NewsStories: Illustrating articles with visual summaries

Reuben Tan[1*] Bryan A. Plummer[1] Kate Saenko[1**] JP Lewis[2]
Avneesh Sud[2] Thomas Leung[2]
[1]Boston University, [2]Google Research
{rxtan, bplum, saenko}@bu.edu, {jplewis, asud, leungt}@google.com

Abstract. Recent self-supervised approaches have used large-scale image-text datasets to learn powerful representations that transfer to many tasks without finetuning. These methods often assume that there is a one-to-one correspondence between images and their (short) captions. However, many tasks require reasoning about multiple images paired with a long text narrative, such as photos in a news article. In this work, we explore a novel setting where the goal is to learn a self-supervised visual-language representation from longer text paired with a set of photos, which we call *visual summaries*. In addition, unlike prior work which assumed captions have a **literal** relation to the image, we assume images only contain loose **illustrative** correspondence with the text. To explore this problem, we introduce a large-scale multimodal dataset called NewsStories containing over 31M articles, 22M images and 1M videos. We show that state-of-the-art image-text alignment methods are not robust to longer narratives paired with multiple images, and introduce an intuitive baseline that outperforms these methods, e.g., by 10% on on zero-shot image-set retrieval in the GoodNews dataset[1].

Keywords: vision-and-language, image-and-text alignment

1 Introduction

State-of-the-art image-and-text representation learning approaches generally focus on learning a one-to-one correspondence between an image and one [20,30,15] or more captions [6,32,35], such as a photo with a caption *"An airplane is flying in the sky"* (Figure 1a). While existing datasets such as MSCOCO [23] and Flickr30K [40] contain multiple captions for an image, the aforementioned approaches still learn a one-to-one correspondence between an image and a short caption that generally has a strong literal relation to it. However, this is unrealistic for longer text narratives containing multiple images (*e.g.* news articles, Wikipedia pages, blogs). To challenge such constraints, Kim *et al.* [18] first introduce the problem of retrieving image sequences based on blog posts that may

* Work done as part of an internship at Google
** Also affliated with MIT-IBM Watson AI Lab
[1] Project page: https://github.com/NewsStoriesData/newsstories.github.io

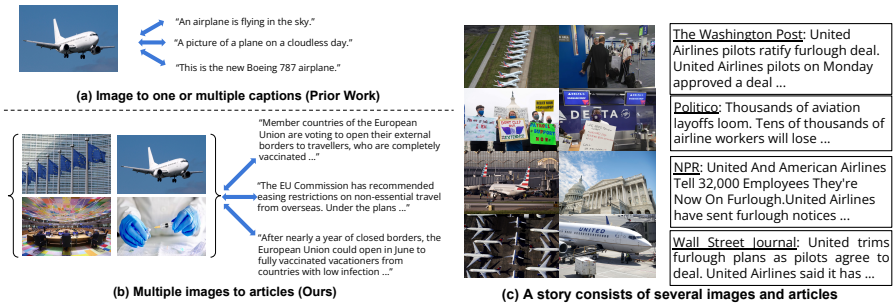

Fig. 1. Unlike prior work (a) which aligns a single image with one or more captions, we study the problem of learning the multiplicity of correspondences between an unordered set of visually diverse images and longer text sequences (b). (c) shows an example story from our NewsStories dataset. For each story, we cluster articles from different media channels and collect images that are used in the articles. In contrast to conventional **literal** caption datasets such as MSCOCO and Flickr30K, the images and text narratives in NewsStories only have **illustrative** relationships

be composed of multiple paragraphs, using the assumption that the image sequence and the paragraphs have the same weak temporal ordering. However, this assumption is quite restrictive due to the prevalence of long narrative texts and groups of relevant images without information about their temporal order, *e.g.*, in the news domain and Wikipedia. More importantly, they do not consider semantically related text narratives that may use similar groups of images.

Motivated by recent representation learning approaches which leverage large-scale data [30,15], we seek to address the important problem of learning visual-semantic representations from text narratives of varying lengths and groups of complementary images from publicly available data. In this paper, we address this research problem in the news domain due to the prevalence of related articles and their corresponding images on the same story from different media channels. However, this is a general problem inherent in other domains including Wikipedia and social media posts on similar events. We define a *story* as an event associated with a *visual summary* consisting of images that illustrate it and articles that describe it, or videos depicting it. In contrast to prior work, this problem requires the capability to reason about the multiplicity of correspondences between sets of complementary images and text sequences of varying lengths. For example, in (Figure 1b), we aim to identify the story that is jointly illustrated by images of an airplane, flags of the European Union (EU) and a nurse preparing a vaccine. Here, one possibility is about traveling to the EU during a pandemic. A story could contain a variety of photos that illustrate a particular concept, and conversely, stock images of an airplane or EU flags could illustrate different stories.

We formulate the problem of visual summarization as a retrieval task, where the image sets are given and the goal is to retrieve the most *relevant* and *illustrative* image set for an article. Our proposed research problem is distinguished from prior work in two ways. First, we must be able to reason about the *many-to-many* correspondences between multiple images and linguistically diverse text

Table 1. Dataset statistics comparison. Our NewsStories dataset is significantly larger than existing datasets with diverse media channels and story clusters that indicate related articles, images and news videos

	GoodNews [4]	NYTimes 800K [36]	Visual News [24]	NewsStories Unfiltered	NewsStories Filtered
# Media channels	1	1	4	28,215	46
# Story clusters	-	-	-	-	350,000
# Articles	257,033	444,914	623,364	31,362,735	931,679
# Images	462,642	792,971	1,080,595	22,905,000	754,732
# Videos	0	0	0	1,020,363	333,357
Avg article length	451	974	773	446	584

narratives in a story. Second, the images in our problem setting often only have *illustrative* correspondences with the text rather than *literal* connections (*e.g.* "travel" or "vacation" rather than "airplane flying"). Extracting complementary information from images and relating them to the concepts embodied by the story is a relatively under-explored problem, especially when the images and text only have loose and symbolic relationships. While existing work such as Visual Storytelling [14] aims to generate a coherent story given a set of images, the images have a temporal ordering and exhibit *literal* relations to the text.

