr30k [4] and COCO [3] datasets.

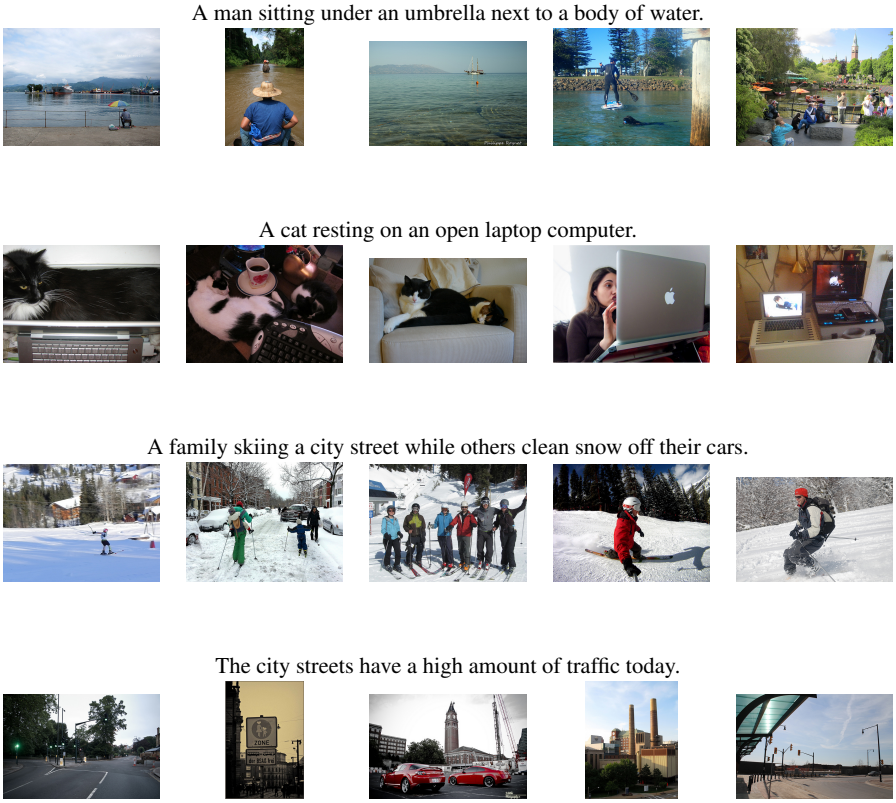


Figure 2: Examples of results from the image search task on the COCO dataset. Each sentence is matched against all images and the top five are presented.

- A man in tan capris , brown sandals , and a white t-shirt , crouches on the trunk of a tree .
- A man in a white shirt and khaki pants crouches on a fallen tree trunk .
- A man is crouched on top of a horizontal tree .
- A man crouched in the bare branches of a fallen tree.
- A man sitting in the middle of branches on a fallen tree.

- Black and white dog jumping up in the snow.
- Shaggy little dog jumps in the snow.
- Three dogs run in the snow.
- a black and white dog jumping in the air surrounded by snow.
- The white and black dog leaped into the air and off the snowy ground.

- An ice hockey goalkeeper in a red and blue strip is on his knees in front of the goal.
- A hockey player in blue and red guarding the goal.
- The ice hockey goal keeper is dressed in a red strip.
- A goalie is crouching in a defensive position in front of the goal.
- A goalie defending the goal in a hockey game.

- A group of various people stand outside of a store called C̈entral Market.
- People stand outside of a market.
- People are strolling around a market.
- People gathered in front of a store named Central market.
- A group of shoppers walk near the produce section of Sunlight Farms store.

- A red dump truck is on scene beside some buildings.
- Four men are performing work in a dirt lot , near a building , using a dump truck.
- A picture of two workers next to a semi truck standing outside the fence of a refinery.
- Two men work next to a cement truck.
- A truck with an advertising billboard stuck in the middle of the road.

Figure 3: Examples of results from the image description task on the Flickr30k dataset. See Fig. 1 for details.

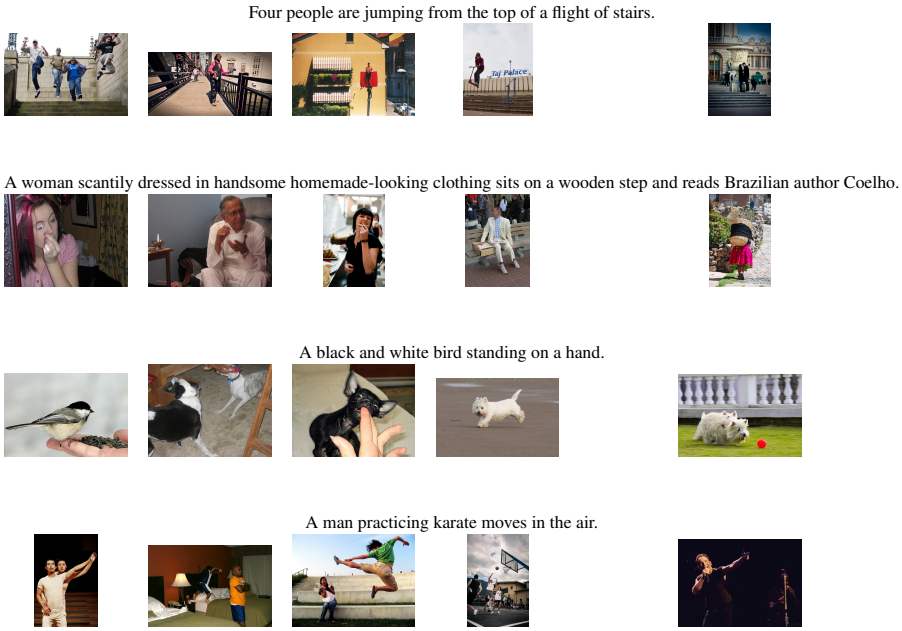
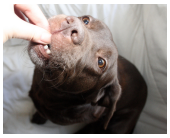


