Combining Vision and Language Representations for Patch-based Identification of Lexico-Semantic Relations

Prince Jha*
princekumar_1901cs42@iitp.ac.in
Indian Intitute of Technology Patna
India

Gaël Dias
gael.dias@unicaen.fr
Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC
France

Alexis Lechervy
alexis.lechervy@unicaen.fr
Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC
France

Jose G. Moreno
jose.moreno@irit.fr
Université de Toulouse, IRIT UMR 5505 CNRS
France

Anubhav Jangra†*
anubhav0603@gmail.com
Indian Institute of Technology Patna
India

Sebastião Pais
sebastiao@di.ubi.pt
University of Beira Interior
Portugal

Sriparna Saha
sriparna@iitp.ac.in
Indian Institute of Technology Patna
India

ABSTRACT

Although a wide range of applications have been proposed in the field of multimodal natural language processing, very few works have been tackling multimodal relational lexical semantics. In this paper, we propose the first attempt to identify lexico-semantic relations with visual clues, which embody linguistic phenomena such as synonymy, co-hyponymy or hypernymy. While traditional methods take advantage of the paradigmatic approach or/and the distributional hypothesis, we hypothesize that visual information can supplement the textual information, relying on the apperceptum subcomponent of the semiotic textology linguistic theory. For that purpose, we automatically extend two gold-standard datasets with visual information, and develop different fusion techniques to combine textual and visual modalities following the patch-based strategy. Experimental results over the multimodal datasets show that the visual information can supplement the missing semantics of textual encodings with reliable performance improvements[1].

CCS CONCEPTS

• **Computing methodologies** → **Lexical semantics**; **Image representations**; **Supervised learning by classification**.

*Work done during internship at Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC.

†Now at Google Research.

[1]Code and datasets are available at https://github.com/Jhaprince/Combining-Vision-and-Language-Representations-for-Patch-based-Identification-of-Lexico-Semantic-Rela

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3548299

KEYWORDS

Lexico-semantic relations, multimodal representations, early and hybrid fusion techniques, multimodal patch-based classification.

ACM Reference Format:
Prince Jha, Gaël Dias, Alexis Lechervy, Jose G. Moreno, Anubhav Jangra, Sebastião Pais, and Sriparna Saha. 2022. Combining Vision and Language Representations for Patch-based Identification of Lexico-Semantic Relations. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3503161.3548299

1 INTRODUCTION

The ability to automatically identify lexico-semantic relations is an important issue for information retrieval and natural language processing applications such as question answering [14], query expansion [25], or text summarization [16]. Lexico-semantic relations embody linguistic phenomena such as synonymy (e.g. phone ↔ telephone), co-hyponymy (e.g. phone ↔ monitor), hypernymy (e.g. phone → speakerphone), but more can be enumerated [63]. To tackle this task, different strategies have been proposed that either define new specific features [1, 50, 59], build specific latent semantic spaces [37, 46, 64], conceptualize multitask architectures [3, 4], or augment input data with textual information [6, 23].

Although many different ideas have been proposed to classify whether two words are in lexico-semantic relation or not, two different input text representations have mostly been used. On the one hand, the paradigmatic approach represents the input data as the lexico-syntactic patterns that connect the two words in a pair [19, 27, 38, 48, 52, 55]. On the other hand, the distributional approach consists in characterizing the semantic relation that exists between two words based on their n-dimensional individual representations [7, 15, 18, 47, 52, 62, 63, 65].

Interestingly, some recent studies have emerged that tackle vision-grounded natural language representations [9, 28, 33, 34, 45] and applications [2, 21, 31, 35, 51, 56]. This idea is founded on the