To facilitate future work in this area, we introduce **NewsStories**, a large-scale multimodal dataset (Fig. 1-c and Table 1). It contains approximately $31M$ articles in English and $22M$ images from more than $28k$ news sources. Unlike existing datasets, NewsStories contains data consisting of three modalities - natural language (articles), images and videos. More importantly, they are loosely grouped into *stories*, providing a rich test-bed for understanding the relations between text sequences of varying lengths and visually diverse images and videos. With an expanding body of recent work on joint representation learning from the language, visual, and audio modalities, we hope that our NewsStories dataset will pave the way for exploring more complex relations between multiple modalities.

Our primary goal is to learn robust visual-semantic representations that can generalize to text narratives of varying lengths and different number of complementary images from uncurated data. We benchmark the capabilities of state-of-the-art image-text alignment approaches to reason about such correspondences in order to understand the challenges of this novel task. Additionally, we compare them to an intuitive Multiple Instance Learning (MIL) approach that aims to maximize the mutual information between the images in a set and the sentences of a related articles. We pretrain these approaches on our NewsStories dataset before transferring the learnt representations to the downstream task of article-to-image-set retrieval on the GoodNews dataset under 3 challenging settings, without further finetuning. Importantly, we empirically demonstrate the utility of our dataset by showing that training on it improves significantly on CLIP and increases the robustness of its learnt representations to text with different numbers of images. To summarize, our contributions are as follows:

1. We propose the novel and challenging problem of aligning a story and a set of illustrative images *without temporal ordering*, as an important step towards advancing general vision-and-language reasoning, and with applications such as automated story illustration and bidirectional multimodal retrieval.
2. We introduce a large-scale news dataset NewsStories that contains over 31M articles from 28K media channels as well as data of three modalities. The news stories provide labels of relevance between articles and images.
3. We experimentally demonstrate that existing approaches for aligning images and text are ineffective for this task and introduce an intuitive MIL approach that outperforms state-of-the art methods as a basis for comparisons. Finally, we show that training on the NewsStories dataset significantly improves the model's capability to transfer its learned knowledge in zero-shot settings.

2 Related Work

To the best of our knowledge, there has been limited work that directly address our problem of learning *many-to-many* correspondence between images and texts. Wang et al. [38] propose to learn an alignment between image regions and its set of related captions. However, the images and captions in their setting have strong literal relationships instead of illustrative correspondences. Current vision-language models have other applications including text-based image retrieval [40], visual question answering [2,42] and visual reasoning [33], and as a tool for detection of anomalous image-text pairing in misinformation [1,34].

Recent vision-language models [30,15] demonstrate excellent zero-shot performance on various downstream tasks, sometimes exceeding the performance of bespoke models for these tasks. This advancement has relied on very large-scale datasets consisting simply of images and their associated captions. Such datasets require little or no curation, whereas the need for training labels has limited the size of datasets in the past [13]. These image-caption datasets are paired with a natural contrastive learning objective, that of associating images with their correct captions [41,30,15]. Previous work has demonstrated that improved visual representations can be learned by predicting captions (or parts of them) from images [16,21,31,8]. Captions provide a semantically richer signal [8] than the restricted number of classes in a dataset such as ImageNet – for example, a caption such as "my dog caught the frisbee" mentions two objects and an action.

Closer to our work are methods that learn one-to-many correspondences from images or videos to captions [6,32,35], or vice-versa. Polysemous Instance Embedding Networks (PVSE) [32] represents a datum from either modality with multiple embeddings representing different aspects of that instance, resulting in $n \times n$ possible matches. They use multiple instance learning (MIL) [9] align a image-sentence pair in a joint visual-semantic embedding [11,17,10] while ignoring mismatching pairs. PCME [6] explicitly generalizes the MIL formulation from a single best match to represent the set of possible mebeddings as a normal distribution, and optimize match probabilities using a soft contrastive loss [28]. In contrast to prior work, in our setting both the visual and text modalities con-

tain semantically distinct concepts that are not appropriately represented with a unimodal probability density.

Finally, while there exist news datasets including GoodNews [4], NYT800k [36] and VisualNews [24], they are only sourced from a single (GoodNews and NYT800K) or four (VisualNews) media channels. Additionally, the VMSMO dataset [22] contains news articles and videos that are downloaded from Weibo, but it does not contain images. Compared to these datasets, our NewsStories dataset not only contains articles from over 28K sources, but also has story labels to indicate related articles and images. Related to our work, [12] released the NewSHead dataset for the task of News Story Headline generation, containing $369k$ stories and $932k$ articles but no images. However, ours contains a much larger corpus of stories and associated articles. Last but not least, the aforementioned datasets generally only contain either images and articles or videos and articles. In contrast, ours provides data from all 3 modalities.

