Unsupervised Discovery of Multimodal Links in Multi-image, Multi-sentence Documents

Jack Hessel  Lillian Lee  David Mimno
Cornell University
{jhessel, llee}@cs.cornell.edu  mimno@cornell.edu

Abstract
Images and text co-occur constantly on the web, but explicit links between images and sentences (or other intra-document textual units) are often not present. We present algorithms that discover image-sentence relationships without relying on explicit multimodal annotation in training. We experiment on seven datasets of varying difficulty, ranging from documents consisting of groups of images captioned post hoc by crowdworkers to naturally-occurring user-generated multimodal documents. We find that a structured training objective based on identifying whether collections of images and sentences co-occur in documents can suffice to predict links between specific sentences and specific images within the same document at test time.

1 Introduction
Images and text act as natural complements on the modern web. News stories include photographs, product listings show multiple images providing detail for online shoppers, and Wikipedia pages include maps, diagrams, and pictures. But the exact matching between words and images is often left implicit. Algorithms that identify document-internal connections between specific images and specific passages of text could have both immediate and long-term promise. On the user-experience front, alt-text for vision-impaired users could be produced automatically (Wu et al., 2017) via intra-document retrieval, and user interfaces could explicitly link images to descriptive sentences, potentially improving the reading experience of sighted users. Also, in terms of improving other applications, the text in multimodal documents can be viewed as a noisy form of image annotation: inferred image-sentence associations can serve as training pairs for vision models, particularly in domains lacking readily-available labeled data.

In this work, we develop unsupervised models that learn to identify multimodal within-document links despite not having access to supervision at the individual image/sentence level during training. Rather, the training documents contain multiple images and multiple sentences1 that are not aligned, as illustrated in Figure 1.

Our intra-document setting poses challenges beyond those encountered in the usual cross-modal retrieval framework, wherein “documents” generally consist of a single image associated with a single piece of text, e.g., an image caption. For the longer documents we consider, a sentence may have many corresponding images or no corresponding images, and vice versa. Furthermore, we expect that images within documents will be, on average, more similar than images across documents, thus making disambiguation more difficult than in the usual one-image/one-sentence case.

Our approach for this difficult setting is ranking-based: we train algorithms to score image collec-

---

1Any discrete textual unit could be used, such as n-grams or paragraphs. We focus on sentences because there exist public sentence-level datasets that we can use for evaluation.
tions and sentence collections that truly co-occur more highly than image collections and sentence collections that do not co-occur. The matching functions we consider predict a latent similarity-weighted bipartite graph over a document’s images and sentences; at test time, we evaluate this internal bipartite graph representation learned by our models for the task of intra-document link prediction.

We work with a variety of datasets (one of which we introduce), ranging from concatenations of individually-captioned images to organically-multimodal documents scraped from noisy, user-generated web content. Despite having no supervision at the individual image-sentence level, our algorithms perform well on the same-document link prediction task. For example, on a visual storytelling dataset, we achieve 90+ AUC, even in the presence of a large number of sentences that do not correspond to any images in the document. Similarly, for organically-multimodal web data, we are able to surpass object-detection baselines by a wide margin, e.g., for a step-by-step recipe dataset, we improve precision by 20 points on link prediction within documents by leveraging document-level co-occurrence during training.

We conclude by using our algorithm to discover links within a Wikipedia image/text dataset that lacks ground-truth image-sentence links. While the predictions are imperfect, the algorithm qualitatively identifies meaningful patterns, such as matching an image of a dodo bird to one of two sentences (out of 100) in the corresponding article that mention “dodo”.

2 Task Formulation

We assume as given a set of documents where each document \( d_i = (S_i, V_i) \) consists of a set \( S_i \) of \( n_i \) sentences and a set \( V_i \) of \( m_i \) images.\(^2\) For example, \( d_i \) could be an article about Paris with \( n_i = 100 \) sentences and \( m_i = 3 \) images of, respectively, the Eiffel Tower, the Arc de Triomphe, and a map of Paris. For each \( d_i \), we are to predict an alignment — where some sentences or images may not be aligned to anything — represented by a (potentially sparse) bipartite graph

\[
\mathcal{M}_i \in \mathbb{R}^{n_i \times m_i}, \quad \forall i \in \{1, 5\}
\]

We produce a dense sentence-to-image association matrix \( \hat{\mathcal{M}}_i \) in which each entry is the confidence that there is an (undirected) edge between the corresponding nodes. Applying different thresholding strategies to \( \hat{\mathcal{M}}_i \)’s values yields different alignment graphs.

