Who are you referring to? Coreference resolution in image narrations

Arushi Goel\textsuperscript{1}, Basura Fernando\textsuperscript{2}, Frank Keller\textsuperscript{1}, and Hakan Bilen\textsuperscript{1}
\textsuperscript{1}School of Informatics, University of Edinburgh, UK
\textsuperscript{2}CFAR, IHPC, A*STAR, Singapore

Abstract

Coreference resolution aims to identify words and phrases which refer to the same entity in a text, a core task in natural language processing. In this paper, we extend this task to resolving coreferences in long-form narrations of visual scenes. First, we introduce a new dataset with annotated coreference chains and their bounding boxes, as most existing image-text datasets only contain short sentences without coreferring expressions or labeled chains. We propose a new technique that learns to identify coreference chains using weak supervision, only from image-text pairs and a regularization using prior linguistic knowledge. Our model yields large performance gains over several strong baselines in resolving coreferences. We also show that coreference resolution helps improve grounding narratives in images.

1. Introduction

Consider the image paired with the long-form description in Figure 1, an example from the Localized Narratives \cite{47}. Can you tell whether \textit{the woman} who is wearing spectacles refers to a \textit{person} or \textit{another woman} in the text? We are remarkably good at identifying referring expressions (or mentions) and determining which of them corefer to the same entity, a task that we regularly perform when we read text or engage in conversations. The text-only version of this problem is known as coreference resolution (CR) \cite{29, 30, 57}, a core task in natural language processing (NLP) with a large literature. While solving text-only CR requires a very good understanding of the syntactic and semantic properties of the language, the visual version of CR shown in the example also demands an understanding of the visual scene. In our example, a language model has to figure out that a \textit{person} can be a woman, has hands, and correctly match it with her [hand] and \textit{the woman}, but not with \textit{another woman}. However, a language model alone cannot answer whether \textit{the woman} refers to a \textit{person} or \textit{the woman}. This can only be disambiguated after visually inspecting which of the two is \textit{wearing spectacles}.

Figure 1: Coreference resolution from an image and narration pair. Each highlighted block of text is referred to as a mention. The mentions in the same color corefer to the same entity, and belong to the same coreference chain.

Text-only CR has been a crucial component for a range of NLP applications including question answering \cite{28, 14}, sentiment analysis \cite{6, 43}, summarization \cite{19, 54} and machine translation \cite{40, 4, 63}. Most text-only CR methods are either rule-based \cite{29, 49} using heuristics such as pronoun match or exact match based on part of speech tagging, or are learned on large annotated text datasets from domains such as news text or Wikipedia articles \cite{5, 31, 30, 21}. State-of-the-art methods \cite{30, 21} fail to resolve coreferences correctly in image narrations for a few reasons. First, CR in image narrations often require image understanding (see Fig. 1). Neural networks trained on text datasets \cite{48, 9} suffer from poor transferability and a significant performance drop when applied to image narrations because of domain shift. Image narrations are unstructured and can be noisy, unlike the well-edited text used during training (such as news or Wikipedia). Moreover, standard image-text datasets \cite{34, 26, 8, 46} only contain short descriptions with very few or no cases of coreference, thus, are not suitable for training text-only CR models.

Some prior work have looked at visual CR for specific tasks. \cite{50} and \cite{53} link character mentions in TV shows or movie descriptions to character occurrences in videos. More recently, the Who’s Waldo dataset \cite{13} links person names in the caption to their occurrence in the image. However, these methods rely on a limited set of object categories and referring expression types (see Table 3 discussed below), require annotated training data and therefore cannot
be applied to long-form unconstrained image narrations that include open-world object categories and multiple types of referring expressions such as pronouns (she), common nouns (another woman), or proper nouns (Peter).

In this paper, we look at the problem of CR in image narrations, i.e., resolving the coreference of mentions in narrative text that is paired with an image. As the prior benchmarks in this domain are limited to either a small vocabulary of objects or specific referring expression types, we introduce a new dataset, Coreferenced Image Narratives, CIN, which augments the rich long-form narrations in the existing Localized Narratives dataset [47]. We add coreference chain annotations and ground each chain by linking it to a bounding box in the corresponding image.

Manually annotating the whole dataset [47] is expensive, hence these annotations are provided only for evaluation and are not available for training. To cope with this setting, we propose a weakly supervised CR method that learns to predict coreference chains from only paired image-text data. Our key idea is to learn the linking of the mentions to image regions in a joint multi-modal embedding space and use the links to form coreference chains during training. To this end, we propose a multimodal pipeline that represents each modality (image regions, text mentions and also mouse traces, additionally provided by [47]) with a modality-specific encoder and then exploit the cross-modal correlations between them to resolve coreference. Finally, inspired by the rule-based CR [29], we incorporate linguistic rules into our learning formulation in a principled way. We report extensive experiments on CIN and demonstrate that our method brings significant improvements in CR and in weakly supervised narrative grounding, a form of disambiguation that has been underexplored in visual grounding1.