3 The NewsStories Dataset

Our NewsStories dataset comprises the following modalities: 1. news articles and meta data including titles and dates 2. images 3. news videos and their corresponding audio. As mentioned above, NewsStories has three main differences from existing datasets. First, it is significantly larger in scale and consists of data from a much wider variety of news media channels. Second, unlike a significant percentage of multimodal datasets, it contains three different modalities – text, image and videos. Third, the text, images, and videos are grouped into stories. This provides story cluster labels that not only help to identify related articles but also create sets of multiple corresponding images for each story.

3.1 Data collection

Learning the multiplicity of correspondences between groups of complementary images and related text narratives requires a dataset that contains multiple *relevant but different* images that correspond to a given text sequence and vice versa. Curating a large-scale dataset with these characteristics is an extremely expensive and time-consuming process. In contrast, the news domain provides a rich source of data due to the proliferation of online multimedia information. We collected news articles and associated media links spanning the period between October 2018 and May 2021[2]. The articles from a filtered subset (Section 3.2) are grouped into *stories* by performing agglomerative clustering on learned document representations [25], similar to [12], which iteratively loads articles published in a recent time window and groups them based on content similarity. We merge clusters that can possibly share similar images via a second round of clustering based on named entities. Specifically, we begin by extracting the named entities in the articles using Spacy and their contextualized representations with a pretrained named-entity aware text encoder [39]. To obtain a single

[2] CommonCrawl [7] can be used to fetch web articles.

vector for the story cluster, we perform average-pooling over all named entity representations across all articles. Finally, these representations are merged into a slightly smaller number of clusters. We extracted video links, however only the text and images are used for alignment in Section 4.

3.2 Dataset filtering

We observed that there is a large amount of noise in the collected datasets due to the prevalence of smaller media channels. To address this, we removed articles and links that do not belong to a curated list of 46 news media channels[3] selected by an independent organization rating media biases. The list contains major news sources including BBC, CNN, and Fox News.

Curated evaluation set. Due to the sparsity of suitable datasets for this task, we curate a subset of the story clusters and use it as our evaluation set for the proxy task of retrieving image sets based on long narrative textual queries. Out of the 350,000 story clusters in the filtered story clusters, we randomly select 5000 clusters with at least 5 images. In the news domain it is common that articles on the same story may use similar photos. For example, different articles on the covid vaccinations may use the same image of a vaccination shot. To ensure that we can visually discriminate between two stories, we adopt a heuristic to ensure that the images are as diverse as possible. We begin by computing a set of detected entities for each image using the Google Web Detection API [4].

To generate a visually diverse image set for a story, we compute all possible combinations of five images from the entire set of images present in the story and compute the intersection set of all detected entities over the images within a combination. Finally, we select the combination with the smallest intersection set as the ground-truth (GT) image set for a story. During training, each set of images is randomly sampled from all available images in the story cluster.

3.3 Quality of story clusters

To evaluate the quality of these story clusters, we use qualified human raters to judge three aspects of our NewsStories dataset. Each data sample is rated by three humans and the rating with the most votes is selected as the final score. We provide the instructions for these evaluations in the supplementary.

Relatedness of articles in a story cluster. For an approach to learn a meaningful alignment between groups of images and groups of related text, it is necessary that the articles in a story cluster are mostly relevant to each other. We randomly sample 100 stories and provide each rater with up to 10 articles from each story. A story is rated as of good quality if at least 80% of the articles are related to each other. Out of 100 randomly selected story clusters, raters determine that 82% of them are of good quality. Note that we do not try to eliminate all noise in the story clusters since we do not have ground-truth targets.

[3] https://www.allsides.com/media-bias/media-bias-chart
[4] https://cloud.google.com/vision/docs/detecting-web

Relevance of images in a story cluster. To rate the semantic relevance of the images to the story, each rater is provided with a maximum of 10 articles and 20 images from each cluster. The image set is labeled as relevant to the articles if at least 80% of the images are plausible in the context of the story. 76% of the randomly selected sets are rated as relevant. Some possible sources of irrelevant images may come from links to other articles or advertisements.

Ambiguity of ground-truth image sets in the evaluation set. A well-known problem of existing bidirectional sentence-to-image retrieval datasets is that some sentences can be relevant to other images in the dataset that are not in the ground truth, which results in inherent noise in the evaluation metrics. We use raters to determine if humans are able to discriminate between the ground-truth image set and others that can potentially be relevant. We randomly sample 150 stories and use the pretrained CLIP model to rank the image sets given the query article. A rater is provided with the GT image set as well as the top-5 sets retrieved by the CLIP model and is asked to select the most relevant image set. 86% of GT sets are selected by at least 2 out of 3 raters as the most relevant, indicating that our annotations are of high quality.

3.4 Data statistics

We compare our NewsStories dataset to existing news datasets such as GoodNews, NYT800K and VisualNews [4,24] in Table 1. We present statistics of both unfiltered and filtered sets. In contrast to existing datasets, the entire dataset contains articles from approximately 28K news media channels, which significantly increases its linguistic diversity. Additionally, it contains over 31M articles and 22M images. We compute the story clusters over articles from the filtered set due to computational constraints, but we release the entire dataset. The articles, images, and videos in the final filtered set are grouped into approximately 350K fine-grained story clusters. The number of articles per story cluster varies greatly across different clusters, ranging from a minimum of 1 to a maximum of about 44,000 (see suppl. for frequency histograms). We observe that story clusters that contain unusually high number of articles tend to come from the entertainment domain, e.g., reality television shows. These story clusters are removed from the final filtered set to eliminate noise. Additionally, we observe that most story clusters contain about 1 to 20 images. This is indicative of the challenges faced in obtaining sufficient data to study the novel problem of learning multiple correspondences between images and text. Finally, please refer to the supplementary material for details on the video statistics.