Evaluation. When we have ground-truth alignment graphs for test documents, we evaluate the correctness of the association matrix \( \hat{\mathcal{M}}_i \) predicted by our algorithms according to two metrics: AUROC (henceforth AUC) and precision-at-C \( \left( p@C \right) \). AUC, commonly used in evaluating link prediction (see Menon and Elkan (2011)) is the area under the curve of the true-positive/false-positive rate produced by sweeping over possible confidence thresholds; random is 50, perfect is 100. \( p@C \) measures the accuracy of the algorithm’s most confident \( C \) predicted edges (in our case, the most confident edges correspond to the largest entries in \( \hat{\mathcal{M}}_i \)). This metric models cases where only a small number of high-confidence predictions need be made per document. We evaluate using \( C \in \{1, 5\} \).

3 Models

Our algorithm is inspired by work in cross-modal retrieval (Rasiwasia et al., 2010; Hodosh et al., 2013; Costa Pereira et al., 2014; Kiros et al., 2014b). Instead of operating at the level of individual images/sentences, however, our training objective encourages image sets and sentence sets appearing in the same document to be more similar than non-co-occurring sets.

3.1 Alignment Model and Loss Function

We assume that the dimensionality \( d_{\text{multi}} \) of the multimodal text-image space is predetermined.

Extracting sentence representations. We pass the words in each sentence through a 300D word-embedding layer initialized with GoogleNews-pretrained word2vec embeddings (Mikolov et al., 2013). We then pass the sequence of word vectors to a GRU (Cho et al., 2014) and extract and L2-normalize a \( d_{\text{multi}} \)-dimensional sentence representation from the final hidden state.
Extracting image representations. We first compute a representation for each image using a convolutional neural network (CNN). The network’s output is then mapped via affine projection to $\mathbb{R}^{d_{\text{multi}}}$ and L2-normalized.

Correspondence prediction. The result of running the two steps above on an image-set/text-set pair $(S, V)$ is $|S| + |V|$ vectors, all in $\mathbb{R}^{d_{\text{multi}}}$. From these, we compute the similarity matrix $\hat{M} \in \mathbb{R}^{|S| \times |V|}$, where the $(j, k)^{th}$ entry is the cosine similarity between the $j^{th}$ sentence vector and the $k^{th}$ image vector.

Training Objective. We train under the assumption that co-occurring image-set/sentence-set pairs should be more similar than non-co-occurring image-set/sentence-set pairs. We hope that use of this document-level objective will produce an $\hat{M}_i$ offering reasonable intra-document information at test time, even though such information is not available at training time.

The training process is modulated by a similarity function $\text{sim}(S, V)$ that measures the similarity between a set of sentences and a set of images by examining the entries of the individual image/sentence similarity matrix $\hat{M}_i$ (specific definitions of $\text{sim}(S, V)$ are proposed in §3.2). We use a max-margin loss with negative sampling: we iterate through true documents $d_i = (S_i, V_i)$, and negatively sample at the document level a set of $b$ sets of images that did not co-occur with $S_i$, $\mathcal{V}' = \{V_i', \ldots, V'_b\}$, and a set of $b$ sets of sentences that did not co-occur with $V_i$, $\mathcal{S}' = \{S_i', \ldots, S'_b\}$.