To summarize our contributions, we introduce (1) the new task of resolving coreferences in multimodal long-form textual descriptions (narrations), (2) a new dataset, CIN, that enables the evaluation of coreference chains in text and the localization of bounding boxes in images, which is provided with multiple baselines and detailed analysis for future work, (3) a new method that learns to resolve coreferences while jointly grounding them from weak supervision and exploiting linguistic knowledge, (4) a rigorous experimental evaluation showing significant improvement over the prior work not only in CR but also in weakly supervised grounding of complex phrases in narrative text.

2. Related Work

Text-only CR in NLP has a long history of rule-based and machine learning-based approaches. Early methods [20, 49] used hand-engineered rules to parse dependency trees, which outperformed all learning-based methods at the time. Recently, neural network methods [62, 61, 12, 21, 30] have achieved significant performance gains. The key idea is to identify all mentions in a document using a parser and then learn a distribution over all the possible antecedents for each mention. SpanBERT [21] uses a span-based masked prediction objective for pre-training and shows improvements on the downstream task of CR. Stolfo et al. [55], on the other hand, transfer the pre-trained representations using rules for CR. It is worth noting that all these learning-based approaches either require large pretraining data or training data annotated with gold standard (ground-truth) coreference chains, such as OntoNotes [48] or PreCo [9].

Visual CR includes learning to associate people or characters mentioned in the text with images or videos [50, 53, 13]. Kong et al. [24] exploit CR to relate texts to 3D scenes. Another direction is to resolve coreferences in visual dialog [25] for developing better question-answering systems. Unlike these works, we focus on learning coreferences from long unconstrained image narrations using weak supervision. A related group of work [64, 66, 33, 16] aims to ground phrases in image parts. In visual phrase grounding [64, 35, 10, 65, 16, 22, 32], the main objective is to localize a single image region given a textual query. These models are trained on visual grounding datasets such as ReferItGame [23], Flickr30K Entities [46], or RefCOCO [65]. However, due to short captions, the grounding of text boils down to mostly salient objects in images. In contrast, grounding narrations which aim at capturing all image regions is significantly more challenging and cannot be effectively solved with those prior methods.

Weakly supervised grounding: learning to ground from image-text pairs only, has recently been used in [37, 39, 38, 36, 59] for referring expression grounding. These methods use phrase reconstruction from visual region features as a training signal. Other methods [60, 18, 15] use contrastive learning by creating many negative queries (based on word replacement) or by mining negative image regions for a given query. Wang et al. [60] is a strong method in this domain, hence we establish it as a baseline in our experiments. Liu et al. [39] parses sentences to scene graphs for capturing visual relation between mentions to improve phrase grounding. However, this cannot be directly applied to our task, as parsing scene graphs from narrations are typically very noisy and incomplete. Wang et al. [59] learns/predicts object class labels from the object detector during training and inference respectively. Due to the open-vocabulary setting in our dataset, we directly rely on predictions from the detector and use them as features to avoid the complexity of open-vocabulary object detection. Furthermore, we show in Sec. 5 that grounding is useful to anchor mentions but it is not sufficient to resolve coreferences without prior linguistic knowledge. Thus, our method also employs contrastive

1Our code and dataset is available at https://github.com/VICO-UoE/CIN.
3. Coreferenced Image Narratives

Our CIN dataset contains 1880 images from the Localized Narratives dataset [47] that come with long-form text descriptions (narrations) and mouse traces. These images are originally a subset of the test and validation set of the Flickr30k dataset [46]. We annotated this subset with coreference chains and bounding boxes in the image that are linked with the textual coreference chains, and use them only for validation and testing. Note that we also include singletons (i.e., coreference chains of length one). Fig. 1 shows an example image from CIN.

**Annotation procedure.** The annotation involved three steps: (1) marking the mentions (sequences of words) that refer to a localized region in the image, (2) identifying coreference chains for the marked mentions, including (a) pronominal words such as *him* or *who* that are used to refer to other mentions, (b) mentions that refer to the same entity (e.g., *a lady* and *that person*), and (c) mentions that do not have any links (e.g. *another woman*), (3) drawing bounding boxes in the image for the coreference chains/mentions identified in steps (1) and (2). We created an annotation interface based on LabelStudio [1], an HTML-based tool that allows us to combine text, image, and bounding box annotation. More details are provided in the supplementary material.

### Table 1: Statistics of relevant noun phrases, pronouns, coreference chains and bounding boxes on Flickr30k Entities [46], RefCOCO [65] and CIN.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#noun phrases</th>
<th>#pronouns</th>
<th>#coreference chains</th>
<th>#bounding boxes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flickr30k Entities [46]</td>
<td>15,252</td>
<td>X</td>
<td>X</td>
<td>17,234</td>
</tr>
<tr>
<td>RefCOCO [65]</td>
<td>10,668</td>
<td>X</td>
<td>X</td>
<td>10,668</td>
</tr>
<tr>
<td>CIN (Ours)</td>
<td>19,587</td>
<td>1,659</td>
<td>3,310</td>
<td>21,246</td>
</tr>
</tbody>
</table>

Figure 2: Numbers of mentions as part of the coreference chain for pronouns *them, he, it, who, she* in CIN.