4 Illustrating articles with visual summaries

The primary objective of this work is to explore the problem of illustrating articles with visual summaries that comprise a varying number of images. By reasoning about the multiplicity of correspondences between longer text sequences and multiple image illustrations, we hope to jointly learn visual-semantic representations that are robust to text narratives of different lengths as well as a

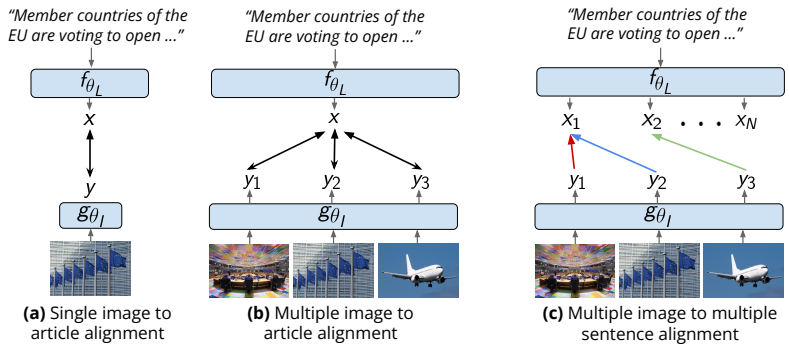

Fig. 2. Comparison of image-and-text alignment objectives. The images and text are encoded by encoders g_{θ_I} and f_{θ_L}, respectively. x and y denote the representations of the encoded article and image respectively. In (c), x is also labeled with a subscript to indicate that it is the representation for a sentence in the article

varying number of images. Specifically, given a set of visually different images, the goal is to learn an aggregated and contextualized representation for the images such that it is able to infer the relevant story, regardless of the exposition styles in the articles. To address this goal, we formulate the task of retrieving the most relevant set of images given a query article.

In this task, given a story consisting of a set of related articles $\mathcal{A}$, and a corresponding set of images $\mathcal{I}$, where $|\mathcal{I}| = N_I$, we aim to maximize the semantic similarity between each article $L \in \mathcal{A}$, and the entire image set $\mathcal{I}$. A language encoder f_{θ_L} is used to encode the entire text sequence L into its representation $x \in \mathbb{R}^{N_L \times D}$. Depending on the number of text segments N_L, L can be used to denote an article-level representation or a set of sentence representations. Each image is encoded to obtain its base representation y_i with an image encoder g_{θ_I} parameterized by weights θ_I, where $i \in \{1, \cdots, N_I\}$.

We describe several existing image-text alignment objectives that can be applied to our task in Sec. 4.1. We then present a Multiple Instance Learning approach for maximizing the correspondence between an image and specific sentences in the article to determine the importance of fine-grained alignment for this problem in Sec. 4.2. See Figure 2 for a high level comparison of these objectives. An effective visual summarization of an article requires the set of retrieved images to be *coherent* and *complete*. Since negative image sets may only overlap with the story partially, we enforce coherence by constraining the models to rank them lower than the positive set for an article. Additionally, our proposed approach seeks to enforce completeness by maximizing the semantic alignment between articles and the sets of relevant images via an article-level loss.

4.1 Existing alignment objectives

Single image-text contrastive loss: We first explore an objective that aligns an image with a caption. Contrastive learning and its variants (InfoNCE and triplet loss) are commonly used to align an image with its corresponding text

sequence by maximizing the similarity score between their respective representations [30,15,3]. We use the popular InfoNCE [5,29] loss, formulated as $\mathcal{L}_{\text{InfoNCE}}(x, y)$:

$$-\log \frac{\exp\left(sim(x,y)/\tau\right)}{\exp\left(sim(x,y)/\tau\right) + \sum_{x'} \exp\left(sim(x',y)/\tau\right) + \sum_{y'} \exp\left(sim(x,y')/\tau\right)} \quad (1)$$

where x', y' denote the non-corresponding text and image representations with respect to a ground-truth pair in a given batch, τ is a temperature value, and $sim(.)$ is a similarity function.

Multiple Instance Learning for Noise Contrastive Estimation: The MIL-NCE formulation [27] provides a natural and intuitive extension to the regular InfoNCE objective by allowing for comparisons between a text sequence and multiple images, formulated as $\mathcal{L}_{\text{MIL-NCE}}(x, y)$:

$$-\log \frac{\sum_{i \in N_I} \exp\left(sim(x,y_i)/\tau\right)}{\exp\left(sim(x,y)/\tau\right) + \sum_{x'} \exp\left(sim(x',y)/\tau\right) + \sum_{y'} \exp\left(sim(x,y')/\tau\right)} \quad (2)$$

When evaluating with sets of images for both the InfoNCE and MIL-NCE baselines, we first compute the similarity score between the text query and each image before taking their average as the final similarity score.

Soft contrastive loss: PCME [6] models each image or text instance as a probability distribution. The probability that an image and a text match ($m = 1$) is computed as:

$$\frac{1}{K^2} \sum_{k}^{K} \sum_{k'}^{K} p(m | z_I^k, z_L^{k'}) \quad (3)$$

where z_I^k is the k-th sampled embedding from the image distribution and K is the number of sampled embeddings from each modality. The probability that two sampled embeddings match is computed using: $p(m|z_I^k, z_L^{k'} = \sigma(-\alpha \| z_I^k - x_L^{k'} \| + \beta)$, where σ is the sigmoid function and α and β are learnable parameters. Finally, the loss for a given pair of image I and text L is formulated as:

$$\mathcal{L}_{\text{soft}} = \begin{cases} -\log p(m|I,L), & \text{if } I \text{ corresponds to } L \\ -\log(1 - p(m|I,L)), & \text{otherwise} \end{cases} \quad (4)$$

During training and evaluation, multiple representations are computed for each image and caption. The final similarity score between a {image, caption} pair is computed from the highest-scoring pair of image and caption representations.