We then compute a loss for $(S_i, V_i)$ by comparing the true similarities to the negative-sample similarities. We find that hard-negative mining (Dalal and Triggs, 2005; Schroff et al., 2015; Faghri et al., 2018), the technique of selecting the negative cases that maximally violate the margin within the minibatch, performs better than simple averaging. The loss for a single positive example is:

$$L(S_i, V_i) = \max_{V' \in \mathcal{V}'} h\left(\text{sim}(S_i, V_i), \text{sim}(S_i, V')\right) + \max_{S' \in \mathcal{S}'} h\left(\text{sim}(S_i, V_i), \text{sim}(S', V_i)\right)$$

for hinge loss $h_\alpha(p, n) = \max(0, \alpha - p + n)$, where we set margin $\alpha = 0.2$ (Kiros et al., 2014a; Faghri et al., 2018).

3.2 Similarity Functions

We explore several functions for measuring how similar a set of $n$ sentences $S$ is to a set of $m$ images $V$. All similarity functions convert the matrix $\hat{M} \in \mathbb{R}^{n \times m}$ corresponding to $\langle S, V \rangle$ into a bipartite graph based on the magnitude of the entries. The functions differ in how they determine which entries $\hat{M}_{ij}$ correspond to edges and edge weights.

Dense Correspondence (DC). The DC function assumes a dense correspondence between images and sentences; each sentence must be aligned to its most similar image, and vice versa, regardless of how small the similarity might be:

$$\text{sim}(S, V) = \frac{1}{n} \sum_{i=0}^{n} \max_{j} \hat{M}_{i,j} + \frac{1}{m} \sum_{j=0}^{m} \max_{i} \hat{M}_{i,j}$$

The underlying assumption of this function can clearly be violated in practice: sentences can have no image, and images no sentence.

Top-K (TK). Instead of assuming that every sentence has a corresponding image and vice versa, in this function only the top $k$ most likely sentence $\Rightarrow$ image (and image $\Rightarrow$ sentence) edges are aligned. This process mitigates the effect of non-visual sentences by allowing algorithms to align them to no image. We discuss choices of $k$ for particular experimental settings in §4.1.

Assignment Problem (AP). We may wish to consider the image-sentence alignment task as a bipartite linear assignment problem (Kuhn, 1955), such that each image/sentence in a document has at most one association. Each time we compute $\text{sim}(S, V)$ in the forward pass of our models, we solve the integer programming problem of maximizing $\sum_{i,j} \hat{M}_{ij} x_{ij}$ subject to the constraints:

$$\forall i, \sum_{j} x_{ij} \leq 1; \forall j, \sum_{i} x_{ij} \leq 1; \forall i, j, x_{ij} \in \{0, 1\}$$

Despite involving a discrete optimization step, the model remains fully differentiable. Our forward pass uses tensorflow’s python interface, t.f. py_func, and the lapjv implementation of the JV algorithm (Jonker and Volgenant, 1987) to solve the integer program itself. Given the solution $x^*_{ij}$, we compute (and backpropagate gradients through) the similarity function $\text{sim}(S, V) = \left(\sum_{i,j} \hat{M}_{ij} x^*_{ij}\right)/r$ where $r$ is the number of non-zero $x^*_{ij}$. Should we want to impose an upper bound

$^4$In some experiments, we use pre-computed image features from a pre-trained CNN (Sharif Razavian et al., 2014). In other cases, we fine-tune the full image network. We specify which representation we choose in a later section.

$^5$Karpathy et al. (2014, §3.3.1) discuss violations in the image fragment/single-word case.
\[ k \text{ on the number of links, we can add the following additional constraint: } 6 \sum_{i,j} x_{ij} \leq k(S, V). \]

The JV algorithm’s runtime is \( O(\max(n, m)^3) \), and each positive example requires computing similarities for the positive case and the \( 2b \) negative samples from Eq. 1, for a per-example runtime of \( O(b \cdot \max(n, m)^3) \). Fortunately, \( \text{lapjv} \) is highly optimized, so despite solving many integer programs, AP often runs faster than DC.

3.3 Baselines

We construct two baseline similarity functions, as we are not aware of existing models that directly address our task in an unsupervised fashion.