**Dataset statistics.** We split the 1880 images in the dataset into a test and validation set using the pre-defined split of [46]. More specifically, we have 1000 images in the test set and 880 images in the validation set. It is important to note that the narrations have a lot of first-person pronouns such as *I can see . . .* We specifically instruct the annotators to exclude such mentions that are not a part of any coreference chain and at the same time cannot be grounded on the image. We elaborate more on the filtering process for these mentions in the supplementary material.

Overall, the dataset has 19,587 noun phrase mentions, 1,659 pronouns, 3,310 coreference chains and 21,246 bounding boxes. In Table 1, we compare the statistics of CIN with the test/val splits of other related datasets. In Fig. 2, we show the distribution over the frequency and types of mention such as *a metal fence* or *few people* that are referred to using a particular pronoun (*them, he, it, who* and *she*). There is a huge diversity in (1) the categories of the mentions and (2) how many times they form a part of the coreference chain.

In Tab. 2, we also report the frequency of the top-6 mentions for each pronoun category to check the diversity of pronoun words in the dataset.

**Figure 2:** Numbers of mentions as part of the coreference chain for pronouns *them, he, it, who, she* in CIN.

**Table 2: Frequency of top-N mentions for each pronoun category.**

<table>
<thead>
<tr>
<th>Category</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>A man (0.28)</td>
<td></td>
</tr>
<tr>
<td>A woman (0.33)</td>
<td></td>
</tr>
<tr>
<td>A person (0.18)</td>
<td></td>
</tr>
<tr>
<td>People (0.11)</td>
<td></td>
</tr>
<tr>
<td>It (0.06)</td>
<td></td>
</tr>
<tr>
<td>A building (0.04)</td>
<td></td>
</tr>
</tbody>
</table>

**Comparison to existing CR datasets.** In Table 3, we compare our proposed CIN dataset to other CR datasets. This comparison shows that most of the other datasets are either from a restricted domain (*shopping, indoor scenes, etc.*), have limited mention types referring to either only *people* or *limited object categories*, or do not cover all possible referring expression types such as common nouns (*a person*) and pronouns (*he*).

**Table 3: Comparison to existing datasets.**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Modality</th>
<th>Domain</th>
<th>Object categories</th>
<th>Referring expression types</th>
</tr>
</thead>
<tbody>
<tr>
<td>NYU-RGBD v2 [24]</td>
<td>Images</td>
<td>Indoor scenes</td>
<td>Household objects, Common nouns</td>
<td>Common nouns</td>
</tr>
<tr>
<td>SIMM-C 2.0 [17]</td>
<td>Images</td>
<td>Shopping</td>
<td>Clothing, People, Proper names</td>
<td>Common nouns, Pronouns</td>
</tr>
<tr>
<td>MPI-INF-IAI [19]</td>
<td>Videos</td>
<td>Movies</td>
<td>People</td>
<td></td>
</tr>
<tr>
<td>CIN (Ours)</td>
<td>Images</td>
<td>Open-world</td>
<td>General objects, Common nouns</td>
<td>Common nouns, Pronouns</td>
</tr>
</tbody>
</table>

4. Method

4.1. Text-only CR

Given a sentence containing a set of mentions (i.e., referential words or phrases), the task of CR is to identify which mentions refer to the same entity. This is fundamentally a clustering problem [57]. In this work, we use an off-the-shelf NLP parser [2] to obtain the mentions. Formally, let
\[ S = \{m_1, m_2, \ldots, m_{|S|}\} \] denote a sentence with \(|S|\) mentions, where each mention \(m\) contains a sequence of words, \(\{w_1, w_2, \ldots, w_{|m|}\}\). We assign a label \(y_{ij}\) to each mention pair \((m_i, m_j)\), which is set to 1 when the pair refers to the same entity, and to 0 otherwise. We wish to learn a compatibility function, a deep network \(f\) that scores high if a pair refers to the same entity, and low otherwise.

Given a training set \(D\) that contains \(|D|\) sentences with their corresponding labels, one can learn \(f\) by optimizing a binary cross-entropy loss:

\[
\min \sum_{S \in D} \sum_{i=0}^{|S|-1} \sum_{j=i+1}^{|S|} \log(y_{ij}(\sigma(f(m_i, m_j)))) - \frac{1}{2} + \frac{1}{2},
\]

where \(\sigma\) is the sigmoid function. Note that prior methods [29, 30, 21] require large labeled datasets for training and are limited to only a single modality, text. These methods typically also combine the learning with fixed rules based on recency and grammatical principles [29].

