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 everywhere on the web, but explicit links between images and sentences (or other intra-document textual units) are often not annotated by users. We present algorithms that successfully discover image-sentence relationships without relying on any explicit multimodal annotation. We explore several variants of our approach on seven datasets of varying difficulty, ranging from images that were captioned post hoc by crowd-workers to naturally-occurring user-generated multimodal documents, wherein correspondences between illustrations and individual textual units may not be one-to-one. We find that a structured training objective based on identifying whether sets of images and sentences co-occur in documents can be sufficient 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. For instance, these models could be used to improve user experience: 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. Furthermore, multimodal documents can be viewed as sources of image labels: inferred image-sentence associations can serve as training pairs for improved image representations, particularly in domains lacking readily-available labeled training data. Thus, exploring the type of multimodal grounding of data we advocate may improve performance in downstream tasks.

In this work, we develop unsupervised models that learn to identify connections between a document’s images and sentences. During training, we are given a set of documents containing multiple images and multiple sentences but with no intra-document image-text associations provided. Based on these co-occurring image sets and sentence sets, our goal is to learn to predict intra-document connections between individual images and individual sentences, despite not being given access to supervision at the individual image/sentence level. Figure 1 illustrates. (Note that the associations

---

1This is a working paper draft; the authors welcome questions and comments about the paper/code/data (which will be made available at www.cs.cornell.edu/~jhessel).

2Any discrete unit of language could be used, including n-grams or paragraphs. We focus on sentences because there are existing public sentence-level datasets that we can use for evaluation.
to be inferred are allowed to result in unaligned sentences and/or images.)

Our approach is ranking-based: we train algorithms to score image sets and sentence sets that truly co-occur more highly than image sets and sentence sets 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 report promising results on multimodal datasets ranging from crowdsourced sets to user-generated organically-multimodal data scraped from real web content. Our algorithms generally perform well on the same-document link prediction task: on a crowdsourced 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 data, we find 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. We conduct an error analysis and find that documents with a higher diversity of images/sentences are generally easier for our algorithms to reason about when compared to more homogeneous documents (at least for secondarily-multimodal data).

We conclude by using our algorithm on a Wikipedia image/text dataset with no ground-truth image-sentence links. While the predictions are imperfect, the algorithm qualitatively discovers meaningful patterns, e.g., matching an image of a dodo bird with one of the 2/100 sentences that mentions “dodo” in the corresponding article. All data and code will be made available.\footnote{www.cs.cornell.edu/˜jhessel}

2 Task Formulation

We assume as given a set of documents where each document $d_i = \langle S_i, V_i \rangle$ consists of a set $S_i$ of $n_i = |S_i|$ sentences and a set $V_i$ of $m_i = |V_i|$ images.\footnote{Sentences and images can be considered as sequences rather than sets in naturally-occurring multimodal documents, but sets are more appropriate for modeling some of the crowdsourced corpora we used in our experiments.} 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 of $n_i$ sentence nodes and $m_i$ image nodes. During training, we are given no access to ground-truth image-sentence association graphs, i.e., we do not know a priori which images correspond to which sentences, only that all images/sentences in a document co-occur together; this is why we refer to our task as unsupervised.

We produce a dense sentence-to-image association matrix $\hat{M}_i \in \mathbb{R}^{n_i \times m_i}$, in which each entry is the confidence that there is an edge between the corresponding nodes. Applying different thresholding strategies to $\hat{M}_i$’s confidence values yields different alignment graphs.

**Evaluation.** When we have ground-truth alignment graphs for test documents, we evaluate the correctness of the edges predicted by our algorithms according to two metrics: AUROC (henceforth AUC) and precision-at-$C$ ($p@C$). AUC, commonly used in evaluating link prediction (see \cite{MenonElkan2011}) is the area under the curve of the true-positive/false-positive rate produced by sweeping over possible confidence thresholds; it ranges from 50 (random performance) to 100. $p@C$ measures the accuracy of the top $C$ predicted edges (in our case, the edges correspond to the largest entries in $\hat{M}_i$). This metric is appropriate for use 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 models combine insights from three directions of prior work.

**Image/text-fragment alignment.** Our image-sentence similarity functions are inspired by work in aligning image fragments (such as object bounding boxes) with portions of sentences (\cite{KarpathyKarpathyFeiFei2015,KarpathyJiang2015,Rohrbach2016}); while follow-up work has tackled similar tasks in a supervised setting (e.g., \cite{Plummer2015}), \cite{Karpathy2014} (and follow-up work) do not rely on explicit image/word labels at training time.\footnote{Rohrbach et al. (2016) address the semi-supervised case, exploring the trade-off between supervision and accuracy.} Our model tackles a similar problem, but at the larger granularity of entire images and entire sentences.

**Cross-modal retrieval.** Our loss function is inspired by work in cross-modal retrieval (\cite{Rasiwasia2014}).

\cite{Karpathy2014} and \cite{KarpathyJiang2015} propose learning functions we consider predict a latent similarity function of entire images and entire sentences.
We assume that the dimensionality of multi-modal text-image space is predetermined.