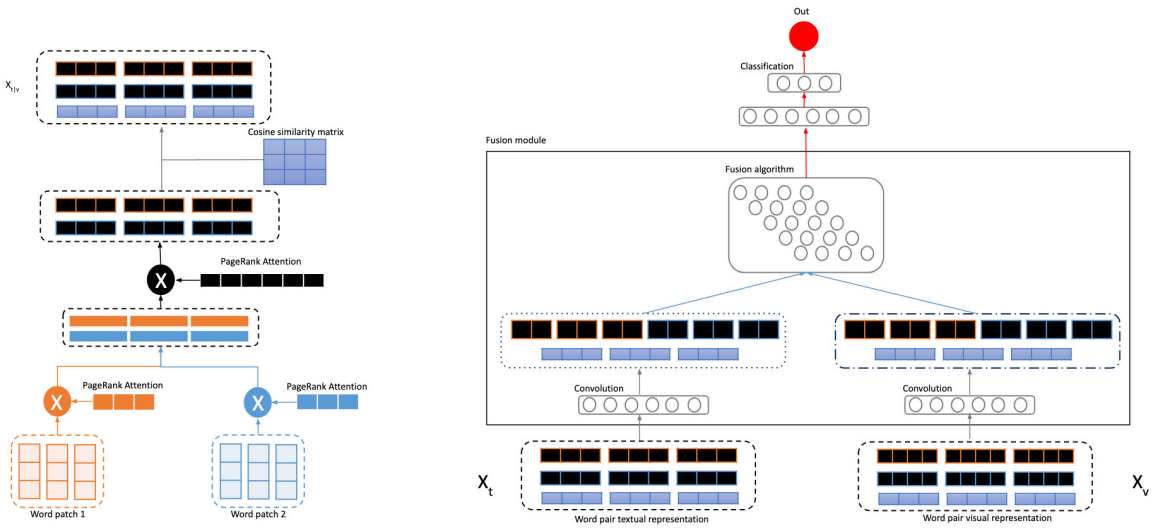

Figure 1: On the left, the patch-based architecture for individual modalities. On the right, the overall multimodal framework.

semiotic textology linguistic theory [17], which lists three subcomponents in order to consider how each textual media produces meaning and the relation between them: dictum (aka. denotation), evocatum (aka. as connotation), and apperceptum (mental images), the latter one embodying the vision-grounded analysis of textual content. However, little research has been endeavoured that combines textual and visual information for relational textual data, to the exception of recent studies on prepositional phrase attachment [11] and relation extraction for knowledge graphs [68].

In this paper, we propose the first attempt to use visual information to identify lexico-semantic relations between word pairs. In particular, we first augment two gold-standard datasets (RUMEN [4] and ROOT9 [49]) with visual information automatically gathered from a search engine. Then, two different fusion techniques, one based on attention fusion [22] and another one based on CentralNet [57], are experimented to combine the textual and visual modalities, where the textual distributional representations are encoded with GloVE [40], and the visual representations are encoded with VGG19 [54]. In order to take advantage of recent multimodal representations, we also propose to encode both modalities with CLIP [45] encodings. Finally, we test our hypothesis following the augmentation data paradigm proposed by [6], by increasing the initial words by their K most similar neighbors within some text representation space, here GloVE, which are then further combined with their visual information. Experiments over the extended multimodal datasets demonstrate that introducing visual information can supplement the missing semantics of textual information with reliable performance improvements.

2 RELATED WORK

Lexico-semantic relation identification. Four major research directions have been proposed for the identification of lexico-semantic relations: (1) feature engineering, (2) fine-tuned semantic spaces, (3) multitask architectures and (4) data augmentation. Within the first topic, [29, 63] propose similar evaluations to combine word input vectors. In particular, word pairs are encoded as the concatenation of the constituent word representations, their vector difference or their sum. [38, 52] propose to overcome domain dependency by representing contextual patterns as continuous vectors, thus successfully combining the paradigmatic approach with the distributional hypothesis. [1, 59] compute specific features over the distributional space (e.g. cosine similarity) in addition to the vector representations themselves, leading to significant improvements. The second research direction aims to build fine-tuned neural latent semantic spaces that embody relational information. [37, 60] learn new embeddings from a background knowlegde of word pairs. To generalize this idea, [8, 24, 64] learn explicit specialization functions that are further injected in the embedding learning process. The third approach tackles this task from the architecture point of view. As semantic relations are known to be closely semantically related, it is likely that multitask learning may improve the decision process. For that purpose, [3] propose a coarse-grained model through a multitask convolutional neural network, while [4] propose a fine-grained methodology, which aims to determine whether the learning process of a given semantic relation can be improved by the concurrent learning of another relation. The fourth strategy aims to augment the initial word pair input with semantically close terms. Within this context, [23] propose a set cardinality-based method, which exploits the WordNet [36] graph, while [6] define a patch-based approach, which augments each constituent word from a latent semantic space.

Vision-grounded language applications and representations. The combination of new multimodal datasets [42] with the definition of new multimodal machine learning models [5] has fostered research in the broad field of multimodal natural language processing [20] and multimodal computer vision [66]. In particular, multimodal machine learning has enabled a wide range of applications, such as multimedia content indexing and retrieval [10], video summarization [51], multimodal sentiment [56] and emotion

[35] analysis, visual question answering [2], image captioning [21], and multimodal dialogue systems [31], to name but a few. Another research direction aims to learn how to represent and summarize multimodal data in a way that exploits the complementarity and redundancy of multiple modalities. In the specifc field of vision-grounded language representations, different models have been proposed [9, 28, 33, 34, 45]. [28] extend the skip-gram model by taking visual information into account. As such, for a restricted set of words, the model is exposed to the visual representations of the objects they denote, and must predict linguistic and visual features jointly. [34] extend the BERT architecture [13] to a multimodal two-stream model by processing both visual and textual inputs in separate streams that interact through co-attentional transformer layers. [9] introduce UNITER, which includes four pretraining tasks over transformers: masked language modeling conditioned on image, masked region modeling conditioned on text, image-text matching, and word-region alignment. [33] present DiMBERT, which takes both visual features from images and textual features from sentences as input, and then apply a single cross-modal transformer to learn vision-language grounded representations. [45] study the behaviors of image classifiers trained with natural language supervision at large scale. Enabled by the large amounts of publicly available data of image-text form on the internet, they create a new dataset of 400 million pairs and demonstrate that a simplified version of ConVIRT [67] trained from scratch, which they call CLIP, is an efficient method of learning from natural language supervision.