### 4.2. CR in image narrations

**Problem definition.** Next, we extend the text-only CR to image-text data in the absence of coreference labels. Let \((I, S)\) denote an image-text pair where \(S\) describes an image \(I\) as illustrated in Figure 1, and assume that coreference labels for mention pairs are not present. As in Sec. 4.1, our goal is to identify the mentions that refer to the same entity in an image-text pair. Each image is defined by \(|I|\) regions \(I = \{r_1, r_2, \ldots, r_{|I|}\}\) which are obtained by running the pre-trained object detector (trained on the COCO [34] and Visual Genome [26] dataset) in [52] on the image. Each region \(r\) is described by its bounding box coordinates \(b\), the text embedding for the detected object category \(o\), and the visual features \(v\). More details are provided in Section 4.3.

**Weak supervision.** We use ‘weak supervision’ to refer to a setting where no coreference label for mention pairs and no grounding of mentions (i.e., bounding boxes are not linked to phrases in the text) are available. Moreover, in contrast to the output space of the object detector (a restricted set of object categories), the sentences describing our images come from the unconstrained vocabulary. Hence, an object instance in a sentence can be referred to with a synonym or may not even be present in the object detector vocabulary [34, 27]. Finally, the object detector can only output category-level labels and hence cannot localize object instances based on the more specific instance-level descriptions provided by the sentences. For instance in Figure 1, a person and the woman both are labeled as person by the object detector.

In addition to image and text, we explore the use of an auxiliary modality, mouse trace segments provided in [47]. Each mouse trace includes a sequence of 2D points over time that relate to a region in the image when de-scribing the scene. As the text in Localized Narratives is a transcription of the speech of the annotators, the mouse traces are synced with spoken words, which we denote as \(T = \{t_1, t_2, \ldots, t_{|T|}\}\) where \(|T| = |S|\). These features are stacked with textual features (see Section 4.3).

In the weakly supervised setting, the key challenge is to replace the coreference label supervision with an alternative one. We hypothesize that each mention in a coreferring pair corresponds to (approximately) the same image region, and it is possible to learn a joint image-text space which is sufficiently rich to capture such correlations. Concretely, let \(g(m, r)\) denote an auxiliary function that is instantiated as a deep network and outputs a score for the mention \(m\) being located at region \(r\) in image \(I\). This grounding score for each mention can be converted into probability values by normalizing them over all regions in the image:

\[
\bar{g}(m, r) = \frac{\exp(g(m, r))}{\sum_{r' \in I} \exp(g(m, r'))}.
\]

The compatibility function \(f\) can be defined as a sum product of a pair’s grounding probabilities over all regions:

\[
f(m, m') = \sum_{r \in I} \bar{g}(m, r)\bar{g}(m', r).
\]

In words, mention pairs with similar region correlations yield bigger compatibility scores and are hence more likely to corefer to each other. The key idea is that we employ the grounding for mentions as anchors to relate coreferring mentions (e.g., a person and the woman). At test time, we compute \(f(m, m')\) for all the pairs and threshold them to predict their pairwise coreference labels.

As no ground-truth bounding box for each mention is available for learning the grounding \(g\), we pose grounding as a weakly supervised localization task as in [18, 60]. To this end, we impute the missing bounding boxes by taking the highest scoring region for a given mention \(m\) at each training iteration:

\[
r_m = \arg \max_{r \in I} g(m, r).
\]

Then we use \(r_m\) as the pseudo-truth to learn \(g\) as following:

\[
\min g \sum_{(I, S) \in D} \sum_{m \in S} -\log \left( \frac{\exp(g(m, r_m))}{\sum_{r' \in I} \exp(g(m, r')}} \right)
\]

where \(r'_m = \arg \max_{r' \in I} g(m, r')\) is the highest scoring region in image \(I\) for mention \(m\). For each mention, we treat the highest scoring region in the original image as positive and other highest scoring regions across different images as negatives, and optimize \(g\) for discriminating between the two. However, as the denominator requires computing \(g\).

\[ \text{15250} \]
Figure 3: Overview of our pipeline. Our model encodes the image regions obtained from an object detector using the image encoder. We parse text mentions and mouse traces from the sentence description, which are then encoded using a text and trace encoder respectively. Finally, a joint text-trace encoder learns a joint embedding for text and traces. A cross-attention module attends to the words given an image region and then we compute the joint probability of the paired mentions, thus forming coreference chains.