4.2 Multiple Instance Learning - Sentence To Image (MIL-SIM)

Inspired by [18], we assume that each image in a given set should correspond to at least one sentence in the the article. Consequently, we adopt a Multiple Instance Learning framework where a bag of image and sentence instances is

labeled as positive if most of the instances are from the same story cluster and negative otherwise. The MIL-NCE loss formulation is a smooth approximation of the max function, thus it learns an alignment between the entire article and the most representative image in a set. In contrast, MIL-SIM tries to learn the illustrative correspondence between each image and the most relevant sentence.

We segment the text article into individual sentences and encode them as $L = \{x_1, \cdots, x_{N_L}\}$. Given a positive pair of image set $I = \{I_1, \cdots, I_{N_I}\}$ and text article L, we aim to maximize the mutual information between them:

$$\max_{\theta} \mathbb{E}[\log \frac{p(I, L)}{p(I)p(L)}], \tag{5}$$

where $p(I)$, $P(L)$ and $P(I, L)$ are the marginal distributions for the images and articles as well as their joint distribution, respectively.

In this setting, we do not have the ground-truth target labels which indicate the sentence that a given image should correspond to. This problem is compounded by the fact that some of the images may not originate from the same text source but from related articles. Consequently, we generate pseudo-targets by selecting the best matching sentence in an article for an image (colored arrows in Figure 2(c)). Then we use Equation 1 to maximize the lower bound of the mutual information between them. Given an image representation y_i and an article L, we compute their similarity as $\max_l x_l^T y_i$ where l denotes the index of the sentence representation. In this formulation, multiple sentences in a corresponding article may be related to the image but will be treated erroneously as irrelevant. We circumvent this by selecting the highest-scoring sentences, with respect to the image, in articles from other clusters as negatives. Additionally, we mitigate the possibility of a different cluster containing a related sentence by reducing the weight of the image-sentence loss.

However, this may introduce a lot of noise to the learning process, especially if an image originates from a weakly-related article. To alleviate this problem, we impose an article-level loss that aims to maximize the general semantic similarity between the entire article and image set. We compute a single representation for the entire article L_f as well as image set I_f by mean-pooling over the sentence and image representations, respectively. Finally, we learn an alignment between them by minimizing the value of InfoNCE(I^f, L^f), where their similarity is computed as: $\text{sim}(I^f, Y^f) = (I^f)^T Y^f$. Our final objective function is formulated as:

$$\mathcal{L}_{\text{MIL-SIM}} = \sum_{b=1}^{B} L_{\text{InfoNCE}}(I_b^f, L_b^f) + \lambda * \sum_{b=1}^{B} \sum_{i=1}^{N_I} L_{\text{InfoNCE}}(I_{b,i}, L_b), \tag{6}$$

where B and λ are the batch size and trade-off weight, respectively.

5 Experiments

5.1 Implementation details

We use the visual and text encoders of CLIP [30] as a starting point and finetune them using the image-and-text alignment objectives described above. The orig-

Table 2. Comparison on the task of article-to-image-set retrieval on the test split of our NewsStories dataset, which contains 5000 unseen stories with image sets of size 5. Higher R@K values and lower median rank indicate more accurate retrievals

Method	Alignment Type	R@1	R@5	R@10	Median Rank
Pretrained CLIP [30]	Single	31.03	53.87	63.53	4
Single Image	Single	35.88	63.58	74.12	3
MIL-NCE [27]	Single	32.84	59.60	70.92	3
PVSE [32]	Single	36.09	64.26	74.90	3
PCME [6]	Single	35.18	65.52	75.65	3
Transformer [37]	Multiple	50.08	78.79	86.10	2
Mean	Multiple	49.12	76.04	85.18	2
MIL-SIM	Multiple	**54.24**	**82.76**	**90.38**	**1**

Table 3. Zero-shot evaluations of article-to-image-set retrieval approaches on our test splits of the GoodNews [4] dataset. Each test split has 3, 4, or 5 images in each article

	R@1			R@5			R@10			Median Rank		
Method	3	4	5	3	4	5	3	4	5	3	4	5
CLIP [30]	22.29	21.13	20.90	41.14	39.25	38.83	49.94	47.33	47.41	11	13	13
Single Image	17.27	16.27	15.61	34.84	32.57	32.55	43.94	41.47	40.95	16	19	20
MIL-NCE [27]	13.75	13.87	13.30	30.14	29.06	28.33	38.30	37.40	36.40	24	26	29
PVSE [32]	19.21	20.29	20.17	38.52	37.72	39.21	47.72	48.04	49.17	14	14	13
PCME [6]	20.08	20.65	20.14	39.36	39.72	39.91	48.12	48.56	49.03	14	14	13
Transformer [37]	29.06	28.69	29.41	51.15	**50.77**	**51.61**	59.83	59.71	60.57	5	5	5
Mean	28.73	28.01	28.77	50.22	49.11	50.24	58.80	58.64	59.39	5	6	5
MIL-SIM	**29.42**	**30.59**	**30.23**	**52.07**	49.82	51.44	**60.51**	**61.73**	**62.58**	**4**	**4**	**5**

inal CLIP model is trained end-to-end using 400 million image-and-text pairs. Due to the scale of the pretraining dataset, its representations have been demonstrated to be transferable to downstream tasks without further finetuning on the target datasets. During training, we finetune the projection layers of the CLIP encoders on the train split of our NewsStories dataset. We extend the max input text length in CLIP from 77 to 256 in our experiments. We set an initial learning rate of 1e-5 and optimize the model using the Adam [19] optimizer, with a linear warm-up of 20,000 steps. The learning rate is gradually annealed using a cosine learning rate scheduler. We tune the hyperparameter settings by averaging the validation performance over 5 splits of 1000 articles that are randomly selected from the entire training set. In the MIL-SIM objective, we use the NLTK [26] library to tokenize the articles into sentences and set the value of λ to 0.1.