Figure 4: Examples of results from the image search task on the Flickr30k dataset. See Fig. 2 for details.

•Two people are walking towards a group of people on the beach at sunset.
•A bunch of people on the beach at sunset.
•Two people walking along the beach at sunset toward a group of people.
•Two people walking along the shore at sunset.
•Two people walking on the beach at sunset.

•Two adults and two children sitting on rocks for a picture.
•The four people sit on a pile of rocks.
•Two adults and two children pose on a pile of rocks.
•Three people sitting on a cliff.
•A family sitting on rocks posing for the camera

•A person is pulled in a cart by a donkey.
•Dogs pulling a sled in a sled race.
•A donkey pulling a cart with a boy in it takes a brake.
•A mule pulling a carriage with a man inside along a dirt path.
•Almost 2 dozen people are riding on the outside of a cart down a road.



•A dog is chewing on someone's finger.
•A dog bites someone's finger.
•A brown dog looks up while a person 's fingers are in its mouth.
•The dog 's mouth is open like he is yawning.
•There is a dog laying on the floor and another standing over him , licking his ear.

•Four swimmers in the crystal blue ocean with yellow surfboards.
•A child is pushed by a large wave while holding his yellow surfboard.
•The man turns with the wave on his surfboard .
•A person on a surfboard in the waves , holding onto a tether.
•A person rides the waves on a surfboard.

Figure 5: Examples of results from the image search task on the Flickr8k dataset. See Fig. 2 for details.

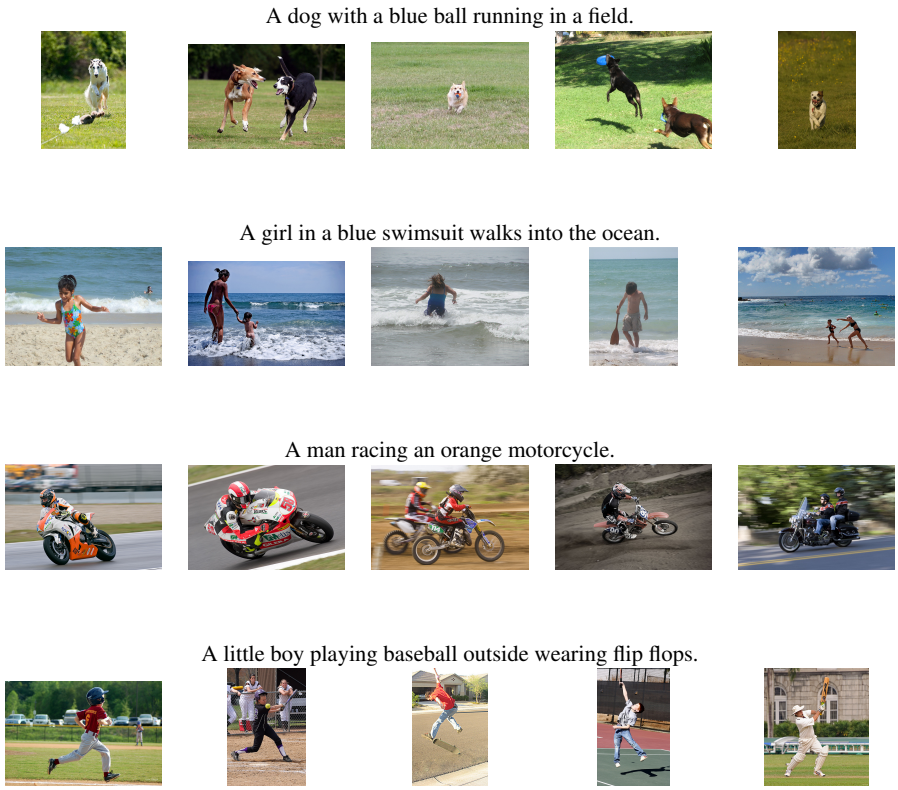


Figure 6: Examples of results from the image search task on the Flickr8k dataset. See Fig. 2 for details.