**Extracting sentence representations.** Each sentence’s words are passed through a 300D word-embedding layer initialized with GoogleNews-pretrained word2vec embeddings (Mikolov et al., 2013). Then, the sequence of word vectors is passed to a GRU (Cho et al., 2014); the result is the hinge loss
\[ L(S_i, V_i) = \max_{V' \in W} \left( \alpha - \max_{V_i' \in W'} (\text{sim}(S_i, V_i), \text{sim}(S_i, V_i')) + \max_{S_i' \in V'} (\text{sim}(S_i, V_i), \text{sim}(S_i', V_i)) \right), \]

where, for margin \( \alpha = 0.2 \) (Kiros et al., 2014a; Faghri et al., 2018), \( h \) is the hinge loss
\[ h(p, n) = \max(0, \alpha - p + n). \]

**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 operate on the matrix \( \tilde{M} \in \mathbb{R}^{n \times m} \) corresponding to \( \langle S, V \rangle \), but have differing interpretations of which entries \( \tilde{M}_{ij} \) correspond to weights for existing edges.

**Dense Correspondence (DC).** Following Karpathy et al. (2014), the DC function assumes a dense correspondence between images and sentences wherein non-aligned items are not allowed: each sentence must be aligned to its most similar image, and vice versa, regardless of how small the

---

6In 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.

7Replacing the maxes with means in Eq. 1 yields the usual negative-sampling algorithm.
similarity might be:
\[
\text{sim}(S, V) = \frac{1}{n} \sum_{i=0}^{n} \max_j \tilde{M}_{i,j} + \frac{1}{m} \sum_{j=0}^{m} \max_i \tilde{M}_{i,j}
\] (2)

The underlying assumption of this function can be violated in many ways (Karpathy et al., 2014, §3.3.1): sentences can have no image, and images no sentence. Nonetheless, DC is simple and has been used effectively in prior related work.

**Top-K (TK).** This is a modification of dense correspondence (DC). Instead of assuming that every image has a corresponding sentence and vice versa, only some sentences or images are aligned:

\[
\text{sim}(S, V) = f_{k_{\max}} \left( \begin{bmatrix} \max_j \tilde{M}_{i,j} \\
\vdots \\
\max_i \tilde{M}_{i,m} \end{bmatrix}, k_1(S, V) \right) + f_{k_{\max}} \left( \begin{bmatrix} \max_i \tilde{M}_{i,1} \\
\vdots \\
\max_i \tilde{M}_{i,m} \end{bmatrix}, k_2(S, V) \right)
\]

where \( f_{k_{\max}}(\vec{v}, k) \) is defined as the mean of the \( k \) maximal values of \( \vec{v} \). The hyperparameter \( k_1 \) (respectively, \( k_2 \)) specifies how many sentences (respectively, images) correspond to images (respectively, sentences). If we expect that half the sentences match with some image, we could set \( k_1(S, V) = \frac{1}{2} \vert S \vert \). We discuss our choices of \( k_1 \) and \( k_2 \) for particular experimental settings in §4.1.

**Assignment Problem (AP).** We may wish to consider the image-sentence association task as a linear assignment problem (Kuhn, 1955), which assumes that each image/sentence has at most one association. We use the \texttt{lapjv} implementation of the JV algorithm (Jonker and Volgenant, 1987) to solve an integer program that assigns each image to at most one sentence, and each sentence to at most one image: maximize \( \sum_{i,j} \tilde{M}_{i,j} x_{ij} \) subject to the constraints that all \( x_{ij} \in \{0, 1\} \) and, considering the \( x_{ij} \) as the entries of a matrix \( X \), each row and column sums to at most 1. Given the solution values \( x^*_{ij} \), as a layer in our neural networks, we compute the differentiable function

\[
\text{sim}(S, V) = \left( \sum_{i,j} \tilde{M}_{ij} x^*_{ij} \right) / r \quad \text{where} \quad r \quad \text{is the number of non-zero} \ x^*_{ij}.
\]

Should we want to impose an upper bound \( k \) on the number of links, we can add the following additional constraint:8

\[
\sum_{i,j} x_{ij} \leq k(S, V) \quad \text{for example, one could set} \quad k(S, V) = \frac{1}{2} \min(|S|, |V|).
\]

### 3.3 Baselines

We are not aware of any existing models that address our task directly, and construct two baselines.

**Object Detection.** For each image in the document, we use DenseNet-169 (Huang et al., 2017) to find the probability that the image belongs to each of the 1K ImageNet classes (e.g., “stingray”). We then represent the image as the average of the word2vec embeddings of the class labels for the top \( k \) objects. We represent each sentence in a document as the mean word2vec embedding of its words. To form the strongest possible baseline we compute the cosine similarity between all sentence-image pairs to form \( \hat{M} \) for \( k \in \{1...20\} \) and report the variant that performs best post-hoc 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. However, perhaps this structural information is not 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 may not truly correspond, we still expect this baseline to produce above-random results when averaged over many random samples (as true correspondences have some probability of being sampled).

### 4 Experiments on Secondarily-Multimodal Data

Our first set of experiments uses four datasets created by crowdworkers adding sentence-long textual descriptions to each image in a collection.9 Since the original data was image-only (e.g., images scraped from Flickr) and the text modality was added via annotation, image-sentence alignments are known by construction. While we do not use these labels at training time, having gold-standard alignment labels at evaluation time allows

---

8Applying Volgenant’s (2004) polynomial-time algorithm.