5.2 Evaluation Datasets and Metrics

We conduct an evaluation of article-to-image set retrieval on both our proposed NewsStories dataset and the GoodNews [4] dataset, which contains approximately 250K articles from the New York Times. For our evaluation on the GoodNews dataset, we create three different evaluation sets of 5000 articles

Table 4. Evaluation of retrieval models on article-to-image-set retrieval on the GoodNews dataset, where the candidate image sets do not contain a fixed number of images

Method	Alignment Type	R@1	R@5	R@10	Median Rank
Pretrained CLIP [30]	Single	18.43	36.59	46.92	12
Single Image [30]	Single	17.14	33.77	43.56	17
MIL-NCE [27]	Single	15.50	28.96	37.60	24
PVSE	Single	18.57	37.70	47.78	12
PCME	Single	19.37	39.19	48.04	12
Mean	Multiple	21.30	41.56	51.66	9
Transformer	Multiple	20.30	40.88	49.24	11
MIL-SIM	Multiple	**25.12**	**46.17**	**56.16**	**7**

with 3, 4 and 5 images with no overlapping articles or images between the sets and evaluate on each separately. Note that compared to [35] which retrieves the ground-truth image from a set of five images, our evaluation setup is more challenging and arguably more realistic for real-world applications. We use two metrics: recall at top-K (R@1, R@5, R@10), where higher recall is better, and Median Rank of the correct sample, where lower is better.

5.3 Quantitative results

NewsStories. Table 2 reports the Recall@K retrieval accuracies and median rank on the test split of our NewsStories dataset. In this setting, each image set has a fixed size of 5 images. As demonstrated by Radford et al.[30], the learnt representations of the pretrained CLIP model transfer effectively to the GoodNews dataset without further finetuning, obtaining a R@1 accuracy of 31.03%. We observe that approaches that align a single image with a text sequence generally perform worse than variants that learn an aggregation function over the images.

In contrast to the video variant of the MIL-NCE approach [27] which reports that aligning a video clip to multiple narrations leads to better performance on downstream tasks, applying such an approach on images and text that only have loose topical relationships does not work out-of-the-box. The retrieval accuracies obtained by MIL-NCE show that maximizing the similarity between the text representation and the most representative image in the set performs worse than training with single images. Despite using a simple average-pooling function, the mean images baseline outperforms the single image and MIL-NCE baselines significantly. This suggests that context between images is crucial.

Although transformers have been shown to be effective at reasoning about context, using an image transformer to compute contextual information across the images only improves 1% over the mean baseline. This indicates that the self-attention mechanism alone is insufficient to capture the complementary relations between related but visually different images. Last but not least, the significant improvements obtained by MIL-SIM over other alignment objectives highlight the importance of maximizing the alignment between an image and the most relevant segment in an article, despite not having access to the ground-truth pairings during training. We also provide results of an ablation study over the length of the input text sequence in the supplementary.

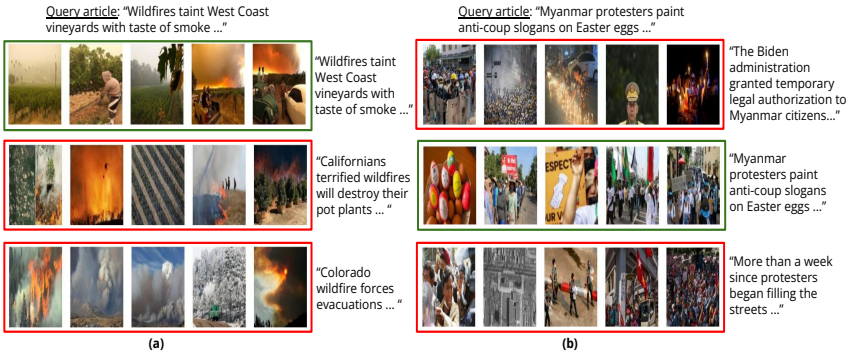

Fig. 3. Qualitative results showing the three top-ranked image sets per query on the test split of our NewsStories dataset. The ground-truth and incorrect image sets, as determined by the cluster labels, are outlined with green and red boxes, respectively

Zero-shot evaluation on GoodNews. Next, we compare the effectiveness of our finetuned models on the curated test splits of the GoodNews dataset without further finetuning. We verify that none of the images and articles are present in our NewsStories train split. Since GoodNews does not group articles into stories, the images in a set are obtained from the same article, instead of related articles as done in NewsStories. Similar to the evaluation results on our NewsStories dataset, Table 3 shows that the pretrained CLIP model already performs well on all 3 test splits of the GoodNews dataset, achieving an average of approximately 21% Recall@1 accuracy without any finetuning on the target dataset. Additionally, finetuning the CLIP model on NewsStories using the standard single image-and-text and MIL-NCE objectives actually leads to significant drops in performance from the pretrained CLIP model. These results suggest that learning a one-to-one correspondence is not optimal for allowing a model to reason about correspondences with multiple related images. This is corroborated by the observation that models trained to learn a one-to-one correspondences between images and text generally perform worse as the number of images increases. In contrast, we observe that training the models to align text with *varying* numbers of images helps them to generalize to sets of images better.