Multimodal lexical semantics. Although a wide range of applications have been proposed in the wide field of multimodal natural language processing, very few works have been tackling multimodal relational lexical semantics, with the rare exceptions of [11, 68]. [11] propose to score alternative prepositional phrase attachments from the caption of an image, previously syntactically-parsed, based on how much the attachments are coherent with the corresponding image. The set of attachments that yields the best score is identified and the corresponding tree is output. [68] present the multimodal relation extraction task that consists in identifying the semantic relations that link two entities in a sentence with visual clues. For that purpose, they propose a multimodal neural network with a graph alignment method that incorporates structural similarity and semantic agreement between visual objects in an image and textual entities in a sentence. Experiments show that improved results can be obtained compared to the concatenation of visual and textual representations. In this paper, we present the first study that tackles multimodal lexico-semantic relation identification.

3 MULTIMODAL METHODOLOGY

The main task at hand consists in deciding whether a given lexico-semantic relation (i.e. synonymy, hypernymy, co-hyponymy) holds between a pair of words (w_0, w_1) or not (i.e. random). For that purpose, we present our methodology, illustrated in Figure 1, which consists in adapting the patch-based approach proposed by [6] in a multimodal environment, thus relying on fusion techniques.

Patch-based Representation. The idea of patch-based classification has been introduced by [6, 23], and consists in augmenting each word in a pair with its K most semantically-related words in some semantic space. While [23] use WordNet for the augmentation, [6] rely on GloVE embeddings. Based on our experiments, we follow the strategy of [6] as it outperforms the one of [23].

Formally, a patch consists of the K most similar words w_j to a source word w_0 in terms of cosine similarity in some latent semantic space, and it is defined in Equation 1. Thus, each input pair (w_0, w_1) is transformed into its patch-based representation $(P^K_{w_0}, P^K_{w_1})$.

$$P^K_{w_0} = \{w_0\} \cup \left\{ w_j | \operatorname*{argmax}^K \cos(w_0, w_j) \right\} \tag{1}$$

All words within a patch are then subject to a fixed attention mechanism, which integrates the notion of centrality. This ensures that the most central words within a patch receive higher attention. This process is performed through the PageRank algorithm [39] over the undirected weighted[2] patch graph, which results in a vector of $(K+1)$ dimensions, where each word within the patch receives a centrality score in $\mathbb{R}$, and it is noted $\langle \alpha_{w_0^0}, \alpha_{w_0^1}, \alpha_{w_0^2}, \ldots, \alpha_{w_0^K} \rangle$.

A second attention mechanism spotlights on word centrality between patches to acknowledge, which words are central to both concepts. The same process is applied with the PageRank algorithm based on the graph that comprises of all $2\times(K+1)$ words as vertices and links all vertices belonging to different patches. This process results in a vector of $2 \times (K+1)$ dimensions, where each word of both patches receives a centrality score in $\mathbb{R}$, and it is noted $\langle \beta_{w_0^1}, \beta_{w_0^2}, \ldots, \beta_{w_0^K}, \beta_{w_1^0}, \beta_{w_1^1}, \ldots, \beta_{w_1^K} \rangle$.

Both attention mechanisms are then combined into a unique learning representation, which is defined in Expression 2, where w^i_x represents a word embedding of patch $P^K_{w_x}$, and $\oplus$ is the concatenation operator. Note that the embeddings are in descending order of cosine similarity with their source word.

$$A_1 = \left(\bigoplus_{i=0}^{K} \alpha_{w_0^i} . \beta_{w_0^i} . w_0^i \right) \oplus \left(\bigoplus_{i=0}^{K} \alpha_{w_1^i} . \beta_{w_1^i} . w_1^i \right) \tag{2}$$

In order to account for domain independence [61], the cosine similarity is measured between all components of both patches, which concatenation is defined in Expression 3.

$$A_2 = \bigoplus_{i=0}^{K} \bigoplus_{j=0}^{K} cos(w_0^i, w_1^j) \tag{3}$$

Finally, each input pair (w_0, w_1) receives two different learning representations, namely X_t for the textual modality and X_v for the visual modality, generically defined in Equation 4. Such representations are then fed to the multimodal fusion module.

