# Domain-Specific Lexical Grounding in Noisy Visual-Textual Documents

**Gregory Yauney**
Cornell University
gyauney@cs.cornell.edu

**Jack Hessel**
Allen Institute for AI
jackh@allenai.org

**David Mimno**
Cornell University
mimno@cornell.edu



## Abstract

Images can give us insights into the contextual meanings of words, but current image-text grounding approaches require detailed annotations. Such granular annotation is rare, expensive, and unavailable in most domain-specific contexts. In contrast, unlabeled multi-image, multi-sentence documents are abundant. Can lexical grounding be learned from such documents, even though they have significant lexical and visual overlap? Working with a case study dataset of real estate listings, we demonstrate the challenge of distinguishing highly correlated grounded terms, such as "kitchen" and "bedroom", and introduce metrics to assess this document similarity. We present a simple unsupervised clustering-based method that increases precision and recall beyond object detection and image tagging baselines when evaluated on labeled subsets of the dataset. The proposed method is particularly effective for local contextual meanings of a word, for example associating "granite" with countertops in the real estate dataset and with rocky landscapes in a Wikipedia dataset.

## 1 Introduction

Multimodal data consisting of text and images is not only ubiquitous but increasingly diverse: libraries are digitizing visual-textual collections (British Library Labs, 2016; The Smithsonian, 2020); news organizations release over 1M images per year to accompany news articles (The Associated Press, 2020); and social media messages are rarely sent without visual accompaniment. In this work, we focus on one such specialized, multimodal domain: New York City real estate listings from the website StreetEasy.

To effectively index image-text datasets for search, retrieval, and other tasks, we need algorithms that learn connections between modalities, doing so from data that is naturally abundant. In documents that contain multiple images and sentences, there may be *no explicit annotations* for image-sentence associations or bounding box-word associations. As a result, existing image captioning/tagging methods are difficult to adapt to *unlabeled* multi-image, multi-sentence documents. Indeed, most prior image captioning work has focused on rare and expensive single-image, single-caption collections such as MSCOCO, which focuses on literal, context-free descriptions for 80 object types (Lin et al., 2014). Similarly, off-the-shelf object detectors may not account for contextual factors: to an ImageNet classifier, "pool" refers to a pool table (Russakovsky et al., 2015). In the specialized real estate context, "pool" commonly refers to a swimming pool.

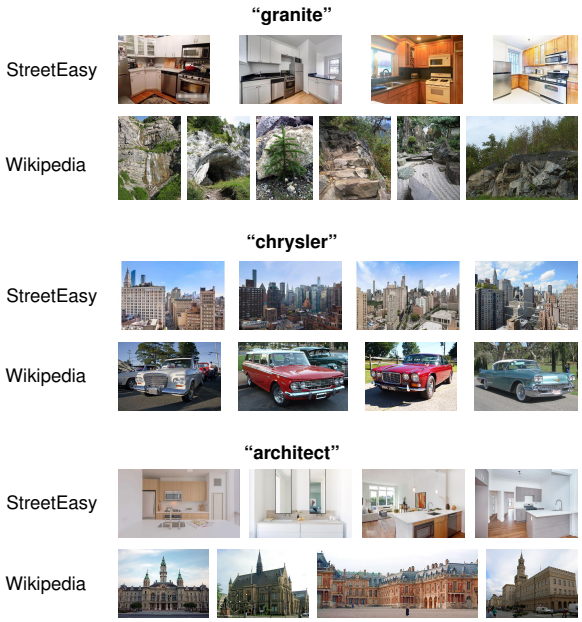


Figure 1: We identify domain-specific associations between words and images from unlabeled multi-sentence, multi-image documents.