9More details about construction of all datasets are given in the supplementary material.
Ingredients Mint Layer 1. 1 sticks butter 2. 1 cup powdered sugar 3. 1 table spoon milk ... *** Chocolate Layer #1 ... 

RecipeQA

MSCOCO

Story-DII

Story-SIS

DII-Stress

WIKI

Table 1: Dataset statistics: top half = secondarily-multimodal 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.

Figure 2: Sample documents from 6 of our datasets. Image sets and sentence sets are sometimes truncated due to space constraints. The Story-DII example is harder than is typical, but we include it to illustrate a point regarding image spread made in §4.1. *** denotes text-chunk delimiters present in the original data. Recall that at training time, algorithms are given no information about relationships between individual images/sentences, aside from their document-level co-occurrence.
with Adam (Kingma and Ba, 2014) using a starting learning rate of 0.0001 for 50 epochs. We measure convergence by computing the loss in Equation 1 over the dev set, and decrease the learning rate by a factor of 5 each time dev loss plateaus for more than 3 epochs. We set \(d_{\text{multi}} = 1024\), and apply dropout with \(p = 0.4\). At test time, we load the model checkpoint with the lowest dev error.

### 4.1 Secondarily-Multimodal-Data Results

We tried all combinations of: \(b \in \{10, 20, 30\}\), \(\text{sim} \in \{\text{DC, TK, AP}\}\), training with negative sampling vs. hard-negative mining, and, for TK and AP, setting \(k_1, k_2,\) and \(k\) to be \(\min(S_v, V_v)\) \(^{11}\) and \(\max(\frac{1}{2} \min(S_v, V_v))\) (denoted \(\frac{1}{2}k\) in the results table).

Table 2 shows test-set prediction results for \(b = 10\). In general, our algorithms outperform both random guessing, object detection, and the NoStruct baseline by a margin. For example, DC improves performance over the object detection baseline on Story-SIS by roughly 9 absolute percentage points for both \(\text{AUC}\) and p@1. These observations suggest that the retrieval-style objectives we consider encourage algorithms to learn useful within-document representations. Furthermore, our algorithms outperform the NoStruct baseline in all cases, e.g., by 9 absolute percentage points in p@1 on Story-DII; thus — incorporating a structured similarity function further improves performance.

Additionally, adding hard-negative mining is almost always beneficial, particularly for precision, irrespective of which similarity algorithm is used. On DII-Stress, for example, incorporating hard negative mining improves p@1 by more than 5 absolute percentage points for DC, TK, and AP.

Figure 3 shows the relationship between the inter-document objective computed over the validation set (Equation 1) and the intra-document prediction performance (AUC). In general, inter-document performance and intra-document performance rise together during training; \(^{12}\) thus for a fixed architecture/loss, models optimizing Equation 1 better also generally produce better intra-document representations.

One surprising observation is that DC achieves high performance on DII-Stress. Because 45 out of 50 sentences in those documents were distractors, and DC assumes that each of these sentences corresponds to one of the images, 90% of the terms in the first summation in Equation 2 were, by construction, noise. Nonetheless, DC was able to find the signal in this difficult setting, and performed mostly on par with AP and TK.

Allowing AP or TK to make fewer connections (i.e., setting \(\frac{1}{2}k\)) had mixed results; in the MSCOCO case, where the true number of links

---

\(^{10}\) Anecdotally, we found that values of 256 and 512 produced similar performance in early testing.

\(^{11}\) For datasets where \(m_i = n_i\) and this choice of \(k\), \(\text{DC}\) and \(\text{TK}\) are the same. We decided to run the duplicate instances, anyway, to get a rough sense of run-to-run variability.
(5) was the same as the number of links accounted for by \( \text{AP}/\text{TK} + \frac{1}{2} k \)
\(^{13}\) performance improved very slightly. In other cases, 3 links were considered, and slight performance drops were observed. So — setting the thresholding parameter can help, e.g., if the documents do truly have few ground-truth connections, but, in our setting, performance is not significantly impacted by considering more connections than truly exist.

**Models have trouble with the same documents.** We can calculate \( \text{AUC} \) for each test document individually. The Spearman correlation between these individual-instance \( \text{AUC} \) values is very high: for \( (\text{DC}/\text{TK}/\text{AP}) + \text{hard neg} \), for \( b = 10 \), for all datasets introduced thus far, the Spearman correlation between instance-wise \( \text{AUC} \) scores was very high \((p > .89)\); the least similar combination was \( \text{DC} + \text{hard neg} \) vs. \( \text{AP} + \text{hard neg} \) on MSCOCO.

**Content vs. Spread.** Why are some instances more difficult to solve? We consider two hypotheses. The “content” hypothesis is that some concepts are more difficult for algorithms to find multimodal relationships between, e.g., “beauty” may be hard to visualize vs. “dog,” which is a concrete concept \( \text{(Hessel et al., 2018; Mahajan et al., 2018).} \) The “spread” hypothesis is that documents with lower diversity among images/sentences may be harder to disambiguate at test time. For example, a document in which all images and all sentences are about horses requires finer-grained distinctions than a document with a horse, a barn, and a tractor. The Story-DII vs. Story-SIS example in Fig. 2 illustrates this contrast.