$$X_{t|v} = A_1 \oplus A_2 \tag{4}$$

While Bannour et al. [6] exclusively focus on textual data augmentation, we need to deal with visual augmentation. For that purpose, we propose that textual data drives the augmentation process[3]. As such, each word pair (w_0, w_1) is transformed into its patch-based representation $(P^K_{w_0}, P^K_{w_1})$, based on finding the K most similar words within some textual semantic space, here GloVE [40], in terms of cosine similarity. Then, each word present in a patch

[2]The weight corresponds to the cosine similarity value.
[3]Other strategies are possible but they remain for future work.

is sent to a search engine, here the Bing Image Search API[4], and the highest ranked image returned by the search engine is taken as the augmented visual information (cf. §4 for more details). Once visual augmentation is performed, the process of [6] is replicated in the exact same way for the visual information but relying on visual n-dimensional representations, VGG19 [53] or CLIP [45].

Multimodal fusion networks. In order to reduce each modality representation X_t and X_v to the same dimension, a reduction process is first performed. Then, two fusion techniques are implemented to combine modalities: early [22] and hybrid fusions [58].

Attention fusion network. Both visual and textual modalities may not equally be relevant for the identification of lexico-semantic relations. This motivates the introduction of an attention fusion network, which weights each modality independently, in the same line of [41, 43]. The attention fusion network is shown in Figure 2.

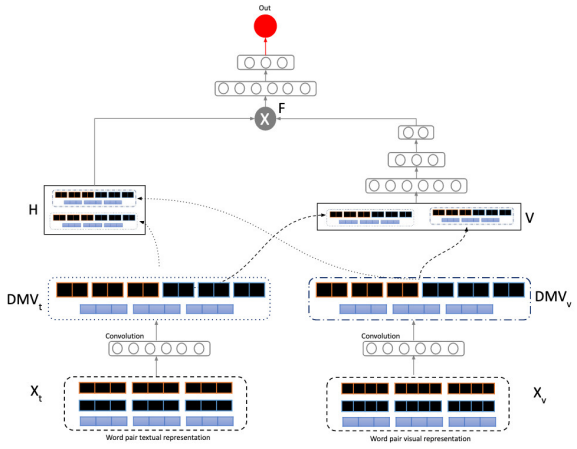

Figure 2: Attention fusion network.

Formally, the input to the attention fusion network is noted MV, the set of modality representations, where the dimension of a modality vector $MV_k \in MV$ is d_k. The first step consists in giving the same dimension d to all the elements of MV. This process is referred to a reduction process, and it is done using a stack of dense layers. The resultant vectors are denoted DMV, such that the reduced modality representations $DMV_k \in DMV$. All DMV_k are then concatenated into a vector V, which is passed through a set of dense layers followed by sigmoid activation layer to calculate attention scores. These attention values weight each modality, and the resulting modality representations are concatenated to build the early fusion vector F. This process is recaped in Equation 5.

$$\begin{aligned} MV &= [MV_t = X_t, MV_v = X_v] \\ DMV_i &= ReLU(W^T_{MV_i} MV_i + b) \\ Att_i &= \sigma(W^T_{DMV_i} V + b) \\ F &= Att_t.V_t \bigoplus Att_v.V_v \end{aligned} \tag{5}$$

[4]https://www.microsoft.com/en-us/bing/apis/bing-image-search-api

F is then passed through further dense layers for the decision process, and the categorical cross-entropy loss function $L_{CE}(\hat{y}, y)$ is used to train the network parameters, where $\hat{y}_i^j$ is the predicted label and y_i^j is the true label.

$$L_{CE}(\hat{y}, y) = -\frac{1}{N} \sum_{j=1}^{C} \sum_{i=1}^{N} y_i^j log(\hat{y}_i^j) \tag{6}$$

CentralNet fusion network. CentralNet [58] is a hybrid fusion network, which mixes early and late fusions into a single architecture as illustrated in Figure 3.

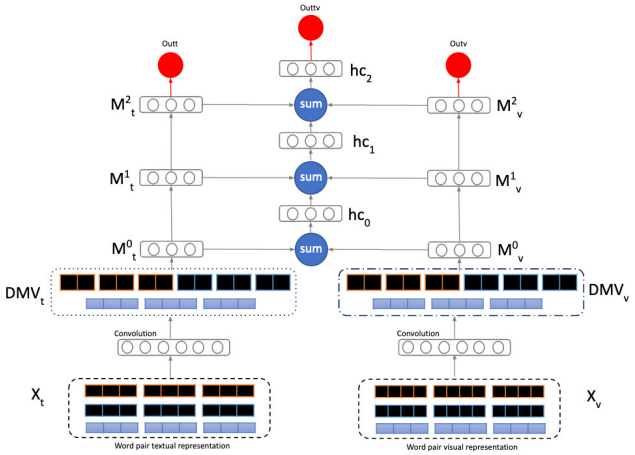

Figure 3: CentralNet fusion network.