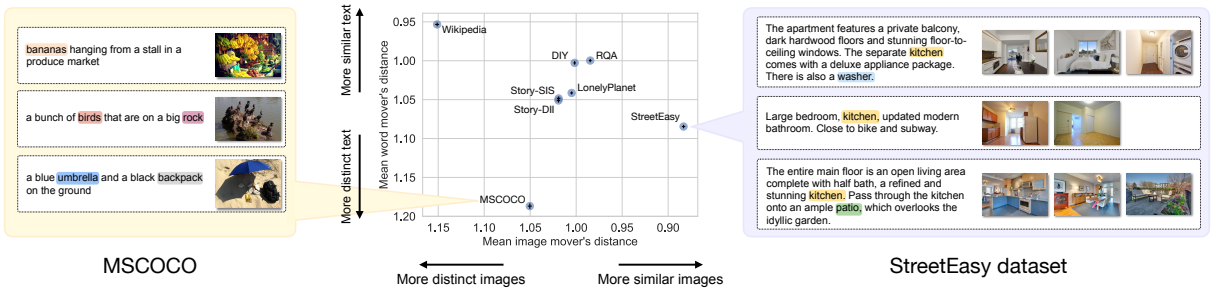


Figure 2: Documents in the StreetEasy dataset are much more visually similar to each other than documents in seven multimodal image-text datasets spanning storytelling, cooking, travel blogs, captioning, etc. (Lin et al., 2014; Huang et al., 2016; Yagcioglu et al., 2018; Hessel et al., 2018, 2019; Nag Chowdhury et al., 2020). Examples from StreetEasy show that words like "kitchen" are frequent and grounded. Black lines represent 99.99% CI.

Consider the task of lexical grounding: given a word, which images in the corpus depict that word? Consider the difficulty in learning a visual grounding for "kitchen" in StreetEasy. First, documents are multi-image, multi-sentence rather than single-image, single-sentence. Second, almost all documents picture a kitchen, a living room, and a dining room. Finally, "kitchen" co-occurs with more than two-thirds of all images, the majority of which are not kitchens. Is this task even possible?

Our first contribution is to map out a landscape of multimodal datasets, placing our real estate case-study in relation to existing corpora. We operationalize this notion in Figure 2 by plotting average across-document visual+textual similarity for our StreetEasy case study compared to several existing multimodal corpora;[1] indeed, images in StreetEasy have very low diversity compared to other corpora. As a result of this self-similarity, in §3, we find that image-text grounding is difficult for off-the-shelf image tagging methods like multinomial/softmax regression, which leverage variation in both lexical and visual features across documents.[2]

Our second contribution is a simple but performant clustering algorithm for this setting, *EntSharp*.[3] We intend this method to learn from $\langle$image, word$\rangle$ co-occurrences collected from multi-image, multi-sentence document collections. The training process iteratively "sharpens" the estimated $\Pr(\text{word} \mid \text{image})$ distributions so that words "compete" to claim responsibility for images. We show that EntSharp outperforms both object detection and image tagging baselines at retrieving relevant images for given word types. We then qualitatively explore EntSharp's predictions on both StreetEasy and a multimodal Wikipedia dataset (Hessel et al., 2018). The algorithm is often able to learn corpus specific relations: as shown in Figure 1, in the context of NYC real estate, "chrysler" refers to a prominent building and "granite" to a kitchen surface, while in Wikipedia the same words are grounded in cars and rocky outcroppings.

**Related work.** Learning image-text relationships is central to many applications, including image captioning/tagging (Kulkarni et al., 2013; Mitchell et al., 2013; Karpathy and Fei-Fei, 2015) and cross-modal retrieval/search (Jeon et al., 2003; Rasiwasia et al., 2010). While most captioning work assumes a supervised one-to-one corpus, recent works consider documents containing multiple images/sentences (Park and Kim, 2015; Shin et al., 2016; Agrawal et al., 2016; Liu et al., 2017; Chu and Kao, 2017; Hessel et al., 2019; Nag Chowdhury et al., 2020). Furthermore, compared to crowd-annotated captioning datasets, web corpora are more challenging, as image-text relationships often transcend literal description (Marsh and White, 2003; Alikhani and Stone, 2019).

## 2 Task and Models

We consider a direct image-text grounding task: for each word type, we aim to retrieve images most-associated with that word. Models are evaluated by their capacity to compute word-image similarities that align with human judgment.

[1] We compute text similarity between documents with a length-controlled version of word mover's distance (WMD) (Kusner et al., 2015) on word2vec token features. We compute visual similarity between documents with "image mover's" distance, which is identical to WMD, but with a CNN feature for each image. More details are given in Appendix A.

[2] Existing unsupervised approaches for this setting (Hessel et al., 2019; Nag Chowdhury et al., 2020) learn within-document matchings of whole sentences/paragraphs, we learn cross-document matchings of word types to images.

[3] Code is at https://github.com/gyauney/domain-specific-lexical-grounding.

**EntSharp.** For each image in a document we iteratively infer a probability distribution over the words present in the document. During training, these distributions are encouraged to have low entropy. The output is an embedding of each word into image space: the model computes word-image similarities in this joint space. This can be thought of as a soft clustering, such that each word type is equivalent to a cluster but only certain clusters are available to certain images. This approach could also be situated within the framework of multiple-instance learning (Carbonneau et al., 2018).