By finetuning the CLIP model on our dataset, the mean baseline shows significantly improved retrieval accuracies despite not observing any articles and images from the GoodNews dataset during training. The much improved performances obtained by the mean images and transformer approaches demonstrate the importance of understanding the complementary relationships between images even in this setting, where all ground-truth images originate from the same text narrative. Similar to the results in the first setting, the results of the best-performing MIL-SIM approach suggests that being able to learn a mapping between each image and specific parts of the text narrative is crucial during training for this research problem, even if the images and text are only weakly-related.

Zero-shot with multiple set sizes on GoodNews. Finally, we evaluate on images sets of variable size in Table 4, where the number of images in a set varies from 3 to 5. This requires a model to not only reason about the general semantic similarity between the query article and images but also determine if

the number of images in a given set provides enough complementary information to discriminate one story from another. To this end, we randomly select 1500 articles from each of the above-mentioned GoodNews evaluation sets to create an evaluation split with different number of images per set.

Despite the inherent noise in obtaining positive text and image pairs using unsupervised clustering, the results suggest that aligning text narratives with varying numbers of complementary images during training is beneficial for learning more robust representations. These learnt representations are better able to discriminate between articles with different number of images.

5.4 Qualitative results

Figure 3 provides an example of correct and incorrect retrievals on the test split of our NewsStories dataset. For each query article, the top 3 retrieved images are displayed in row order from top to bottom. Interestingly, in Figure 3a, our mean image model is able to retrieve other set of images that are relevant to the notion of "fire", despite the fact that they belong to relatively different stories. Figure 3b shows a hard example since the other two retrieved image sets are related to the query article, with the exception of not containing the image of the easter eggs. We include more retrieval visualizations in the supplementary.

5.5 Practical application of retrieving sets of images

While we formulate the evaluation of this research problem as an article-to-image-set retrieval task, it may be impractical to find suitable images that have already been grouped into sets. Hence, we present an algorithm to find individual candidate images before grouping them into sets and ranking them using our trained models. We refer interested readers to the supplementary for more details as well as visualizations of the retrieval results.

6 Conclusion

In this work, we propose the important and challenging problem of illustrating stories with visual summarizes. This task entails learning many-to-many *illustrative* correspondences between relevant images and text. To study this problem in detail, we introduce a large-scale multimodal news dataset that contains over 31M news articles and 22M images. Finally, we benchmark the effectiveness of state-of-the-art image-and-text alignment approaches at learning the many-to-many correspondences between the two modalities and draw useful insights from our empirical results for future research in this direction.

Limitations and societal impact. Data and algorithms dealing with media can potentially be repurposed for misinformation. The selection of media channels in NewsStories reflects the judgment of an independent organization. Our models are trained on top of the CLIP model and may inherit its bias (if any).

Acknowledgements: This material is based upon work supported, in part, by DARPA under agreement number HR00112020054.

References

1. Aneja, S., Bregler, C., Nießner, M.: COSMOS: catching out-of-context misinformation with self-supervised learning. CoRR **abs/2101.06278** (2021), https://arxiv.org/abs/2101.06278
2. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: visual question answering. In: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015. pp. 2425–2433. IEEE Computer Society (2015). https://doi.org/10.1109/ICCV.2015.279, https://doi.org/10.1109/ICCV.2015.279
3. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. CoRR **abs/2104.00650** (2021), https://arxiv.org/abs/2104.00650
4. Biten, A.F., Gomez, L., Rusinol, M., Karatzas, D.: Good news, everyone! context driven entity-aware captioning for news images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12466–12475 (2019)
5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.E.: A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event. pp. 1597–1607 (2020), http://proceedings.mlr.press/v119/chen20j.html
6. Chun, S., Oh, S.J., de Rezende, R.S., Kalantidis, Y., Larlus, D.: Probabilistic embeddings for cross-modal retrieval. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021 (2021), https://openaccess.thecvf.com/content/CVPR2021/html/Chun_Probabilistic_Embeddings_for_Cross-Modal_Retrieval_CVPR_2021_paper.html
7. http://commoncrawl.org
8. Desai, K., Johnson, J.: Virtex: Learning visual representations from textual annotations. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021. pp. 11162–11173. Computer Vision Foundation / IEEE (2021), https://openaccess.thecvf.com/content/CVPR2021/html/Desai_VirTex_Learning_Visual_Representations_From_Textual_Annotations_CVPR_2021_paper.html
9. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artif. Intell. **89**(1-2), 31–71 (1997). https://doi.org/10.1016/S0004-3702(96)00034-3, https://doi.org/10.1016/S0004-3702(96)00034-3
10. Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: VSE++: improving visual-semantic embeddings with hard negatives. In: British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018. p. 12. BMVA Press (2018), http://bmvc2018.org/contents/papers/0344.pdf
11. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., Mikolov, T.: DeViSE: A deep visual-semantic embedding model. In: Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States (2013), https://proceedings.neurips.cc/paper/2013/hash/7cce53cf90577442771720a370c3c723-Abstract.html
12. Gu, X., Mao, Y., Han, J., Liu, J., Yu, H., Wu, Y., Yu, C., Finnie, D., Zhai, J., Zukoski, N.: Generating Representative Headlines for News Stories. In: Proc. of the the Web Conf. 2020 (2020)