The architecture consists of k independent networks corresponding to each modality, and one central network. In particular, the central network combines the features generated from the different modalities by considering the weighted sum of unimodal hidden representations, and its own previous layer. Such fusion layers are defined in Equation 7, where α_p are scalar trainable weights, M_k^i is the hidden representation of k^{th} modality at the i^{th} layer, and hc_i is the central hidden representation at the i^{th} layer. Note that fusing at a low-level layer stands for early fusion, while fusing at a last layer means late fusion.

$$hc_{i+1} = ReLU\left(\alpha_c hc_i + \sum_{k=t}^{v} \alpha_{M_i^k} M_i^k\right) \tag{7}$$

Each layer hc_{i+1} is fed to an operating layer composed of a dense layer followed by a $ReLU$ activation function. Note that the input to the first layer of the central network is only the weighted sum of the modalities hidden representations $M_0^t = DMV_t$ and $M_0^v = DMV_v$ as there is no previous central hidden representation. The final output representation of the central network represents the fusion vector F, which is used for the final prediction. In particular, we employ the categorical cross-entropy loss function (Equation 6) to train the network parameters, and the final $Loss$ function is defined in Equation 8, where L_{CE}^C is the loss computed from the output of central network, and $L_{CE}^{M^k}$ is the loss of modality k.

$$Loss(\hat{y}, y) = L_{CE}^{C}(\hat{y}, y) \; + \; \sum_{k=t}^{v} L_{CE}^{M^k}(\hat{y}, y) \tag{8}$$

Note that our model differs from the one presented in Vielzeuf et al. [58] in the sense that each unimodal network is first pre-trained independently, and then frozen to learn the central network. As such, only the central network is trainable, and the remaining parts of the architecture are kept non-trainable, i.e. frozen. Indeed, the frozen architecture showed stronger performances compared to the all-trainable model for the sake of our experiments.

Implementation details and experimental setups of all the modules of the methodology are given in Apppendix A.

4 NEW DATASETS: IXRUMEN AND IXROOT9

Since the community lacks a multimodal dataset for the task of lexico-semantic relation identification, we propose the extension of two gold-standard datasets, namely RUMEN [4] and ROOT9 [49]. RUMEN is a dataset comprising of 3213 instances for synonymy detection and 3375 instances for hypernymy detection, whereas ROOT9 comprises of 1636 instances for co-hyponymy detection and 1256 instances for hypernymy detection. As we follow the patch-based data augmentation strategy due to its empirical effectiveness, where each word instance is augmented by its K-nearest neighbors in the GloVE [40] embedding space, the visual augmentation must deal with the original words within the pair plus the K augmented words that form the respective patches.

To extend the two datasets in a multimodal setting, we propose to scrap the web for exemplar images by using the Bing Image Search API, such that for each of the $K+1$ words within a patch, we download exactly 3 images ordered by their retrieval rank[5]. This multimodal augmentation strategy is performed for a patch size up to $K = 5$, and we adopt lexical split [4, 6, 30], which avoids vocabulary intersection between the train and test splits, thus bypassing the lexical memorization issue [30]. As a consequence, it is clear that some of the initial word pairs contained in RUMEN and ROOT9 must be withdrawn from their original datasets, if they cannot provide up to 3 visual clues. The statistics for the image-extended RUMEN dataset (IxRUMEN) and the image-extended ROOT9 dataset (IxROOT9) can be found in Table 1.

Dataset	Train	Test	Total
RUMEN (Synonym)	2256	957	3213
RUMEN (Hypernym)	2638	737	3375
RUMEN (Random)	2227	969	3196
IxRUMEN (Synonym)	2031	860	2891
IxRUMEN (Hypernym)	2393	648	3041
IxRUMEN (Random)	2006	830	2836
ROOT9 (Co-hyponym)	1070	566	1636
ROOT9 (Hypernym)	826	430	1256
ROOT9 (Random)	381	129	510
IxROOT9 (Co-hyponym)	975	531	1506
IxROOT9 (Hypernym)	717	392	1109
IxROOT9 (Random)	335	103	438

Table 1: Statistics for RUMEN, ROOT9, IxRUMEN and IxROOT9 datasets.

[5]Note that we explored existing large-scale corpora like the MSCOCO dataset [32] for better reproducibility, but due to its limited lexical coverage, an open-domain retrieval strategy was opted for.

To overcome privacy concerns, we decided to release the image encodings for each image in the dataset over the actual image. For that purpose, we use VGG19 [53] and CLIP [44] embeddings. VGG19 [53] shows state-of-the-art performances in image classification tasks. It is 19 layers deep convolutional network, which is pre-trained on ImageNet [12] to predict 1000 object classes. Thus, VGG19 embeddings have the ability to represent robust visual concepts. Here, each image is encoded as a 4096-dimensional vector. CLIP (Contrastive Language-Image Pre-training) [44] is a pre-trained visual-linguistic model that can encode image-text pairs. CLIP was pre-trained on 400 million image-text pairs, where for a given batch of N (image,text) pairs, the model had to predict N correct matches out of $N \times N$ possible pairings. In particular, CLIP maximizes the cosine similarity of N real pairs by training image and text encoders together to create an efficient multimodal embedding space. Here, each image is encoded as a 512-dimensional vector and note that the image information is combined with the textual pattern "*a photo of <source word>*" as suggested in [44] to get full advantage of the contextualized multimodal model.