Object Detection. For each document, we use DenseNet169 (Huang et al., 2017) to find its \( K \) most probable ImageNet classes (e.g., “stingray”), and represent the image as the average of the word2vec embeddings of those \( K \) labels. We represent each sentence in a document as the word2vec embedding of its words. To form the cosine similarity between all sentence-image pairs to form \( M \) for \( K \in \{1 \ldots 20\} \) and report the variant with the best post-hoc performance on the test set.

NoStruct. The similarity functions described in §3.2 rely on document-level, structural information, i.e., for a single image in a document, the other images in a document affect the overall similarity (and vice versa for sentences). However, this structural information may not be worth incorporating. Thus, we train a baseline that solely relies on single image/single sentence co-occurrence statistics. At training time, we randomly sample a single image and a single sentence from a document, compute the cosine similarity of their vector representations, and treat that value as the document similarity. While the randomly sampled image/sentence will not truly correspond for every sample, we still expect this baseline to produce above-random results when averaged over many iterations, as true correspondences have some (low) probability of being sampled.\(^7\)

4 Experiments on Crowdlabeled Data

Our first set of experiments uses four pre-existing datasets created by asking crowdworkers to add sentence-long textual descriptions to images in a collection. Image-sentence alignments are therefore known by construction. We do not use these labels at training time: gold-standard alignments are only used at evaluation time to compare performance between algorithms.\(^8\) Statistics of these datasets are given in the top half of Table 1, and example documents are given in Figure 2. Each crowdlabeled dataset is constructed to address a different question about our learning setting.

Q: Is this task even possible? Test: MSCOCO. MSCOCO (Lin et al., 2014) was created by crowd-sourced manual captioning of single images. We construct “documents” from this data by first randomly aggregating five image-caption pairs. We then add five “distractor” images with no captions and five “distractor” captions with no images. Thus,

\(^7\)This probability is equal to the density of the ground-truth, underlying image-sentence association graph.

\(^8\)The supplementary material gives more details.
a non-distractor image truly corresponds to the single caption that was written about it, and not to the other 9 captions in the document. There are a total of 10 images/sentences per document, and 5 ground-truth image-sentence links. A priori, we expect this to be the easiest setting for within-document disambiguation because mismatched images and sentences are completely independent.

Q: What if the images/sentences within a document are similar? Test: Story-DII. Huang et al. (2016) asked crowdworkers to collect subsets of images contained in the same Flickr album (Thomee et al., 2016) that could be arranged into a visual story. In the Story-DII (= “descriptions in isolation”) case, (possibly different) crowdworkers subsequently captioned the images, but only saw each image in isolation. We construct a set of documents from Story-DII so that each contains five images and five sentences. Because images come from the same album, images and captions in our Story-DII “documents” are more similar to each other than those in our MSCOCO “documents.”

Q: What if the sentences are cohesive and refer to each other? Test: Story-SIS. Huang et al. (2016) also presented all the images in a subset from the same Flickr album to crowdworkers simultaneously and asked them to caption the image subsets collectively to form a story (SIS = “story in sequence”). In contrast to Story-DII, the generated sentences are generally not stand-alone descriptions of the corresponding image’s contents, and may, for example, use pronouns to refer to elements from neighboring sentences and images.

Q: What if there are many sentences with no corresponding images? Test: DII-Stress. Because documents often have many sentences that do not directly refer to visual content, we constructed a setting with many more sentences than images. We augment documents from Story-DII with 45 randomly negatively sampled distractor captions. The resulting documents have five images and fifty sentences, where only five sentences truly describe images in the document.

Experiment Protocols. We conduct our evaluations over a single randomly sampled train/dev/test split. For image features, we extract the pre-classification layer of DenseNet169 (Huang et al., 2017) pretrained on the ImageNet (Russakovsky et al., 2015) classification task, unless otherwise specified. We train with Adam (Kingma and Ba, 2014) using a starting learning rate of .0001 for 50 epochs. We decrease the learning rate by a factor of 5 each time the loss in Eq. 1 over the dev set plateaus for more than 3 epochs. We set $d_{multi} = 1024$, and apply dropout with $p = .4$. At test time, we use the model checkpoint with the lowest dev error.