To quantify the spread of a document, we first extract vector representations of each test image/sentence.\(^{14}\) Then, after L2-normalizing, we compute the mean squared distance of each vector to their centroid; higher “spread” values indicate that a document’s sentences/images are more diverse. To quantify the content of a document, for simplicity, we simply mean-pool the image/sentence representations, and PCA the results to 20D.

We first compute an OLS regression of image spread + text spread on test \( \text{AUC} \) scores for Story-DII/Story-SIS/DII-Stress\(^{15}\) for AP+HN with \( b = 10 \), and find that 42/23/16\% respectively (F-

1. Without this modification, 10 links are considered.
2. We use DenseNet-169 features for images and mean word2vec for sentences. We don’t use internal model representations as we aim to quantify aspects of the dataset itself.
3. MSCOCO is omitted because the \( \text{AUC} \) scores are all large.

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

|                | RQA | DIY |
|----------------|-----|-----|
|                | \( \text{AUC} \) | p@1/p@5 | \( \text{AUC} \) | p@1/p@5 |
| Random         | 49.4 | 17.8/16.7 | 49.8 | 6.3/6.8 |
| Obj Detect     | 58.7 | 25.1/21.5 | 53.4 | 17.9/11.8 |
| NoStruct       | 60.5 | 33.8/27.0 | 57.0 | 13.3/11.8 |
| DC             | 66.9 | 41.7/34.8 | 59.5 | 19.4/14.9 |
| \( \downarrow \) hard neg | 63.5 | 38.3/30.6 | 59.3 | 20.8/16.1 |
| TK             | 65.6 | 39.7/33.3 | 60.0 | 19.7/15.8 |
| \( \downarrow \) hard neg | 67.9 | 44.0/35.8 | 60.5 | 21.2/16.0 |
| \( \downarrow \) + \( \frac{1}{2} k \) | 68.1 | 44.5/35.4 | 56.0 | 14.1/12.5 |
| AP             | 66.6 | 40.0/34.4 | 59.3 | 16.5/14.1 |
| \( \downarrow \) hard neg | 69.3 | 47.3/37.3 | 61.8 | 22.5/17.2 |
| \( \downarrow \) + \( \frac{1}{2} k \) | 68.7 | 47.2/36.2 | 59.4 | 21.6/15.3 |

---

5. **Experiments on Organically-Multimodal Data**

The previous datasets had captions added by crowdsources. We now consider three datasets (one of which we constructed ourselves and will release) of organically-multimodal documents, where authors created documents that were multimodal at the outset. Two of these come with implicit image-text associations provided by their authors, allowing for quantitative evaluation; for the third, we are limited to qualitative exploration. Statistics of these datasets are given in the bottom half of Table 1.

**RQA.** RecipeQA \( \text{(Yagcioglu et al., 2018)} \) is a question-answering dataset scraped from instructibles.com consisting of images/descriptions of food preparation steps. For our purposes, we do not directly undertake the question-answering tasks posed, but instead construct documents by treating each recipe step as a sentence.\(^{16}\)

**DIY (new).** We downloaded a sample of 9K Reddit posts made to the community DIY (“do it your-
self”). These posts\textsuperscript{17} 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 may explicitly annotate individual images with captions,\textsuperscript{18} and, for evaluation, we assume that a caption written alongside a given image corresponds to a true link.

**WIKI.** We construct this dataset from English sentence-tokenized Wikipedia articles (not including captions) and their associated images from ImageCLEF (Popescu et al., 2010). In contrast to RQA and DIY, there are no observed connections between individual images and individual sentences, so we cannot compute \(AUC\) or precision.

### 5.1 Quantitative Organically-Multimodal Results

We adopt the same experimental protocols as in §4, but increase the maximum sentence token-length to 50 from 20. The test-set performance of our algorithms is given in Table 3. In general, the algorithms we introduce outperform the NoStruct baseline for the organically-multimodal data, as well. In contrast to the secondarily-multimodal experiments, one algorithm (slightly) outperformed the others, specifically: AP + hard negative mining.\textsuperscript{19} DIY is the most difficult among the datasets we consider, and remains a challenge for the approaches we consider here.

Following our earlier analysis for secondarily-multimodal data, we computed the Spearman correlation between test-instance \(AUC\) scores for (DC/TK/AP)+hard neg, for \(b = 10\). While algorithms do generally have trouble with the same instances, the performance of different approaches on organically-multimodal data is more varied than on secondarily-multimodal data, e.g., for RQA, DC and AP only have a Spearman correlation of .64.

### 5.2 CNN Finetuning

We experiment with fine-tuning the parameters of our image model instead of extracting features from a pretrained network. However — given that hundreds of images and sentences need to fit in GPU memory for each batch, we needed to switch our CNN from DenseNet169 to one with a smaller memory footprint; we chose NASNetSmall (Zoph et al., 2018).\textsuperscript{20} We trained models with AP + hard negative mining using fixed, NASNetSmall pre-extracted features and compared those models to ones where we fine-tune the additional 5M CNN parameters. The resulting test \(AUC\)/objectives are:

|            | RQA | DIY | WIKI |
|------------|-----|-----|------|
| \(AUC\)   | Obj | Obj | Obj  |
| Fixed CNN  | 67.6 - .37 | 60.9 - .37 | N/A - .26 |
| Finetuned CNN | 65.7 - .40 | 57.9 - .39 | N/A - .21 |

We did not observe intra-document performance increases with fine-tuning the CNN for DIY and RQA for the experiment settings we consider. However — even though we do not have ground-truth links to compute \(AUC\) on WIKI, the fine-tuned model resulted in much better retrieval performance for that dataset.\textsuperscript{21} Given that better retrieval performance resulted in better intra-document for a fixed algorithm (see Fig. 3) we expect that fine-tuning was beneficial for WIKI.