5 RESULTS AND DISCUSSION

In this section, we first present the results of the unimodal models, where each textual and visual modalities are taken individually for the decision process. Then, we present the results obtained for the early and hybrid fusions. Finally, we present a qualitative analysis that shows the benefits and drawbacks of the multimodal fusion.

5.1 Unimodal Models

Results for unimodal models are given in Table 2 for the textual modality and in Table 3 for the visual modality. For the textual modality, results confirm the findings of [6] and show that the patch-based approach outperforms the baseline strategy, where no word augmentation is performed, i.e. $K = 0$. In particular, larger values of K steadily improve results for the identification of symmetric relations (synonymy and co-hyponymy), while such is not true for asymmetric relations such as hypernymy. This can be explained by the fact that larger values of K might lead to concept shift for the hypernym relation, thus noising the input data. This is particularly true for GloVE embeddings, although such does not stand for CLIP embeddings. Results also show that CLIP multimodal embeddings do not provide a sustainable alternative for the sake of the identification of lexico-semantic relations, as results drastically drop, when compared to text-based embeddings, especially for the case of the IxRUMEN dataset. This can be explained by the fact that CLIP embeddings have been tuned to better represent visual information at the expense of textual information [45]. Moreover, as the IxRUMEN dataset contains a wide spectrum of abstract words [4] (e.g. destiny $\leftrightarrow$ fate), this might lead to difficulties in visually representing such information. As a consequence, multimodal embeddings might not correctly encode this information.

For the visual modality, different situations occur. While the use of VGG19 encodings clearly evidences the positive impact of using the patch-based strategy with steady improvements for high values of K independently of the lexico-semantic relation and the dataset at hand, similar results are not exactly observable for multimodal representations. Indeed, while the use of CLIP multimodal