<table>
<thead>
<tr>
<th>train/val/test</th>
<th>$n_i/m_i$</th>
<th># imgs density</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSCOCO</td>
<td>25K/2K/2K</td>
<td>10/10</td>
</tr>
<tr>
<td>Story-DII</td>
<td>22K/3K/3K</td>
<td>3/5</td>
</tr>
<tr>
<td>Story-SIS</td>
<td>37K/5K/5K</td>
<td>5/5</td>
</tr>
<tr>
<td>DII-Stress</td>
<td>22K/3K/3K</td>
<td>50/5</td>
</tr>
<tr>
<td>DIY</td>
<td>7K/1K/1K</td>
<td>15/16</td>
</tr>
<tr>
<td>RQA</td>
<td>7K/1K/1K</td>
<td>6/8</td>
</tr>
<tr>
<td>WIKI</td>
<td>14K/1K/1K</td>
<td>86/5</td>
</tr>
</tbody>
</table>

Table 1: Dataset statistics: top half = crowdlabeled datasets; bottom half = organically-multimodal datasets. Density measures the sparsity of the ground truth graph as the number of ground-truth edges divided by the number of possible edges.

4.1 Crowdlabeled-Data Results

We tried all combinations of $b \in \{10, 20, 30\}$, $\text{sim} \in \{DC, TK, AP\}$. For TK and AP we set the maximum link threshold $k$ to $\min(S_i, V_i)$ or $\left\lceil \frac{1}{2} \min(S_i, V_i) \right\rceil$ (denoted $\frac{k}{2}$ in the results table).\(^9\)

Table 2 shows test-set prediction results for $b = 10$ (results for $b \in \{20, 30\}$ are similar). The retrieval-style objectives we consider encourage algorithms to learn useful within-document representations, and incorporating a structured similarity is beneficial. All our algorithms outperform the strongest baseline (NoStruct) in all cases, e.g., by at least 10 absolute percentage points in p@1 on Story-DII.

We next show, as a sanity check, that our inter-document training objective function (Eq. 1) corresponds to intra-document prediction performance (the actual function of interest). Figure 3 plots how both functions vary with number of epochs, for two different validation datasets. In general, inter-document performance and intra-document performance rise together during training;\(^11\) for a fixed neural architecture, models better at optimiz-

\(^9\)Anecdotally, we found that values of 256 and 512 produced similar performance in early testing.

\(^{10}\)For datasets where $m_i = n_i$ and the first choice of definition for $k$ is used, DC and TK are the same. But running the duplicate algorithms anyway provides us with a rough sense of run-to-run variability.

\(^{11}\)See the supplementary material for plots for all datasets; while the general pattern is the same, some of the training curves exhibit additional interesting patterns.
Table 2: Results for crowdlabeled datasets (similar results for other settings are included in the supplementary material). Values are bolded if they are within 1% of the best-in-column performance.

<table>
<thead>
<tr>
<th></th>
<th>MSCOCO</th>
<th>Story-DII</th>
<th>Story-SIS</th>
<th>DII-Stress</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC</td>
<td>p@1/p@5</td>
<td>p@1/p@5</td>
<td>p@1/p@5</td>
<td>p@1/p@5</td>
</tr>
<tr>
<td>Random</td>
<td>49.7</td>
<td>50.0/4.6</td>
<td>49.4</td>
<td>19.5/19.2</td>
</tr>
<tr>
<td>Obj Detect</td>
<td>89.5</td>
<td>67.7/45.9</td>
<td>65.3</td>
<td>50.2/35.2</td>
</tr>
<tr>
<td>NoStruct</td>
<td>87.5</td>
<td>50.6/34.6</td>
<td>76.6</td>
<td>60.1/46.2</td>
</tr>
<tr>
<td>DC</td>
<td>98.9</td>
<td>93.6/80.1</td>
<td>82.8</td>
<td>71.5/55.5</td>
</tr>
<tr>
<td>TK</td>
<td>99.0</td>
<td>95.0/81.1</td>
<td>82.0</td>
<td>72.6/54.9</td>
</tr>
<tr>
<td>$\downarrow$ + $\frac{1}{k}$</td>
<td>98.7</td>
<td>91.0/78.0</td>
<td>82.6</td>
<td>70.5/55.0</td>
</tr>
<tr>
<td>AP</td>
<td>98.9</td>
<td>93.9/80.4</td>
<td>81.6</td>
<td>72.4/54.4</td>
</tr>
</tbody>
</table>