over all training samples at each iteration, which is not computationally feasible, we instead sample the negatives only from the randomly sampled minibatch.
Linguistic constraints. Learning the associations between textual and visual features helps with disambiguating coreferring mentions, especially when mentions contain visually salient and discriminative features. However, resolving coreferences when it comes to pronouns (e.g., her, their) or ambiguous phrases (e.g., one man or another man) remains challenging. To address such cases, we propose to incorporate a regularizer into the compatibility function $f(m, m')$ based on various linguistic priors. Concretely, we construct a look-up table for each mention pair $q(m, m')$ based on the following set of rules [29]:
(a) Exact String Match. Two mentions corefer if they exactly match and are noun phrases (not pronouns).
(b) Pronoun Resolution. Based on the part-of-speech tags for the mentions, we set $q(m, m')$ to 1 if $m$ is a pronoun and $m'$ is the antecedent noun that occurs before the pronoun.
(c) Distance between mentions. Smaller distance is more likely to indicate coreference since mentions occur close together if they refer to the same entity.
(d) Last word match. In certain cases, the entire phrases might not match but only the last word of the phrases.
(e) Overlap between mentions. If two mentions have one or more overlapping words, then they are likely to corefer.

Finally, we include $q(m, m')$ as a regularizer in Eq. (5):

$$
\begin{align*}
\min_{g} & \sum_{(I, S) \in D} \sum_{m \in S} \left( - \log \left( \frac{\exp(g(m, r_m))}{\sum_{l \in D_i} \exp(g(m, r_{m,l}))} \right) + \lambda \sum_{m' \in S} ||f(m, m') - q(m, m')||_F^2 \right) \\
\end{align*}
$$

where $\lambda$ is a scalar weight for the Frobenius norm term. Note that $f$ is a function of $g$ (see Eq. (3)). We show in Section 6 that incorporating this term results in steady and significant improvements in CR performance.

4.3. Network modules

Our model (illustrated in Figure 3) consists of an image encoder $e_i$ and text encoder $e_t$ to extract visual and linguistic information respectively, and a cross-attention module $a$ for their fusion.

Image encoder $e_i$ takes in a $d_i$-dimensional vector for each region $r$ that consists of a vector consisting of bounding box coordinates $b \in R^4$, text embedding for the detected object category $o \in R^{do}$ and visual features $v \in R^{dv}$. The regions are extracted from a pre-trained object detector [52] for the given image $I$. The image encoder applies a nonlinear transformation to this vector to obtain a $d$-dimensional embedding for each region $r$.

Text encoder $e_t$ takes in the multiple mentions from a parsed multi-sentence image description $S$ produced by an NLP parser [2] and outputs a $d$-dimensional embedding for each word in the parsed mentions. Note that the parser does not only extract nouns but also pronouns as mentions.

Mouse trace encoder $e_m$ takes in the mouse traces for each mention parsed above after it is preprocessed into a 5D vector of coordinates and area, $(x_m, y_m, y_{max}, y_{max}, \text{area})$ [44] and outputs a $d_m$-dimensional embedding. In [7, 47], mouse trace embeddings have been exploited for image retrieval, however, we use them to resolve coreferences. We concatenate each mention embedding extracted from $e_t$ with the mouse trace encoding $e_m$, denoted as $e_{tm}$ and ap-
ply additional nonlinear transformations (Joint encoder in Fig. 3) before feeding into the cross-attention module.

**Cross-attention module** 

This module takes in the joint text-trace embeddings for all the words in each mention and returns a d-dimensional vector for each m by taking a weighted average of them based on their correlations with the image regions. Concretely, in this module, we first compute the correlation between each word and all regions, take the highest correlation over the regions through an auxiliary function \( \bar{a} \):

\[
\bar{a}(w) = \max_{r \in I} \left( \frac{\exp(e_{tm}(w) \cdot e_i(r))}{\sum_{r' \in I} \exp(e_{tm}(w) \cdot e_i(r'))} \right)
\]

where \( \cdot \) is dot product. The transformation can be interpreted as probability of word \( w \) being present in image \( I \). Then we compute a weighted average of the word embeddings for each mention \( m \):

\[
a(m) = \sum_{w \in m} \bar{a}(w)e_{tm}(w).
\]

Similarly, \( a(m) \) can be seen as probability of mention \( m \) being present in image \( I \).

**Scoring function** \( g(m, r) \) can be written as a dot product between the output of the attention module and region embeddings:

\[
g(m, r) = a(m) \cdot e_i(r).
\]

While taking a dot product between the two embeddings seemingly ignores the correlation between text and image data, the region embedding \( e_i(r) \) encodes the semantic information about the detected object category in addition to other visual features and hence results in a high score only when the mention and region are semantically close. Further implementation details about the modules can be found in Section 5 and the supplementary.

5. Experiments

We train our models on the Flickr30k subset of the Localized Narratives [47] which consists of 30k image-narration pairs, and evaluate on the proposed CIN dataset, which contains 1000/880 pairs for test/val respectively.

**Evaluation metrics.** To evaluate the CR performance, we use the standard link-based metrics MUC [58] and BLANC [51].

(a) **MUC F-measure** counts the coreference links (pairs of mentions) common to the predicted chain \( R \) and the ground-truth chain \( K \) by computing MUC-R (recall) and MUC-P (precision).