Each image $i$ starts with a fixed feature vector $\vec{i} \in \mathbb{R}^d$. Let $\mathcal{I}$ be the set of these image embeddings. For each word $w$ we initialize a cluster centroid $\vec{w} \in \mathbb{R}^d$ to the average of co-occurring images' embeddings. Let $\mathbb{1}_{i,w}$ be 1 if image $i$ co-occurs with word $w$ in any document and 0 otherwise. Each image $\vec{i}$ is assumed to have a membership distribution $\vec{p}_i$ over words, where $\vec{p}_i$ is initially uniform over co-occurring words. At each iteration, cluster centroids are updated to the weighted average of co-occurring images' embeddings: $\vec{w} := \sum_{\vec{i} \in \mathcal{I}} p_i(w) \cdot \vec{i}$ followed by normalization. Each image's distribution over clusters is updated by taking a softmax of the cosine similarity between pairs of image and word embeddings, first multiplying similarities by a sharpness coefficient[4] equal to the iteration number, and finally masking for co-occurrence: $p_i(w) \propto \mathbb{1}_{i,w} \cdot \exp\left(\text{sharpness} \cdot (\vec{i} \cdot \vec{w})\right)$. After training, we calculate the cosine similarity between image embeddings and the learned word-cluster embedding.

**Untrained EntSharp baseline.** We consider a simple averaging baseline, corresponding to the cluster center initializations of EntSharp: each word embedding is set to the mean of the features for all its co-occurring images.

**Object detection baselines.** We can use ImageNet to identify objects, but most words in the full vocabulary are not in the ImageNet labels. We implement two object detection baselines that map images to object names and then match object names to words in documents (Hessel et al., 2019). For each image, we first get the image's top class predictions from DenseNet169 (Huang et al., 2017) pretrained on the ImageNet classification task (Russakovsky et al., 2015). These predictions are for a whole image and are restricted to the 1000 ImageNet labels. We bridge the gap between ImageNet labels and the vocabulary by then creating an *image vector* by averaging the word vectors corresponding to these predictions. Finally, for each word in the full vocabulary, we rank images by the cosine similarity between the word's vector and these image vectors. Words are represented in one baseline by word2vec embeddings (Mikolov et al., 2013) and in the other by the output of RoBERTa (Liu et al., 2019) when fed a single token as input.

**Image tagging baselines.** Inspired by Mahajan et al. (2018), we implement softmax and multinomial regression models. The former, softmax regression, takes image features and predicts a distribution over the words in the vocabulary with a softmax loss. It computes the word type indicator vector for each document, i.e., 1 if word $w$ was in the document else 0, and then $\ell_1$ normalizes. Multinomial regression computes the word type indicator vector, and—instead of normalizing—computes the logistic sigmoid loss treating the labels as 0/1 indicators. This is equivalent to training a separate logistic regression for each word type to predict the presence/absence of a word type in each document, given the image features. Both models finally use the predicted conditional distributions to produce a ranking of images for each word.

## 3 Experiments

**StreetEasy dataset.** The StreetEasy dataset comprises 29,347 real estate listings in New York City in June 2019. Document excerpts are shown in Figure 2: each consists of both images and English-language sentences. Documents contain an average of 128 word tokens and 10 images, for totals of 3,773,608 word tokens and 294,279 images. There are no image-specific captions or labels. For our quantitative word-image retrieval evaluations, we augment StreetEasy with 17,658 human relevance judgements. After initial experiments, we selected words with a a variety of frequencies and degree of lexical/visual overlap with ImageNet categories: "kitchen" (co-occurs with 200k images), "bedroom" (175k), "washer" (65k), "outdoor" (50k), "fitness" (49k), and "pool" (29k). For each of these words of interest, we labeled a different random 1% subset of all images (2,943 images each): an image in a sample was labeled true if it corresponded with any sense of the associated word and false otherwise. For each model, we rank images for each query

[4] Sharpness is equal to the inverse of softmax temperature; thus EntSharp equivalently decreases softmax temperature during training.


Figure 3: Top images for EntSharp and object detection baselines on the StreetEasy dataset. Images in each word's section come from the same evaluation set, and each row is ranked in decreasing order from left to right. For example, the three rows in the "kitchen" section are different orderings of the same 2,943 images. Images with dark blue borders were labeled true with respect to the word, and those with light red borders were labeled false. E: EntSharp. W: word2vec object detection baseline. R: RoBERTa object detection baseline.

word and calculate the area under the precision-recall curve (PR AUC: perfect performance is 100, and random performance is the percentage of images with true labels). Each of the six evaluation words co-occurred with only some of their sampled images, ranging from kitchen (co-occurred with 1,997 images) to pool (310 images). We perform evaluations on the entire samples of 2,943 images (not just those that co-occur with each word) in order to avoid overstating performance.