We qualitatively explore the predictions of the fine-tuned, AP+hard neg WIKI model. Many of the model’s predictions seem reasonable. To visualize predictions given \(\hat{M}_i\) for a given document, we use a solution to the linear assignment problem described in §3.2. Figure 4 shows the model’s 5 most confident predictions on the 100-sentence Wikipedia article about Mauritius, which

\textsuperscript{17}We required at least 25 upvotes per Reddit post to filter out spam and low-quality submissions.

\textsuperscript{18}Similarly to RQA, DIY captions are not always strictly grammatical sentences.

\textsuperscript{19}This observation holds even when varying the number of negatively sampled documents; see the supplementary material for full results.

\textsuperscript{20}Additionally, at training time, for documents with more than 10 images/sentences, we randomly downsample images/sentences to a set of 10 (though at test time, longer documents are kept intact).

\textsuperscript{21}A model trained using DenseNet-169 image features was also outperformed by fine-tuning NASNetSmall.
we cherry-picked because of its high image/text spread.

6 Additional related work

Prior work has addressed the task of identifying objects in single images that are referred to by natural language descriptions (Mitchell et al., 2010, 2013; Kazemzadeh et al., 2014; Karpathy et al., 2014; Plummer et al., 2015; Hu et al., 2016b; Rohrbach et al., 2016; Nagaraja et al., 2016; Hu et al., 2016a; Yu et al., 2016; Peyre et al., 2017; Margffoy-Tuay et al., 2018). In general, a supervised approach is taken (Mao et al., 2016; Krishna et al., 2017; Johnson et al., 2017).

Several prior works have considered streams of multiple images (or videos) for generating captions/longer stories (Park and Kim, 2015; Huang et al., 2016; Shin et al., 2016; Liu et al., 2017). Related tasks involving multi-image/multi-sentence data, include: Agrawal et al. (2016), who consider sorting aligned (image, caption) pairs into stories, image/textual cloze tasks (Iyyer et al., 2017; Yagcioglu et al., 2018), and question-answering tasks (Kembhavi et al., 2017); these tasks are usually supervised. Our work is similar in spirit to Zhu et al. (2015), who use align video clips/subtitles with full books using labeled data.

7 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 coarse-grained document-level co-occurrence information 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 (i.e., the contribution of each image-sentence pair to the loss depends on other images/sentences in the same document), image and sentence representations themselves have no awareness of neighboring images/sentences at test time; this information should prove useful if modeled appropriately.22

Acknowledgments We thank Jon Kleinberg, Vlad Niculae, and Zillow Group for the helpful discussions. We additionally thank NVidia Corporation for the GPUs used in this study. This work was supported in part by NSF grant SES-1741441. 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.

References

Harsh Agrawal, Arjun Chandrasekaran, Dhruv Batra, Devi Parikh, and Mohit Bansal. 2016. Sort story: Sorting jumbled images and captions into stories. In EMNLP.

Malihe Alikhani and Matthew Stone. 2018. Exploring coherence in visual explanations. In Multimedia Information Processing and Retrieval.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In EMNLP.

Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Nikhil Rasiwasia, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. 2014. On the role of correlation and abstraction in cross-modal multimedia retrieval. TPAMI.

Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In CVPR.

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. VSE++: Improving visual-semantic embeddings with hard negatives. In British Machine Vision Conference.

Jack Hessel, David Mimno, 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, Marcus Rohrbach, and Trevor Darrell. 2016a. Segmentation from natural language expressions. In ECCV.

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

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. 2017. Densely connected convolutional networks. In CVPR.

Ting-Hao (Kenneth) Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. Visual storytelling. In NAACL.

22We tried to incorporate text-specific document context information directly by passing sentence-level vectors produced by the word-level RNN through a sentence-level GRU (Li et al., 2015; Yang et al., 2016). However, we were unable to achieve performance improvements over context-less models in our early attempts.
Mohit Iyyer, Varun Manjunatha, Anupam Guha, Yogarshi Vyas, Jordan Boyd-Graber, Hal Daume, and Larry S. Davis. 2017. The amazing mysteries of the gutter: Drawing inferences between panels in comic book narratives. 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.

Roy Jonker and Anton Volgenant. 1987. A shortest augmenting path algorithm for dense and sparse linear assignment problems. Computing.

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

Andrej Karpathy, Armand Joulin, and Li 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 photographs of natural scenes. In EMNLP.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In CVPR.

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

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. 2014a. Multimodal neural language models. In ICML.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2014b. Unifying visual-semantic embeddings with multimodal neural language models. arXiv preprint arXiv:1411.2539.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision.

Harold W Kuhn. 1955. The hungarian method for the assignment problem. Naval research logistics quarterly.

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In ECCV.

Yu Liu, Jianlong Fu, Tao Mei, and Chang Wen Chen. 2017. Let your photos talk: Generating narrative paragraph for photo stream via bidirectional attention recurrent neural networks. In AAAI.

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.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In CVPR.