(b) **BLANC** measures the precision (BLANC-P) and recall (BLANC-R) between the ground-truth and predicted coreference links and also between non-coreferent links.

(c) **Narrative grounding.** For evaluating narrative grounding in images, we consider a prediction to be correct if the IoU (Intersection over Union) between the predicted bounding box and the ground truth box is larger than 0.5 [60, 18]. We report percentage accuracy for evaluating narrative grounding for both noun phrases and pronouns. Further details about the metrics is in the supplementary material.

**Inputs and modules.** For the image modeling, we extract bounding box regions, visual features and object class labels using the Faster-RCNN object detector [52]. For the text modeling, we use Glove embeddings [45] to encode the object class labels and the mentions from the textual branch. For the mouse traces, we follow [47] and extract the trace for each word in the sentence and then convert it into bounding box coordinates for the initial representation. The model discussed in Sec. 4 referred to as ‘Ours’ in Sec. 6 uses the transformer backbone for the image, text and trace encoders (more details in supplementary).

**Baselines.** We consider the following baselines to fairly compare and evaluate our proposed method:

(a) **Text-only CR:** For all these methods, we directly evaluate the coreference chains using the narration only without the image. (1) **Rule-based** [29]: In this method, a multi-sieve rule based system is used to find mentions in the sentence and the coreference chains, (2) **Neural-Coref** [30]: Instead of rules, this method is trained end-to-end using a neural network on a large corpus of wikipedia data to detect mentions and coreferences, and (3) **Similarity-based:** We compute cosine similarity between mentions using Glove word features and threshold them to get coreference chains.

(b) **Visual-text models:** The baselines discussed below are not trained for CR and hence we post-process their output in order to evaluate for CR. (1) **VisualBert** [56], UNITER [11] and VinVL [67] are vision-language models trained on image-caption data and fine-tuned on downstream tasks such as VQA, NLVR. To test it for CR, we compute the cosine similarity for the multi-modal mention embeddings in a zero-shot way. (2) **GLIP** [32]: GLIP is trained on large-scale image-text paired data with bounding box annotations and shows improvement in object detection and visual phrase grounding. To evaluate it for CR, we predict bounding boxes for the mentions in the narrations from GLIP. If the IoU overlap between the mentions is greater than 0.7, then we consider them to form a coreference chain. (3) **MAF** [60]: MAF is a weakly supervised phrase grounding method, originally trained on the Flickr30k-Entities [46]. We train this model on narrations data and evaluate CR by computing Eq. (3). (4) **MAF++:** We retrain the MAF† model on the narrations with our reg-
Table 4: CR performance on CIN dataset. MT denotes mouse trace and † denotes our trained model.

<table>
<thead>
<tr>
<th>Method</th>
<th>Text</th>
<th>Image</th>
<th>MT</th>
<th>MUC-R</th>
<th>MUC-P</th>
<th>MUC-F1</th>
<th>BLANC-R</th>
<th>BLANC-P</th>
<th>BLANC-F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule-Based [29]</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>5.6</td>
<td>10.13</td>
<td>6.4</td>
<td>3.3</td>
<td>4.1</td>
<td>4.9</td>
</tr>
<tr>
<td>Neural-Coref [30]</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>0.11</td>
<td>0.17</td>
<td>0.13</td>
<td>1.59</td>
<td>36.99</td>
<td>3.23</td>
</tr>
<tr>
<td>Similarity-based</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>7.07</td>
<td>14.43</td>
<td>9.06</td>
<td>37.48</td>
<td>65.17</td>
<td>45.98</td>
</tr>
<tr>
<td>GLIP [32]</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
<td>0.13</td>
<td>0.12</td>
<td>0.12</td>
<td>21.71</td>
<td>61.40</td>
<td>31.66</td>
</tr>
<tr>
<td>VisualBERT [56]</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
<td>18.17</td>
<td>6.08</td>
<td>8.06</td>
<td>23.90</td>
<td>44.07</td>
<td>22.38</td>
</tr>
<tr>
<td>UNITER [11]</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
<td>16.76</td>
<td>7.15</td>
<td>8.83</td>
<td>27.64</td>
<td>55.30</td>
<td>29.20</td>
</tr>
<tr>
<td>VinVL [67]</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
<td>16.76</td>
<td>8.60</td>
<td>9.75</td>
<td>34.56</td>
<td>58.07</td>
<td>36.56</td>
</tr>
<tr>
<td>MAF† [60]</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
<td>25.86</td>
<td>10.18</td>
<td>13.21</td>
<td>37.68</td>
<td>61.14</td>
<td>38.17</td>
</tr>
<tr>
<td>MAF++</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>19.07</td>
<td>15.62</td>
<td>15.65</td>
<td>41.25</td>
<td>65.04</td>
<td>47.21</td>
</tr>
<tr>
<td>Ours</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
<td>22.07</td>
<td>17.10</td>
<td>17.58</td>
<td>42.72</td>
<td>65.92</td>
<td>48.29</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>24.87</td>
<td>18.34</td>
<td>19.19</td>
<td>43.81</td>
<td>66.35</td>
<td>48.53</td>
</tr>
</tbody>
</table>