Figure 3: Inter-document objective \( (\text{AP}, b = 10) \) and intra-document \text{AUC} increase together during training.
image captions were explicitly instructed to provide literal descriptions and “not describe what a person might say” (Chen et al., 2015). The manner in which users interact with multimodal content “in the wild” significantly differs from crowdlabeled data: Marsh and Domas White’s (2003) 49-element taxonomy of multimodal relationships (e.g., “decorate”, “reiterate”, “humanize”) observed in 45 web documents highlights the diversity of possible image-text relationships.

We thus consider two datasets (one of which we release ourselves) of organically-multimodal documents scraped from web data, where the original authors created or selected both images and sentences. Statistics of these datasets are given in the bottom half of Table 1.

**RQA.** RecipeQA (Yagcioglu et al., 2018) is a question-answering dataset scraped from instructibles.com consisting of images/descriptions of food preparation steps; we construct documents by treating each recipe step as a sentence. Users of the Instructibles web interface put images and recipe steps in direct correspondence, which gives us a graph for test time evaluation.

**DIY (new).** We downloaded a sample of 9K Reddit posts made to the community DIY (“do it yourself”). These posts consist of multiple images that users have taken of the progression of their construction projects, e.g., building a rock climbing wall (see Figure 2). Users are encouraged to explicitly annotate individual images with captions, and, for evaluation, we treat a caption written alongside a given image as corresponding to a true link.

We adopt the same experimental protocols as in §4, but increase the maximum sentence token-length from 20 to 50; Table 3 shows the test-set results. In general, the algorithms we introduce again outperform the NoStruct baseline. In contrast to the crowdlabeled experiments, AP (slightly) outperformed the other algorithms. DIY is the most difficult among the datasets we consider.

To see if the algorithms err on the same instances, we again compute the Spearman correlation $\rho$ between test-instance AUC scores for DC/TK/AP, for $b = 10$. We find greater variation in performance on organically-multimodal compared to crowdlabeled data. For example, on RQA, DC and AP have a $\rho$ of only .64. We also repeat the regression on test-instance AUC scores introduced in §4.1 with different results; content generally explains more variance than spread, e.g., for AP, for RQA/DIY respectively, only 2/1% is explained by spread alone, but 18/13% is explained by spread+content.

### 6. Qualitative Exploration

To visualize the within-document prediction for document $i$, we compute $\hat{M}_i$ and solve the linear assignment problem described in §3.2, taking the edges with highest selected weights to be the most confident. Figure 4 contains example test predictions (along with $\hat{M}_i$) from the datasets with ground-truth annotation. In an effort to provide representative cases, the selected examples have AUC scores close to average performance for their corresponding datasets.

The model mostly succeeds at associating literal objects and their descriptions: tennis players in MSCOCO, castles in Story-DII, a stapler in DIY, and bacon in a blender in RQA. Errors are often justifiable. For example, for the MSCOCO document, the chosen caption for a picture of two people playing baseball accurately describes the image, despite it having been written for a different image and thus counting as an error in our quantitative evaluation.

Similarly, for RQA, a container of maple syrup is associated with a caption mentioning “syrup”, which seems reasonable even though the recipe’s author did not link that image/sentence.

In other cases, the algorithm struggles with what part of the image to “pay attention” to. In the Story-DII case (Figure 4b), the algorithm erroneously

---

Table 3: Performance on the organically-multimodal data; values within 1% of best-in-column are bolded.