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

Aditya Krishna Menon and Charles Elkan. 2011. Link prediction via matrix factorization. In ECML PKDD.

Margaret Mitchell, Kees van Deemter, and Ehud Reiter. 2010. Natural reference to objects in a visual domain. In International Natural Language Generation Conference.

Margaret Mitchell, Ehud Reiter, and Kees Van Deemter. 2013. Typicality and object reference. In Proceedings of the Annual Meeting of the Cognitive Science Society.

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.

Julia Peyre, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2017. Weakly-supervised learning of visual relations. In ICCV.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In ICCV.
Adrian Popescu, Theodora Tsikrika, and Jana Kludas. 2010. Overview of the wikipedia retrieval task at imageCLEF 2010. In *CLEF*.

Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RQ Lanckriet, Roger Levy, and Nuno Vasconcelos. 2010. A new approach to cross-modal multimedia retrieval. In *ACM MM*.

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

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*.

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

Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. Cnn features off-the-shelf: an astounding baseline for recognition. In *CVPR*.

Andrew Shin, Katsunori Ohnishi, and Tatsuya Harada. 2016. Beyond caption to narrative: Video captioning with multiple sentences. In *International Conference on Image Processing (ICIP)*.

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*.

Anton Volgenant. 2004. Solving the k-cardinality assignment problem by transformation. *European Journal of Operational Research*.

Shaoimei 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*.

Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. 2018. Recipeqa: A challenge dataset for multimodal comprehension of cooking recipes. In *EMNLP*.

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

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *ECCV*.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhudinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *ICCV*.

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

1 Data preprocessing details

MSCOCO. We downloaded the train/val 2017 images, and the train/val annotations from 2014 and 2017 from the mscoco website (but create our own training and validation splits). Then, we randomly designate half of the images as “true” images (which will eventually be paired with their true captions in documents) and half of the images as “fake” images, which will not be paired with their true captions in documents. Then, we randomly group all true images into groups of 5, and all fake images into groups of five. Then, we pair each real image set with a fake image set, and divide the resulting groups of 10 images into train/val/test splits. Then, for each of the training/validation/testing document sets independently, for each document, we create (usually) 5 true versions of each document (for testing and validation, we only sample a single version of each document, and do not consider the alternate true captions provided by MSCOCO) because (in general) each MSCOCO image comes with 5 caption annotations. For each of these true versions, we randomly sample captions from a pool of all captions written on all images not in that document (but from the train/val/test pools independently, so that there is no overlap between these sets, except in cases where captions happen to be identical). Then, we shuffle the sampled captions for each version. The result is 4968/1655/1655 train/val/test documents, but each training “document” generally consists of 5 versions because MSCOCO images generally come with 5 captions each.

Story-DII/Story-SIS. We downloaded the Story-DII/Story-SIS train/val/test splits along with all images from the visual storytelling website; we preserve these splits for our train/val/test sets. DII stories have multiple annotations per fixed image set, whereas SIS stories have slightly varying images because human annotators were allowed to select images for their story from a single album. We discard any story with any invalid or missing image (the FAQ page on the data download website mentions that images may be missing because users deleted them).

DII-Stress. We augmented the documents from Story-DII with 45 distractor captions (i.e., captions that were not written about any of the images in the document) selected uniformly at random. To preserve train/val/test splits, we limit these uniform selections to within-split samples, i.e., training document distractor captions are sampled only from training documents.

RQA. We download the train and validation questions (29.6K/3.5K) and extract the “context” of each question, which consists of a list of recipe steps and their associated images; without filtering, there are 8.1K unique recipes in the training set, and 983 unique recipes in the validation data. We also download the training/validation images provided. We treat the provided validation split as the test data.

We concatenate the title and the body of the step provided with a space separator. We discard recipe steps that do not contain any tokens, and discard recipes for which there are no images that correspond to steps (e.g., if the only steps for which there were images contained empty text). Then, we reserve training recipes to act as our validation split. Then, we discard all recipes with less than 2 images/recipe steps. The result is 6502/946/878 training/validation/test recipes, with 69K total images. The sizes of the documents are: mean/med/max number of images 11/8/93; and mean/med/max number of sentences 7/6/20

DIY. We downloaded all the submissions on pushshift.io’s files page from Jan. 2013-Oct. 2018. We looped over all of them and found the ones available made to the subreddit “DIY,” which was 241K posts. Then, we discard posts with less than 25 score. While the semantics of the “score” field have changed over time on reddit (other confounding factors: reddit has become more popular over time, DIY has likely changed in popularity, etc.) we intend for this filtration step to be a basic spam filter. We only consider link submissions to imgur urls with “/a/” in the url, indicating that the imgur link is an album, rather than a single image. We then scrape the associated imgur album page and search for all “div” html fields that are “post-image-container,” and extract both the image associated with that field and its associated caption, if it’s not empty; users may leave image captions empty, but may not upload a caption without an associated image. We ignore imgur albums with no “post-image-container” fields. There are 13K documents after this step. We attempt to scrape all images for these documents, discarding gifs and invalid images for simplicity. After this step, we are left with 295K images.