6. Results

Coreference resolution. In Table 4, we report the CR performance of the baselines and our method. Our method significantly outperforms all the text-only and vision-text baselines on all the metrics. The text-only CR baselines in the first three rows fail to effectively resolve conferences from narrations. It is important to note that a relatively high number in BLANC scores (compared to MUC) occurs because this measure also counts non-coreferent links (i.e., mentions that are not paired with anything), whereas MUC only measures pairs that are resolved.

The rule-based method [29] uses exact match noun phrases, pronoun-noun matches, and the distance between mentions as hard constraints. It achieves low scores on all metrics and especially on BLANC. The reason for this is the limitation of the rule-based heuristics: For instance, in long narrations, if a pronoun such as she occurs farther to its referent (e.g., the woman) than the predefined distance, it will not form a coreference chain. In contrast, as we apply rules as a soft constraint, we are able to make more flexible decisions from our method. Neural-Coref [30], a deep network on a pre-trained large-corpus of labeled CR data, obtains low scores on CIN for both MUC and BLANC. This is due to the large domain gap between the source and target data as well as the ambiguity in resolving the mentions without visual cues. Similar observations are made when pre-trained CR methods are applied to other domains such as biomedical text [41] or social media [3]. Lastly, the similarity-based baseline performs poorly, as the utilized off-the-shelf word vectors are not trained to cluster corefering mentions. The relatively high scores on BLANC are due to the frequent non-coreferents in our narratives. This kind of approach clusters words with similar meanings together e.g., woman and another woman (both representing female entities) or he and she (both pronouns).

Next, we compare our method to the visual grounding baselines that use both image and text input. Our method also outperforms these baselines: Though GLIP is pre-trained on large-scale data with ground-truth boxes for each object in captions, these captions are usually short and do not contain multiple mentions of entities, unlike in our data. Hence GLIP acts more like an object detector, fails to link coreferring pairs (low MUC scores) and merely identifies singletons (higher BLANC scores). While it is nontrivial to finetune GLIP on our data without groundtruth boxes, we denote this as MAF†. This is the strongest baseline on our task, as training it on narrations including the pronouns reduces the domain gap and enables it to resolve coreferences well. However, this method obtains low precision by incorrectly linking visually similar mentions (that do not belong together) such as trees, plant, flowers. When the training is regularized with the linguistic priors from our method, denoted as MAF++, its performance significantly improves on both MUC and BLANC. The constraint helps to push away the negative mentions (trees, plant, etc.) and encourages the model to learn unique embeddings for them. Due to the self-attention in the transformer architectures, Ours without mouse traces (MT) achieves better performance than MAF++, a simple MLP baseline. The performance difference between our method without using mouse traces and MAF++ can be explained by the better architecture described previously. Finally, our method achieves the best performance gains in CR thanks to the mouse traces and improved architecture over MAF.

Ablation on mouse traces In Table 4, we also analyze the contribution of modeling mouse traces (second last row). Adding the mouse traces improves performance on CR
across all metrics. We hypothesize that the mouse traces provide a strong discriminative location prior to the textual mentions, which helps the model to learn a better compatibility score. To visualize qualitatively, consider the example in Figure 3, the same mention this person points to two different visual regions – one with the person holding the ball and the other person standing next to the door. In such cases, mouse traces provide a strong signal for disambiguation. But in many cases, mouse traces are noisy and can link mentions that are very close to each other in the image, referring to two different regions. In the above example, mouse traces for these persons and this person have a significant overlap and hence act as a noisy prior. Therefore, without the visual/image region features, it is very challenging to address the problem with mouse traces alone.

**Narrative grounding** Not only does our method show performance gains on CR but also outperforms the baselines on another challenging task of narrative grounding. Table 6 compares results from our methods and baselines. We directly compare with the weakly supervised method for a fair comparison. MAF† [60] is originally evaluated on the Flickr30k-Entities [46] dataset where the textual descriptions are significantly shorter (i.e., single sentence) than the image narrations in our dataset. The performance of MAF on our dataset is significantly lower (21% vs 61% on Flickr30k-Entities), which indicates that narrative grounding is a challenge in itself and cannot be addressed off-the-shelf by phrase grounding methods. When trained with the regularizer, the localization performance improves for both nouns and pronouns with our method and MAF++. With the help of regularization, the model learns to attend to different regions of the image for semantically similar mentions as they might be two separate entities (e.g., five people and the people in Fig. 4).