**Experimental details for EntSharp.** For each image, features are extracted from the final pre-classification layer of DenseNet169 pre-trained on ImageNet (Russakovsky et al., 2015) and then randomly projected from 1,664 dimensions to 256.[5] We use a vocabulary of 7,971 words that occur at least ten times across this corpus and Wikipedia (to eliminate misspellings). We run EntSharp for 100 iterations.[6] Setups for baselines are comparable, and more details are available in Appendix C.

[5]Random projection is a time and memory optimization. The baseline approaches have access to full feature vectors.

[6]The average runtime is 198 $\pm$ 3.6 minutes on an Intel Xeon Gold 6134 (3.20GHz) CPU with 512 GB RAM.

**Results.** As shown in Table 1, EntSharp outperforms all baselines on PR AUC on all six of the evaluation words. The uniform initialization (Untrained EntSharp) is strong for frequent words ("kitchen", "bedroom") but poor otherwise. The word2vec baseline is also superior to the RoBERTa baseline in four of six evaluations. The baselines do best on "kitchen", "bedroom", and "washer". Table 2 shows the ImageNet object labels associated with each word in manually selected images. Though "kitchen" is not a category in the ImageNet dataset, "microwave", "refrigerator", and "dishwasher" are, and these words are sufficiently close to "kitchen" to learn an association. Nevertheless, EntSharp achieves the highest PR AUC even in the case of "washer", which is a category learned by the object detection baselines. EntSharp's performance increase is most pronounced for the words "outdoor", "bedroom", "pool", and especially "fitness", which have dissimilar visual manifestations in StreetEasy and ImageNet.

Qualitatively (Figure 3), we see that EntSharp associates "bedroom" with empty rooms containing a door and a window while the word2vec baseline as-

| | washer | kitchen | outdoor | fitness | bedroom | pool |
|---|---|---|---|---|---|---|
| Random | 1.6 | 18.4 | 16.9 | 1.8 | 22.9 | 1.3 |
| word2vec | 49.3 | 52.7 | 20.0 | 1.5 | 39.0 | 20.1 |
| RoBERTa | 62.1 | 21.1 | 13.2 | 2.4 | 34.4 | 17.2 |
| Softmax regression | 2.0 | 19.9 | 21.6 | 3.9 | 23.0 | 13.6 |
| Multinomial regression | 1.8 | 17.4 | 23.6 | 8.1 | 22.8 | 19.3 |
| Untrained EntSharp | 1.0 | 21.6 | 10.1 | 1.4 | 42.7 | 1.2 |
| EntSharp | **70.7** | **72.9** | **68.5** | **77.2** | **64.3** | **49.6** |

Table 1: Area under the precision-recall curve (AUC) for each grounding method on each labeled random image subset. Best-in-column is bolded. Random performance results in an AUC equal to the percentage labeled true.

| Evaluation word | Image | Top DenseNet169 predictions |
|---|---|---|
| "kitchen" |  | 'dishwasher', 'microwave', 'refrigerator' |
| "bedroom" |  | 'sliding_door', 'wardrobe', 'window_shade' |
| "outdoor" |  | 'mountain_bike', 'bicycle-built-for-two' |
| "pool" |  | 'pool_table', 'fountain', 'tub' |
| "washer" |  | 'washer', 'microwave', 'reflex_camera' |
| "fitness" |  | 'shoe_shop', 'dumbbell', 'barbell' |

Table 2: Top DenseNet169 ImageNet class predictions for selected example images.

sociates the word with rooms that contain a bed or a sofa. Similarly, "outdoor" manifests in StreetEasy as building exteriors, but the RoBERTa baseline returns images of bike rooms, presumably because bicycles are usually seen outdoors. In StreetEasy the word "pool" more frequently refers to swimming pools rather than the billiards tables seen in ImageNet. The baseline is not technically wrong in this case (indeed, we marked pool tables as correct), but it misses the more common contextual meaning of the word in the local collection. Finally, none of the baselines are able to handle"fitness".