Next, we search for any image du-
plicates using findimagedupes (https://gitlab.com/opennota/findimagedupes) with a neighbor threshold of 3. We discard any documents with any duplicate images. Then, we discard all documents without at least 2 image captions with at least 5 tokens, and discard documents without at least 2 valid images. Because a small number of documents are quite long, we discard documents with more than 40 images or more than 40 captions.\footnote{At this step, it's possible for there to be more captions than images in a document, e.g., because we discard animated gifs that may have been associated with captions.} We split the remaining documents into 6.8K/1K/1K train/val/test documents. Between these documents, there are 154K unique images. The sizes of the documents are: mean/med/max number of images 17.4/16.0/40, mean/med/max number of sentences 16.4/15.0/40.

**WIKI.** We downloaded the ImageClef 2011 wiki retrieval data as a starting point (https://www.imageclef.org/wikidata), and only considered English documents. This dataset contains the fulltext of wikipedia articles, alongside a list of images in each article. We then stripped out wiki formatting, and used Spacy’s (https://spacy.io/) English sentence tokenizer to split documents into sentences (the resulting sentence tokenization is imperfect, but sufficient). We keep only the first 100 identified sentences in a document. We discarded documents with less than 10 sentences, and documents with less than 3 images. The result is 16K articles, we use a 14K/1K/1K train/val/test split. For the results discussed in the paper, we explore same-document predictions on training documents using a model checkpoint with low validation error. The sizes of the documents are: mean/med/max number of images 6/5/108, mean/med/max number of sentences 72/86/100.

## 2 Image model Training Details

NASNetSmall was selected for the image finetuning experiments because of its high performance relative to its memory requirements. Our subsampling process mentioned in the paper, which selects images/sentences to 10 randomly, ensures that at most 110 images are in GPU memory at a time (for 10 negative samples). We found this to be a reasonable practical upper-bound for NASNetSmall on a single GPU with 12GB of RAM, given that a word embedding matrix and a 1024-D GRU must also be in memory.

We perform random data augmentation to help regularize our training. We first resize images to 256 by 256, and, at training time, perform the following data augmentation: random horizontal flipping, up to 20 degree random image rotation, and a random crop to 224 by 224. At validation/test time, we use a center crop.

## 3 Additional Results
Table 4: Results for secondarily-multimodal data with ground-truth annotation with $b = 20$ negative samples.

| Method       | MSCOCO | Story-DII | Story-SIS | DII-Stress |
|--------------|--------|-----------|-----------|------------|
|              | AUC    | p@1/p@5  | AUC       | p@1/p@5   | AUC        | p@1/p@5 |
| Random       | 49.7   | 5.0/4.6   | 49.4      | 19.5/19.2 | 50.0       | 19.4/19.7 | 50.0 | 2.0/2.0 |
| Obj Detect   | 89.5   | 67.4/54.9 | 65.3      | 50.2/35.2 | 58.4       | 40.8/28.6 | 76.9 | 25.7/17.5 |
| NoStruct     | 88.3   | 53.4/35.8 | 76.6      | 60.4/46.2 | 64.9       | 43.3/33.8 | 84.2 | 21.4/15.6 |
| NoStruct+ hard neg | 51.8 | 8.3/5.9   | 75.9      | 63.0/45.0 | 63.3       | 45.1/31.9 | 51.9 | 4.3/3.1 |
| DC           | 98.8   | 92.0/78.6 | 81.8      | 69.1/53.7 | 68.0       | 49.7/37.6 | 93.8 | 58.3/40.1 |
| DC+ hard neg | 98.9   | 93.1/79.9 | 82.9      | 71.9/55.7 | 68.8       | 52.2/38.7 | 95.0 | 65.2/44.9 |
| TK           | 98.8   | 92.1/78.6 | 81.8      | 69.6/53.8 | 68.0       | 49.7/37.6 | 94.4 | 60.2/42.2 |
| TK+ hard neg | 99.0   | 95.0/81.4 | 81.9      | 71.4/54.5 | 67.6       | 51.5/37.8 | 94.7 | 64.5/43.4 |
| AP           | 98.5   | 87.6/75.3 | 81.7      | 68.3/53.5 | 67.3       | 47.1/36.6 | 93.5 | 58.3/39.7 |
| AP+ hard neg | 98.7   | 91.1/77.9 | 82.6      | 70.7/55.0 | 68.6       | 50.6/38.3 | 95.4 | 65.4/45.5 |
| AP+ hard neg+ $\frac{1}{2}k$ | 98.9   | 94.1/80.7 | 81.5      | 72.2/54.2 | 67.4       | 51.9/37.7 | 94.6 | 64.7/43.7 |

Table 5: Results for secondarily-multimodal data with $b = 30$ negative samples.