**Further ablations.** In Table 5, we start by exploring the impact of training using our proposed architecture without incorporating mouse traces and the regularizer. This leads to a decrease in both coreference (CR) and grounding performance. Although the model manages to capture certain coreference links, it also generates a noticeable number of incorrect associations (resulting in lower precision scores, row 1) when compared to the model trained with mouse traces (row 2).

In subsequent rows, we investigate the effects of using different regularizers during training. Notably, using the Frobenius norm as a constraint brings improvements in performance, in contrast to using L1 and mean squared error (MSE) regularizers. This improvement can be attributed to the Frobenius norm’s ability to impose a more solid constraint on the learned coreference matrix. It’s important to note that the final row in the table corresponds to our proposed model, which combines mouse traces and the Frobenius norm regularizer (MT+Frobenius norm).

Table 5: Ablation study with different regularizer types and with/without mouse traces.

<table>
<thead>
<tr>
<th>Method</th>
<th>Reg</th>
<th>CR</th>
<th>Grounding</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAF† [60]</td>
<td>✗</td>
<td>21.60</td>
<td>18.31</td>
</tr>
<tr>
<td>MAF++</td>
<td>✓</td>
<td>25.58</td>
<td>22.36</td>
</tr>
<tr>
<td>Ours</td>
<td>✓</td>
<td>27.62</td>
<td>23.46</td>
</tr>
</tbody>
</table>

Table 6: Grounding accuracy (%) for noun phrases and pronouns and the overall accuracy on the CIN dataset.

Table 7: Our method with/without cross attention.

In Tab. 7, we compare the performance of our final method under two settings: (1) directly averaging the word features or (2) attending over the words by using the image as the query as discussed in Sec. 4. The MUC-F1 and narrative grounding scores are 17.02 and 28.83 respectively for the (1) setting and 19.19 and 29.36 respectively for the (2) setting. Both the narrative grounding accuracy and the coreference evaluation get a boost in performance for visually aware word features. More often than not, the word phrases are relatively short (e.g., the machine) and hence the model does not always learn to disambiguate better with attention for the grounding. On the other hand, this technique
in the middle of the picture, we see a person who is wearing the costumes is walking on the road[1]. at the bottom, we see a baby trolley, and a baby is sitting in the baby trolley.

in front of them, we see a baby trolley, and a baby is sitting in the baby trolley.

Figure 4: Qualitative results of predictions on the CIN dataset. The colored mentions in the text indicate the ground truth coreference chains. The solid and dotted bounding boxes on the image denote the correct and incorrect grounding respectively for our proposed method. We also show the predicted coreference chains for our final method with and without regularizer.

is especially useful for CR because the flow of visual information to the word features acts as a prior to cluster mentions that refer to the same region but with are referred to with different mentions/entities in the text (e.g. the machine and an equipment).

Qualitative results Figure 4 qualitatively analyzes CR and narrative grounding. We visualize the narrative grounding results from our proposed method on the images. The model correctly resolves and localizes phrases such as a person, who, the people, them and a baby trolley, the baby trolley. Whereas, the model fails to ground and chain the instance a baby. It is interesting to note that our model pairs an object and water sprinkler, thereby resolving ambiguity in what the object might refer to. But it fails to add which to this coreference chain. Moreover, without the language regularizer, our method fails to link them to the people. It is very hard to learn coreferences for these pronouns as they come with a weak language prior and hence are difficult for the model to disambiguate. Our model (without regularization) misses the referring expression of the baby trolley to refer to the instance of the trolley before. With the help of rules (e.g. last token match), we can resolve these pairs more often than not. Hence, we clearly show the challenging problem of coreferences we are dealing with and indicate the great potential for developing models with strong contextual reasoning.

7. Conclusion

We introduce a novel task of resolving coreferences in image narrations, clustering mention pairs referring to the same entity. For benchmarking and enabling the progress, we introduce a dataset – CIN – that contains images with narrations annotated with coreference chains and their grounding in the images. We formulate the problem of learning CR by using weak supervision from image-text pairs to disambiguate coreference chains and linguistic priors to avoid learning grammatically wrong chains.

We demonstrate strong experimental results in multiple settings. In the future, we plan to address the noise induced by the language rules during learning and also reduce the errors coming from the mouse traces. We hope that our proposed task definition, dataset, and the weakly supervised method will advance the research in multi-modal understanding.

Acknowledgement: AG is supported by the Armeane Choksi Scholarship and HB is supported by the EPSRC programme grant Visual AI EP/T02572/1 and HB and FK are supported by Edinburgh Laboratory for Integrated Artificial Intelligence (ELIAI). This research/project is supported by the National Research Foundation, Singapore, under its NRF Fellowship (Award NRF-NRFF14-2022-0001). We thank the anonymous reviewers for their constructive feedback.

References


[59] Liwei Wang, Jing Huang, Yin Li, Kun Xu, Zhengyuan Yang, and Dong Yu. Improving weakly supervised visual grounding by contrastive knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14090–14100, 2021. 2