<table>
<thead>
<tr>
<th></th>
<th>RQA AUC p@1/p@5</th>
<th>DIY AUC p@1/p@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>49.4/17.8/16.7</td>
<td>49.8/6.3/6.8</td>
</tr>
<tr>
<td>Obj Detect</td>
<td>58.7/25.1/21.5</td>
<td>53.4/17.9/11.8</td>
</tr>
<tr>
<td>NoStruct</td>
<td>60.5/33.8/27.0</td>
<td>57.0/13.3/11.8</td>
</tr>
<tr>
<td>DC</td>
<td>63.5/38.3/30.6</td>
<td>59.3/20.8/16.1</td>
</tr>
<tr>
<td>TK</td>
<td>67.9/44.0/35.8</td>
<td>60.5/21.2/16.0</td>
</tr>
<tr>
<td>$\hat{M}_i + \frac{1}{k}$</td>
<td>68.1/44.5/35.4</td>
<td>56.0/14.1/12.5</td>
</tr>
<tr>
<td>AP</td>
<td>69.3/47.3/37.3</td>
<td>61.8/22.5/17.2</td>
</tr>
<tr>
<td>$\hat{M}_i + \frac{1}{k}$</td>
<td>68.7/47.2/36.2</td>
<td>59.4/21.6/15.3</td>
</tr>
</tbody>
</table>

---

14 Recipe steps have variable length, are often not strictly grammatical sentences, and can contain lists, linebreaks, etc.
15 We required at least 25 upvotes per Reddit post to filter out spam and low-quality submissions.
16 As with RQA, DIY captions are not always grammatical.
17 This holds even when varying the number of negatively sampled documents; see the supplementary material.

2040
My boss is great and makes me laugh. I don’t have to waste my time making extra trips after work to go shopping ... lame but I love my job. The store even carries my favorite brand of soup, and look at that price, what a deal!

I work at a grocery store, some may think it’s tame but I love my job. The store even carries my favorite brand of soup, and look at that price, what a deal! After a long day of dealing with customers, I get very much relieved from it. Making extra trips after work to go shopping ... lame but I love my job. The store even carries my favorite brand of soup, and look at that price, what a deal!

While I made a triple batch for competition, this recipe is scaled... This layer will be your “meat” strip in the ... Combine ... This one is just syrup and smoke. Combine syrup and smoke.

Pour the quart of half-and-half into the blender. Weigh out about 120g... First, fry up a pound of your favorite thin-sliced bacon. For this dish, ... for a nice relaxing evening at home. I bought some fabric, enough to fail on the first try. Temporary stool. I bought some fabric, enough to fail on the first try. Temporary stool. I bought some fabric, enough to fail on the first try. Temporary stool.


For more difficult datasets like Story-SIS (Figure 4c), the algorithm struggles with ambiguity. For 2/5 sentences that refer to literal objects/actions (soup cans/laughter), the algorithm works well. The remaining 3 captions are general musings about working at a grocery store that could be matched to any of the three remaining images depicting grocery store aisles. DIY is similarly difficult, as many images/sentences could reasonably be assigned to each other.

WIKI. We also constructed a dataset from English sentence-tokenized Wikipedia articles (not including captions) and their associated images from ImageCLEF2010 (Popescu et al., 2010). In contrast to RQA and DIY, there are no explicit connections between individual images and individual sentences, so we cannot compute AUC or precision, but this corpus represents an important organically-multimodal setting. We follow the same experimental settings as in §4 at training time, but instead of using pre-extracted features, we fine-tune the vision model’s parameters. Examining the predictions of the AP+fine-tuned CNN model trained on WIKI shows many of the model’s predictions to be reasonable. Figure 5 shows the model’s 5 most confident predictions on the 100-sentence Wikipedia article about Mauritius, chosen for its high image/text spread.