2 Appendix B - Lemma's Proofs

This section contains proofs for the lemmas described in the paper

Lemma 1. *Let $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$ denote two paired lists of n matching samples from two random variables with zero mean and σ_x^2 and σ_y^2 variances. Then, the correlation between the two n dimensional samples x and y equals $\frac{\sigma_x}{2\sigma_y} + \frac{\sigma_y}{2\sigma_x} - \frac{\|x-y\|^2}{2n\sigma_x\sigma_y}$.*

Proof. Given two n-dimensional vectors x and y we consider the squared Euclidean distance

$$\|x-y\|^2 = \sum_{j=1}^{n}(x_j^2) + \sum_{j=1}^{n}(y_j^2) - 2\sum_{j=1}^{n}(x_j y_j)$$

Thus:

$$\sum_{j=1}^{n}(x_j y_j) = \frac{n\sigma_x^2}{2} + \frac{n\sigma_y^2}{2} - \frac{\|x-y\|^2}{2} \tag{1}$$

For zero mean variables, the correlation between x and y is given by $c = \frac{1}{n}\frac{\sum_{j=1}^{n}(x_j y_j)}{\sigma_x\sigma_y}$. Combining with 1 results in what had to be proven. □

Lemma 2. *Given two matching hidden layers, h_j and $\hat{h_j}$ with m neurons each. a_k is the activation vector of neuron k from h_j with standard deviation σ_{a_k} and b_k is the activation vector of neuron k from $\hat{h}_j$ with standard deviation σ_{b_k}. Each vector is produced by feeding a batch of samples of size n from views x and y through channels H and $\hat{H}$ respectively. The sum of correlations C is bounded by:*

$$\begin{aligned}\sum_{k=1}^{m} C_k \geq \frac{1}{2}\sum_{k=1}^{m}(\frac{\sigma_{a_k}^2 + \sigma_{b_k}^2}{\sigma_{a_k}\sigma_{b_k}}) \\ - \frac{1}{2n}\sum_{k=1}^{m}\|a_k - b_k\|^2 \sum_{k=1}^{m}\sigma_{a_k}^{-1}\sigma_{b_k}^{-1}\end{aligned} \tag{2}$$

Proof. From lemma 1, we get:

$$\sum_{k=1}^{m} C_k = \frac{1}{2}\sum_{k=1}^{m}(\frac{\sigma_{a_k}^2 + \sigma_{b_k}^2}{\sigma_{a_k}\sigma_{b_k}}) - \frac{1}{2n}\sum_{k=1}^{m}(\frac{\|a_k - b_k\|^2}{\sigma_{a_k}\sigma_{b_k}}) \tag{3}$$

We will define $G_m = \sum_{k=1}^{m} \|a_k - b_k\|^2$ and $f_k = \sigma_{a_k}^{-1}\sigma_{b_k}^{-1}$. Using Abel transform:

$$\begin{aligned}
\sum_{k=1}^{m} \frac{\|a_k - b_k\|^2}{\sigma_{a_k}\sigma_{b_k}} &= f_m G_m - \sum_{k=1}^{m-1} G_k f_{k+1} + \sum_{k=1}^{m-1} G_k f_k \\
&\leq f_m G_m + \sum_{k=1}^{m-1} G_k f_k \\
&\leq f_m G_m + G_m \sum_{k=1}^{m-1} f_k = G_m \sum_{k=1}^{m} f_k \\
&= \sum_{k=1}^{m} \|a_k - b_k\|^2 \sum_{k=1}^{m} \sigma_{a_k}^{-1}\sigma_{b_k}^{-1} \qquad (4)
\end{aligned}$$

Note that both $\sigma_{a_k}\sigma_{b_k}$ and $\|a_k - b_k\|^2$ are positive for all k which makes the above inequalities valid. Inserting 4 in 3 results in what had to be proven. □

Lemma 3. *Assume that u_i and v_i are drawn from a multivariate normal distribution with zero mean and the identity covariance matrix, such that the correlation between $u_i(k)$ and $v_i(k)$ for all k is $\rho_k = \rho$. Then, $E\left(|s_i \cap \hat{s}_i|\right) = d\left[\frac{1}{4} + \frac{\sin^{-1}\rho}{2\pi}\right]$.*

Proof. To estimate the size of c, let us look at the quadrant probability p of $u_i(k)$ and $v_i(k)$ which is given analytically by [1],

$$p = P(u_i(k) > 0, v_i(k) > 0) = \frac{1}{4} + \frac{\sin^{-1}\rho}{2\pi}$$

Given that the variables in $u_i(k)$ and $v_i(k)$ are drawn independently, the probability of $P(|c| = t)$ has a binomial distribution with probability p, thus the mean of the size of c is equal to $E(|c|) = dp = d\left[\frac{1}{4} + \frac{\sin^{-1}\rho}{2\pi}\right]$. □

References

[1] Oliver D Anderson, A Stuart, and JK Ord. Kendall's advanced theory of statistics, volume 1: Distribution theory., 1988.

[2] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013.

[3] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014.

[4] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.