13. Gurevych, I., Miyao, Y. (eds.): Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers. Association for Computational Linguistics (2018), https://aclanthology.org/volumes/P18-1/
14. Huang, T.K., Ferraro, F., Mostafazadeh, N., Misra, I., Agrawal, A., Devlin, J., Girshick, R.B., He, X., Kohli, P., Batra, D., Zitnick, C.L., Parikh, D., Vanderwende, L., Galley, M., Mitchell, M.: Visual storytelling. CoRR **abs/1604.03968** (2016), http://arxiv.org/abs/1604.03968
15. Jia, C., Yang, Y., Xia, Y., Chen, Y., Parekh, Z., Pham, H., Le, Q.V., Sung, Y., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: Proceedings of the 38th International Conference on Machine Learning, ICML (2021)
16. Joulin, A., van der Maaten, L., Jabri, A., Vasilache, N.: Learning visual features from large weakly supervised data. In: Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VII (2016). https://doi.org/10.1007/978-3-319-46478-7_5, https://doi.org/10.1007/978-3-319-46478-7_5
17. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015. pp. 3128–3137. IEEE Computer Society (2015). https://doi.org/10.1109/CVPR.2015.7298932, https://doi.org/10.1109/CVPR.2015.7298932
18. Kim, G., Moon, S., Sigal, L.: Ranking and retrieval of image sequences from multiple paragraph queries. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1993–2001 (2015)
19. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015), http://arxiv.org/abs/1412.6980
20. Lee, K.H., Chen, X., Hua, G., Hu, H., He, X.: Stacked cross attention for image-text matching. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 201–216 (2018)
21. Li, A., Jabri, A., Joulin, A., van der Maaten, L.: Learning visual n-grams from web data. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017. pp. 4193–4202. IEEE Computer Society (2017). https://doi.org/10.1109/ICCV.2017.449, http://doi.ieeecomputersociety.org/10.1109/ICCV.2017.449
22. Li, M., Chen, X., Gao, S., Chan, Z., Zhao, D., Yan, R.: Vmsmo: Learning to generate multimodal summary for video-based news articles. arXiv preprint arXiv:2010.05406 (2020)
23. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
24. Liu, F., Wang, Y., Wang, T., Ordonez, V.: Visual News: Benchmark and challenges in news image captioning. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 6761–6771 (2021)
25. Liu, J., Liu, T., Yu, C.: NewsEmbed: Modeling news through pre-trained document representations. arXiv preprint arXiv:2106.00590 (2021)
26. Loper, E., Bird, S.: NLTK: The natural language toolkit. CoRR **cs.CL/0205028** (2002), http://dblp.uni-trier.de/db/journals/corr/corr0205.html#cs-CL-0205028

27. Miech, A., Alayrac, J., Smaira, L., Laptev, I., Sivic, J., Zisserman, A.: End-to-end learning of visual representations from uncurated instructional videos. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020 (2020). https://doi.org/10.1109/CVPR42600.2020.00990, https://openaccess.thecvf.com/content_CVPR_2020/html/Miech_End-to-End_Learning_of_Visual_Representations_From_Uncurated_Instructional_Videos_CVPR_2020_paper.html
28. Oh, S.J., Murphy, K.P., Pan, J., Roth, J., Schroff, F., Gallagher, A.C.: Modeling uncertainty with hedged instance embeddings. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019 (2019), https://openreview.net/forum?id=r1xQQhAqKX
29. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. CoRR **abs/1807.03748** (2018), http://arxiv.org/abs/1807.03748
30. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning, ICML (2021)
31. Sariyildiz, M.B., Perez, J., Larlus, D.: Learning visual representations with caption annotations. In: Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VIII (2020). https://doi.org/10.1007/978-3-030-58598-3_10, https://doi.org/10.1007/978-3-030-58598-3_10
32. Song, Y., Soleymani, M.: Polysemous visual-semantic embedding for cross-modal retrieval. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2019)
33. Suhr, A., Zhou, S., Zhang, A., Zhang, I., Bai, H., Artzi, Y.: A corpus for reasoning about natural language grounded in photographs. In: Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers (2019). https://doi.org/10.18653/v1/p19-1644, https://doi.org/10.18653/v1/p19-1644
34. Tan, R., Plummer, B., Saenko, K.: Detecting cross-modal inconsistency to defend against neural fake news. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 2081–2106. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.emnlp-main.163, https://aclanthology.org/2020.emnlp-main.163
35. Thomas, C., Kovashka, A.: Preserving semantic neighborhoods for robust cross-modal retrieval. In: European Conference on Computer Vision. pp. 317–335. Springer (2020)
36. Tran, A., Mathews, A., Xie, L.: Transform and tell: Entity-aware news image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13035–13045 (2020)
37. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
38. Wang, L., Li, Y., Lazebnik, S.: Learning deep structure-preserving image-text embeddings. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5005–5013 (2016)

39. Yamada, I., Asai, A., Shindo, H., Takeda, H., Matsumoto, Y.: LUKE: deep contextualized entity representations with entity-aware self-attention. arXiv preprint arXiv:2010.01057 (2020)
40. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics **2**, 67–78 (2014)
41. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. CoRR **abs/2010.00747** (2020), https://arxiv.org/abs/2010.00747
42. Zhu, Y., Groth, O., Bernstein, M.S., Fei-Fei, L.: Visual7W: Grounded question answering in images. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. pp. 4995–5004. IEEE Computer Society (2016). https://doi.org/10.1109/CVPR.2016.540, https://doi.org/10.1109/CVPR.2016.540