7 Additional related work

Our similarity functions are inspired by work in aligning image fragments, such as object bounding boxes, with portions of sentences without ex-
This archipelago was formed in a series of undersea volcanic eruptions 8-10 million years ago... (93.9)

The island is well known for its natural beauty. (92.1)

First sighted by Europeans around 1600 on Mauritius, the dodo became extinct less than eighty years later. (84.5)

... a significant migrant population of Bhumihar Brahmins in Mauritius who have made a mark for themselves in different fields. (79.8)

Mauritian Créole, which is spoken by 90 per cent of the population, is considered to be the native tongue... (68.3)

Figure 5: Predicted sentences, with cosine similarities, for images in a 100-sentence ImageCLEF Wikipedia article on Mauritius. The first three predictions are reasonable, the last two are not. The third result is particularly good given that only two sentences mention dodos; for comparison, the object-detection’s choice began “(Mauritian Créole people usually known as ‘Creoles’)

8 Conclusion and Future Directions

We have demonstrated that a family of models for learning fine-grained image-sentence links within documents can produce good test-time results even if only given access to document-level co-occurrence at training time. Future work could incorporate better models of sequence within document context (Kim et al., 2015; Alikhani and Stone, 2018). While using structured loss functions improved performance, image and sentence representations themselves have no awareness of neighboring images/sentences; this information should prove useful if modeled appropriately.19

Acknowledgments. We thank Yoav Artzi, Cristian Danescu-Niculescu-Mizil, Jon Kleinberg, Vlad Niculae, Justine Zhang, the Cornell NLP seminar, the reviewers, and Ondrej Linda, Randy Puttick, Ramin Mehran, and Grant Long of Zillow for helpful comments. We additionally thank the NVidia Corporation for the GPUs used in this study. This material is supported by the U.S. National Science Foundation under grants BIGDATA SES-1741441, 1526155, 1652536, and the Alfred P. Sloan Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

19 Attempts to incorporate document context information by passing the word-level RNN’s output through a sentence-level RNN (Li et al., 2015; Yang et al., 2016) did not improve performance.
References


Tamara L Berg, Alexander C Berg, and Jonathan Shih. 2010. Automatic attribute discovery and characterization from noisy web data. In ECCV.

Piotr Bojanowski, Rémi Lajugie, Edouard Grave, Francis Bach, Ivan Lavtev, Jean Ponce, and Cordelia Schmid. 2015. Weakly-supervised alignment of video with text. In ICCV.


Jack Hessel, David Mimmo, and Lillian Lee. 2018. Quantifying the visual concreteness of words and topics in multimodal datasets. In NAACL.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research.


Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. 2016b. Natural language object retrieval. In CVPR.


Xinyang Jiang, Fei Wu, Xi Li, Zhou Zhao, Weiming Lu, Siliang Tang, and Yueting Zhuang. 2015. Deep compositional cross-modal learning to rank via local-global alignment. In ACM MM.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In CVPR.


Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In CVPR.

Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. 2014. Deep fragment embeddings for bidirectional image sentence mapping. In NeurIPS.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: Referring to objects in natural scenes. In EMNLP.


Gunhee Kim, Seunghwan Moon, and Leonid Sigal. 2015. Ranking and retrieval of image sequences from multiple paragraph queries. In CVPR.


Jiwei Li, Minh-Thang Luong, and Dan Jurafsky. 2015. A hierarchical neural autoencoder for paragraphs and documents. In ACL.


Yijuan Lu, Lei Zhang, Qi Tian, and Wei-Ying Ma. 2008. What are the high-level concepts with small semantic gaps? In CVPR.

Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. 2018. Exploring the limits of weakly supervised pretraining. In ECCV.


Edgar Margffoy-Tuay, Juan C Pérez, Emilio Botero, and Pablo Arbeláez. 2018. Dynamic multimodal instance segmentation guided by natural language queries. In ECCV.


Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In NeurIPS.


Varun K. Nagaraja, Vlad I. Morariu, and Larry S. Davis. 2016. Modeling context between objects for referring expression understanding. In ECCV.


Cesc C. Park and Gunhee Kim. 2015. Expressing an image stream with a sequence of natural sentences. In NeurIPS.


Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. 2016. Grounding of textual phrases in images by reconstruction. In ECCV.


Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In CVPR.


Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. YFCC100M: the new data in multimedia research. Communications of the ACM.


Shaomei Wu, Jeffrey Wieland, Omid Farivar, and Julie Schiller. 2017. Automatic alt-text: Computer-generated image descriptions for blind users on a social network service. In CSCW.


Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In NAACL.


Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. 2018. Learning transferable architectures for scalable image recognition. In CVPR.