**Wikipedia experiments.** We also ran EntSharp on a multimodal Wikipedia dataset (Hessel et al., 2018). Figure 1 shows that the algorithm often grounds words differently in Wikipedia's much broader range of images than it does in the StreetEasy dataset. Similarly, top ranked images in Wikipedia for "fitness" included marathon runners rather than the StreetEasy dataset's exercise rooms.

## 4 Discussion

We present EntSharp, a simple clustering-based algorithm for learning image groundings for words. It is motivated by the unlabeled multimodal data that exists in abundance rather than relying on expensive custom datasets. By encouraging words to compete to claim responsibility for images, we "sharpen" the resulting image/word associations. The method is effective at finding contextual lexical groundings of words in unlabeled multi-image, multi-sentence documents even in the presence of high cross-document similarity.

One area for future work would be to better identify and model words that either *don't* have a visual grounding or whose identified visual grounding doesn't align with human expectation. For example, the word "Gristedes" (the name of a supermarket chain) appears in StreetEasy documents, but users rarely post photographs of the supermarkets themselves. Conversely, the word "bright" outside the context of StreetEasy may not be "visually concrete" (according to human judgment); nonetheless, it frequently co-occurs with images of sunlit hardwood floors. Given the lexical and visual identifiability issues explored in §1, incorporating prior human concreteness judgments (e.g., Nelson et al. (2004)) for vocabulary items might enable EntSharp to learn for these sorts of ambiguous lexical items. However, finding an appropriate balance of domain-specific flexibility versus alignment with human priors could pose a significant challenge.

## Acknowledgments

We would particularly like to thank Grant Long for putting together the StreetEasy data as well as Ondrej Linda, Ramin Mehran, and Randy Puttick for helpful conversations. We would like to thank Maria Antoniak for valuable feedback. Data provided by StreetEasy, an affiliate of Zillow Group. The results and opinions are those of the authors and do not reflect the position of StreetEasy, Zillow Group, or any of their affiliates. This work was supported by Zillow Group and NSF #1652536. JH performed this work while at Cornell University.

## 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. 2019. "caption" as a coherence relation: Evidence and implications.

In *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, pages 58–67.

British Library Labs. 2016. Digitised books. https://data.bl.uk/digbks/.

Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. 2018. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77:329–353.

Wei-Ta Chu and Ming-Chih Kao. 2017. Blog article summarization with image-text alignment techniques. In *2017 IEEE International Symposium on Multimedia (ISM)*, pages 244–247. IEEE.

Jack Hessel, Lillian Lee, and David Mimno. 2019. Unsupervised discovery of multimodal links in multi-image, multi-sentence documents. In *EMNLP*.

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

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *CVPR*, pages 4700–4708.

Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. 2016. Visual storytelling. In *NAACL*, pages 1233–1239.

Jiwoon Jeon, Victor Lavrenko, and Raghavan Manmatha. 2003. Automatic image annotation and retrieval using cross-media relevance m odels. In *SIGIR*.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. 2013. Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *ICML*.

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*, pages 740–755. Springer.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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*, pages 181–196.

Emily E Marsh and Marilyn Domas White. 2003. A taxonomy of relationships between images and text. *Journal of Documentation*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NeurIPS*, pages 3111–3119.

Margaret Mitchell, Kees Van Deemter, and Ehud Reiter. 2013. Generating expressions that refer to visible objects. In *NAACL*. Association for Computational Linguistics (ACL).

Sreyasi Nag Chowdhury, William Cheng, Gerard De Melo, Simon Razniewski, and Gerhard Weikum. 2020. Illustrate your story: Enriching text with images. In *ACM WSDM*.

Douglas L Nelson, Cathy L McEvoy, and Thomas A Schreiber. 2004. The university of south florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36(3):402–407.

Cesc C Park and Gunhee Kim. 2015. Expressing an image stream with a sequence of natural sentences. In *NeurIPS*, pages 73–81.

Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. 2010. A new approach to cross-modal multimedia retrieval. In *ACM MM*, pages 251–260.

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. *IJCV*, 115(3):211–252.

Andrew Shin, Katsunori Ohnishi, and Tatsuya Harada. 2016. Beyond caption to narrative: Video captioning with multiple sentences. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3364–3368. IEEE.

The Associated Press. 2020. AP information: https://www.ap.org/en-us/, accessed may 14, 20 20.

The Smithsonian. 2020. Smithsonian open access dataset.

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

## A Document similarity metrics

We compute a length-controlled version of word mover's distance (Kusner et al., 2015) to measure the textual distances between documents. This was inspired by the simple extension to "image mover's distance" enabled by swapping the word2vec token representations to CNN image representations.