| Method       | MSCOCO | Story-DII | Story-SIS | DII-Stress |
|--------------|--------|-----------|-----------|------------|
|              | AUC    | p@1/p@5  | AUC       | p@1/p@5   | AUC        | p@1/p@5 |
| Random       | 49.7   | 5.0/4.6   | 49.4      | 19.5/19.2 | 50.0       | 19.4/19.7 | 50.0 | 2.0/2.0 |
| Obj Detect   | 89.5   | 67.4/54.9 | 65.3      | 50.2/35.2 | 58.4       | 40.8/28.6 | 76.9 | 25.7/17.5 |
| NoStruct     | 87.5   | 50.8/34.7 | 76.6      | 59.9/46.2 | 64.9       | 43.4/33.7 | 84.1 | 21.3/15.6 |
| NoStruct+ hard neg | 52.0 | 10.3/6.0  | 75.9      | 63.0/45.0 | 63.0       | 44.5/31.5 | 51.8 | 4.0/2.0 |
| DC           | 98.8   | 92.0/78.7 | 82.2      | 70.5/54.6 | 68.0       | 49.7/37.7 | 93.9 | 58.6/40.3 |
| DC+ hard neg | 98.9   | 93.4/79.9 | 82.8      | 71.3/55.5 | 68.8       | 52.1/38.6 | 95.0 | 63.8/44.5 |
| TK           | 98.8   | 91.6/78.7 | 81.8      | 69.5/53.9 | 68.0       | 49.9/37.7 | 94.4 | 60.5/42.4 |
| TK+ hard neg | 99.0   | 95.2/81.5 | 82.1      | 73.1/55.1 | 67.7       | 51.9/37.8 | 94.7 | 64.2/43.6 |
| AP           | 98.5   | 87.3/75.4 | 81.7      | 67.7/53.4 | 67.3       | 47.1/36.6 | 93.4 | 57.2/39.8 |
| AP+ hard neg | 98.7   | 91.2/78.0 | 82.6      | 71.1/55.0 | 68.5       | 50.3/38.2 | 95.3 | 65.3/45.6 |
| AP+ hard neg+ $\frac{1}{2}k$ | 98.9   | 94.1/80.5 | 81.6      | 72.8/54.4 | 67.4       | 51.8/37.8 | 94.4 | 64.3/43.2 |

Table 6: Results for organically-multimodal data with ground-truth annotation with $b = 20$ negative samples.

| Method       | RQA | AUC    | p@1/p@5 | DIY | AUC    | p@1/p@5 |
|--------------|-----|--------|---------|-----|--------|---------|
| Random       |     | 49.4   | 17.8/16.7 |     | 49.8   | 6.3/6.8 |
| Obj Detect   |     | 58.7   | 25.1/21.5 |     | 53.4   | 17.9/11.8 |
| NoStruct     |     | 60.5   | 34.3/26.8 |     | 56.9   | 13.8/12.2 |
| NoStruct+ hard neg |     | 60.1   | 35.0/26.7 |     | 56.3   | 15.0/12.5 |
| DC           |     | 67.1   | 43.8/34.9 |     | 59.5   | 19.3/15.2 |
| DC+ hard neg |     | 63.4   | 36.6/31.0 |     | 59.3   | 21.0/16.0 |
| TK           |     | 65.2   | 41.6/33.1 |     | 60.0   | 20.4/15.5 |
| TK+ hard neg |     | 67.9   | 45.2/36.0 |     | 60.5   | 20.3/16.2 |
| TK+ hard neg+ $\frac{1}{2}k$ |     | 67.7   | 44.4/35.0 |     | 56.1   | 14.8/12.0 |
| AP           |     | 66.9   | 37.8/34.2 |     | 59.1   | 16.9/13.9 |
| AP+ hard neg |     | 69.4   | 45.9/37.8 |     | 61.9   | 23.3/17.9 |
| AP+ hard neg+ $\frac{1}{2}k$ |     | 68.5   | 44.9/36.4 |     | 59.6   | 21.7/15.7 |
Table 7: Results for organically-multimodal data with $b = 30$ negative samples.

|                      | RQA        | DIY        |
|----------------------|------------|------------|
|                      | AUC p@1/p@5| AUC p@1/p@5|
| Random               | 49.4      | 49.8       |
|                      | 17.8/16.7 | 6.3/6.8    |
| Obj Detect           | 58.7      | 53.4       |
|                      | 25.1/21.5 | 17.9/11.8  |
| NoStruct             | 60.4      | 56.9       |
|                      | 34.5/26.7 | 13.3/11.9  |
| NoStruct+ hard neg   | 59.7      | 55.9       |
|                      | 31.8/27.0 | 14.7/12.4  |
| DC                   | 66.7      | 59.5       |
|                      | 42.7/34.1 | 18.9/14.7  |
| DC+ hard neg         | 63.5      | 59.4       |
|                      | 37.6/30.6 | 20.8/16.4  |
| TK                   | 65.3      | 60.1       |
|                      | 41.2/32.8 | 20.0/15.9  |
| TK+ hard neg         | 68.0      | 60.5       |
|                      | 44.0/36.2 | 21.4/16.1  |
| TK+ hard neg+ $\frac{1}{2}k$ | 67.8  | 57.3       |
|                      | 43.2/35.1 | 19.1/13.5  |
| AP                   | 66.5      | 59.2       |
|                      | 41.0/33.8 | 15.7/14.0  |
| AP+ hard neg         | 69.3      | 61.9       |
|                      | 47.5/37.4 | 24.4/17.8  |
| AP+ hard neg+ $\frac{1}{2}k$ | 68.7  | 59.4       |
|                      | 45.2/36.2 | 22.0/15.7  |

Figure 5: Inter-document objective (AP, $b = 10$) and intra-document AUC during 50 epochs of training for all datasets we consider with ground-truth, intra-document annotations. While there are some interesting discontinuities, e.g., in DII-Stress’s training curves, in general, for a fixed neural architecture/similarity function, better retrieval performance, as measured by the training objective computed over the validation set, equates to better intra-document performance, as measured by AUC.