			Synonym v/s Random				Hypernym v/s Random			
		Patch Size	Accuracy (avg/stddev/max)	F1 Score (avg/stddev/max)	Precision (avg/stddev/max)	Recall (avg/stddev/max)	Accuracy (avg/stddev/max)	F1 Score (avg/stddev/max)	Precision (avg/stddev/max)	Recall (avg/stddev/max)
RUMEN	GLOVE	K=0	80.98/0.35/81.52	80.98/0.35/81.52	80.98/0.35/81.52	80.978/0.35/81.52	79.52/0.15/79.66	79.51/0.15/79.66	79.52/0.15/79.66	79.50/0.16/79.66
		K=1	84.17/0.23/84.53	84.17/0.23/84.52	84.18/0.23/84.54	84.19/0.22/84.53	82.24/0.27/82.47	82.20/0.27/82.44	82.24/0.27/82.47	82.19/0.27/82.43
		K=2	84.81/0.12/84.94	84.80/0.12/84.94	84.81/0.12/84.94	84.84/0.12/84.97	82.13/0.17/84.42	82.05/0.16/82.33	82.13/0.17/82.42	82.09/0.17/82.38
		K=3	83.97/0.07/84.06	83.96/0.07/84.05	83.97/0.07/84.06	84.02/0.07/84.12	82.33/0.19/82.59	82.27/0.19/82.53	82.33/0.19/82.59	82.29/0.19/82.55
		K=4	84.67/0.10/84.79	84.67/0.10/84.78	84.67/0.10/84.79	84.70/0.10/84.82	**82.55**/0.13/82.71	**82.49**/0.13/82.65	**82.55**/0.13/82.71	**82.50**/0.13/82.67
		K=5	**84.90**/0.04/84.94	**84.90**/0.04/84.94	**84.90**/0.04/84.94	**84.91**/0.04/84.95	81.82/0.08/81.95	81.74/0.07/81.86	81.82/0.08/81.95	81.78/0.08/81.91
	CLIP-Text	K=0	56.09/0.13/56.33	56.09/0.13/56.33	56.09/0.13/56.33	56.11/0.13/56.34	54.24/0.28/54.57	51.76/1.69/54.76	54.24/0.28/54.57	52.52/1.50/55.15
		K=1	59.54/0.09/59.61	59.54/0.09/59.60	59.55/0.09/59.61	59.54/0.09/59.60	59.60/0.30/59.85	57.48/0.31/57.75	59.60/0.30/59.85	58.62/0.35/58.91
		K=2	63.66/0.04/63.71	63.66/0.04/63.71	63.66/0.04/63.71	63.66/0.05/63.71	62.56/0.24/62.95	61.37/0.23/61.74	62.56/0.23/62.95	61.98/0.26/62.41
		K=3	64.99/0.13/65.21	64.99/0.13/65.21	64.99/0.13/65.21	64.99/0.13/65.21	**64.37**/0.18/64.54	**62.94**/0.20/63.19	**64.37**/0.18/64.54	64.07/0.21/64.23
		K=4	66.51/0.31/66.87	66.50/0.31/66.87	66.51/0.31/66.87	66.50/0.31/66.87	62.09/0.21/62.31	60.25/0.24/60.51	62.09/0.21/62.31	61.54/0.24/61.79
		K=5	**66.64**/0.10/66.72	**66.63**/0.10/66.71	**66.64**/0.10/66.72	**66.65**/0.10/66.73	64.35/0.19/64.48	62.59/0.20/62.75	64.35/0.19/64.48	**64.19**/0.23/64.39
			Co-hyponym v/s Random				Hypernym v/s Random			
ROOT9	GLOVE	K=0	92.17/0.25/92.46	92.24/0.24/92.56	92.16/0.25/92.46	92.47/0.22/92.74	85.06/0.07/85.11	85.15/0.06/85.2	85.06/0.07/85.11	85.32/0.30/85.54
		K=1	93.72/0.15/93.81	93.72/0.15/93.82	93.72/0.15/93.81	93.73/0.16/93.83	91.10/0.11/91.27	91.04/0.11/91.2	91.10/0.11/91.27	91.02/0.12/91.18
		K=2	93.86/0.07/93.94	93.87/0.08/93.95	93.86/0.07/93.94	93.88/0.08/93.96	91.86/0.15/92.09	91.82/0.15/92.05	91.86/0.15/92.09	91.80/0.15/92.03
		K=3	93.41/0.10/93.54	93.47/0.09/93.60	93.41/0.10/93.54	93.59/0.08/93.70	**92.69**/0.09/92.75	**92.64**/0.09/92.70	**92.69**/0.09/92.75	**92.63**/0.09/92.69
		K=4	93.46/0.12/93.67	93.52/0.12/93.74	93.46/0.12/93.67	93.62/0.13/93.85	92.42/0.12/92.59	92.43/0.11/92.59	92.42/0.12/92.59	92.45/0.11/92.60
		K=5	**94.00**/0.12/94.08	**94.05**/0.16/94.14	**94.00**/0.12/94.08	**94.16**/0.11/94.27	91.96/0.14/92.09	91.98/0.14/92.11	91.96/0.14/92.09	92.00/0.12/92.12
	CLIP-Text	K=0	74.32/0.62/75.24	75.90/0.54/76.68	74.32/0.63/75.24	80.05/0.29/80.34	67.41/0.39/67.87	67.65/0.45/68.2	67.41/0.39/67.87	67.93/0.53/68.58
		K=1	82.26/0.22/82.50	82.97/0.21/83.18	82.26/0.22/82.50	84.51/0.22/84.68	73.01/0.18/73.31	73.20/0.16/73.44	73.01/0.18/73.31	73.43/0.14/73.58
		K=2	86.00/0.17/86.14	86.40/0.15/86.52	86.00/0.17/86.14	87.21/0.11/87.28	75.98/0.14/76.11	76.09/0.13/76.21	75.98/0.14/76.11	76.22/0.11/76.32
		K=3	85.82/0.23/86.14	86.24/0.22/86.56	85.82/0.23/86.14	87.10/0.23/87.46	77.66/0.15/77.76	77.59/0.15/77.74	77.66/0.15/77.76	77.53/0.15/77.72
		K=4	87.92/0.31/88.29	88.26/0.29/88.61	87.92/0.31/88.29	89.02/0.29/89.33	77.92/0.12/78.09	77.89/0.13/78.07	77.92/0.12/78.09	77.85/0.14/78.05
		K=5	**89.10**/0.00/89.10	**89.43**/0.01/89.43	**89.10**/0.00/89.10	**90.24**/0.04/90.26	**79.61**/0.14/79.74	**79.55**/0.11/79.69	**79.61**/0.14/79.74	**79.51**/0.10/79.64

Table 2: Results for IxRUMEN and IxROOT9 based on textual unimodality represented either by GloVE or CLIP embeddings.