After computing image/word mover's distances, we noticed that these metrics were slightly correlated with document length; this correlation was also noted by Kusner et al. (2015), who mention that longer documents might be closer to others "as longer documents may contain several similar words." To account for this, we implemented a version of mover's distances that selects a bootstrap sample of $b_1 = 50$ words and $b_2 = 10$ images before computing distances. The scatterplot we report in Figure 2 is insensitive to reasonable choices of these parameters, as it looks largely the same for any $\langle b_1, b_2 \rangle \in \{10, 30, 50\} \times \{3, 5, 10\}$.

To compute a corpus-level statistic, it's computationally infeasible to compute distances between all possible pairs; some calculations based on the EMD library we are using shows that full computation would take at least a few months. Instead, we randomly sample 10K pairs and report confidence intervals for the mean in the figure.

## B StreetEasy dataset preprocessing

The dataset consists of 29,347 English-language real estate listings from the StreetEasy website from June 2019. They contain a total of 294,279 images and 24,078,190 word tokens across 34,564 word types. We preprocess the text by removing numbers, punctuation, hyphens, and capitalization. We restrict the vocabulary to word types that occur at least ten times in StreetEasy and in the multimodal Wikipedia dataset. This results in 3,773,608 word tokens across 7,971 word types. Figure 2 shows a few excerpts of listings, and Figure 4 shows additional listing excerpts.

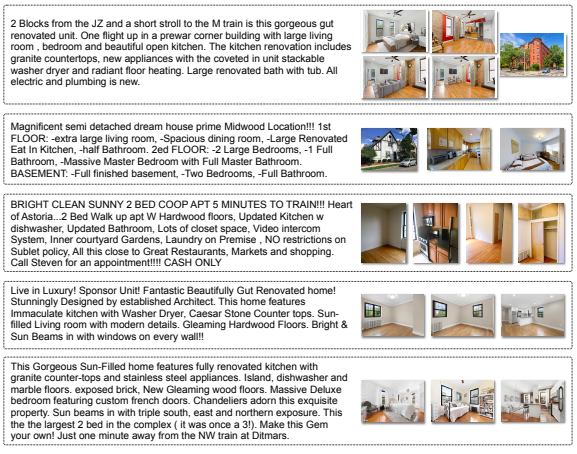


Figure 4: Additional excerpts of documents in the StreetEasy dataset.

## C Baselines

**Object detection.** An image is represented as the mean of the word vectors of its top $K$ class predictions from DenseNet169. We report each model's performance with the $K \in \{1, \ldots, 20\}$ that resulted in the highest average PR AUC across evalation words to create the strongest baselines ($K = 2$ for word2vec and $K = 1$ for RoBERTa). For words not in the word2vec vocabulary, we use a random vector as the word embedding. All six evaluation words are present in the word2vec vocabulary. Average runtimes are $80.9 \pm 1.6$ seconds for word2vec and $458.8 \pm 1.6$ seconds for RoBERTa.

**Image tagging.** We reserved 20% of the StreetEasy corpus as a validation set. We don't hold out a test set: this tasks the algorithms only with fitting the dataset, not generalizing beyond it. We use the validation set for early stopping, model selection, and hyperparameter optimization. We optimize learning rate (in $\{0.001, 0.0005, 0.0007\}$) and number of layers (in $\{0, 1, 2, 3, 4, 5\}$). We decay learning rate upon validation loss plateau. We use the Adam optimizer (Kingma and Ba, 2015).

## D EntSharp training

We run EntSharp for 100 iterations. Figure 5 shows that PR AUC converges at different rates for the different evaluation words.

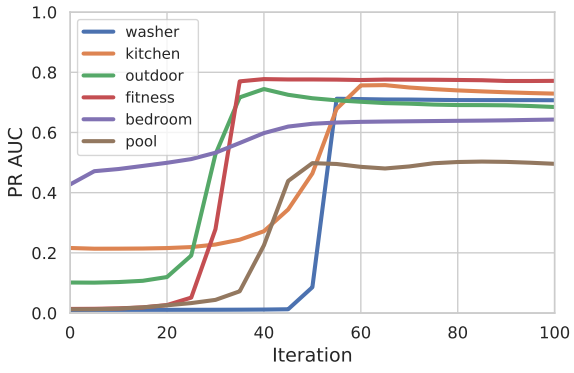


Figure 5: During EntSharp training, PR AUC plateaus at a different rate for each evaluation word.