			Synonym v/s Random				Hypernym v/s Random			
		Patch Size	Accuracy (avg/stddev/max)	F1 Score (avg/stddev/max)	Precision (avg/stddev/max)	Recall (avg/stddev/max)	Accuracy (avg/stddev/max)	F1 Score (avg/stddev/max)	Precision (avg/stddev/max)	Recall (avg/stddev/max)
RUMEN	VGG19	K=0	52.65 / 0.41 / 53.37	50.56 / 1.04 / 52.35	52.65 / 0.41 / 53.37	53.34 / 0.35 / 53.80	53.52 / 0.28 / 53.75	45.53 / 0.75 / 46.80	53.52 / 0.28 / 53.75	47.53 / 0.78 / 48.76
		K=1	56.49 / 1.51 / 57.84	56.15 / 1.91 / 57.80	56.49 / 1.51 / 57.84	56.59 / 1.37 / 57.90	53.89 / 0.34 / 54.34	50.10 / 0.08 / 50.16	53.89 / 0.34 / 54.34	51.09 / 0.28 / 51.45
		K=2	57.25 / 2.28 / 59.40	57.18 / 2.34 / 59.38	57.25 / 2.34 / 59.38	57.27 / 2.27 / 59.43	57.09 / 0.53 / 57.50	53.43 / 0.59 / 54.04	57.09 / 0.53 / 57.50	55.34 / 0.71 / 55.88
		K=3	59.21 / 0.30 / 59.50	59.17 / 0.31 / 59.46	59.21 / 0.30 / 59.50	59.28 / 0.30 / 59.56	56.62 / 0.62 / 57.56	56.57 / 0.68 / 57.64	56.62 / 0.62 / 57.56	57.21 / 0.78 / 57.91
		K=4	**60.60** / 0.40 / 61.01	**60.56** / 0.39 / 60.95	**60.60** / 0.40 / 61.01	**60.68** / 0.41 / 61.10	**58.49** / 0.16 / 58.68	57.41 / 0.76 / 57.82	**58.49** / 0.16 / 58.68	57.67 / 0.16 / 57.81
		K=5	60.59 / 0.44 / 60.90	**60.56** / 0.46 / 60.88	60.59 / 0.44 / 60.90	60.66 / 0.42 / 60.96	58.38 / 0.38 / 58.73	**57.92** / 0.45 / 58.36	58.38 / 0.38 / 58.73	**57.96** / 0.27 / 58.25
	CLIP-Image	K=0	**77.11** / 0.29 / 77.52	**77.11** / 0.29 / 77.52	**77.11** / 0.29 / 77.52	**77.12** / 0.29 / 77.53	**75.91** / 0.08 / 75.97	**75.94** / 0.08 / 75.99	**75.91** / 0.08 / 75.97	75.97 / 0.07 / 76.02
		K=1	75.41 / 0.20 / 75.65	75.46 / 0.30 / 75.90	75.41 / 0.20 / 75.64	75.46 / 0.21 / 75.72	73.77 / 0.42 / 74.09	73.56 / 0.44 / 73.85	73.77 / 0.42 / 74.09	73.61 / 0.42 / 73.97
		K=2	74.76 / 0.17 / 74.97	74.76 / 0.17 / 74.97	74.76 / 0.17 / 74.97	74.77 / 0.17 / 74.98	74.57 / 0.11 / 74.74	74.22 / 0.12 / 74.41	74.57 / 0.11 / 74.74	74.54 / 0.10 / 74.69
		K=3	74.15 / 0.15 / 74.25	74.15 / 0.15 / 74.25	74.15 / 0.15 / 74.25	74.15 / 0.15 / 74.25	74.86 / 0.07 / 74.91	74.38 / 0.07 / 74.44	74.86 / 0.07 / 74.91	75.01 / 0.06 / 75.07
		K=4	74.91 / 0.16 / 75.08	74.91 / 0.16 / 75.08	74.91 / 0.16 / 75.08	74.91 / 0.16 / 75.08	75.19 / 0.21 / 75.38	75.24 / 0.67 / 75.82	75.19 / 0.21 / 75.38	75.46 / 0.23 / 75.68
		K=5	76.52 / 0.10 / 76.69	76.52 / 0.10 / 76.69	76.52 / 0.10 / 76.69	76.52 / 0.10 / 76.69	75.66 / 0.05 / 75.73	75.40 / 0.50 / 75.94	75.66 / 0.05 / 75.73	**76.18** / 0.04 / 76.22
			Co-hyponym v/s Random				Hypernym v/s Random			
ROOT9	VGG19	K=0	62.19 / 3.13 / 65.95	65.00 / 2.83 / 68.42	62.26 / 3.00 / 65.95	75.46 / 0.74 / 75.97	62.41 / 1.39 / 64.42	60.30 / 0.48 / 60.73	62.41 / 1.39 / 64.42	58.96 / 0.27 / 59.32
		K=1	70.53 / 0.19 / 70.66	72.57 / 0.15 / 72.67	70.53 / 0.19 / 70.66	78.61 / 0.14 / 78.71	61.08 / 0.44 / 61.45	61.66 / 0.46 / 62.02	61.08 / 0.44 / 61.45	62.37 / 0.52 / 62.74
		K=2	77.93 / 0.25 / 78.20	79.08 / 0.20 / 79.30	77.93 / 0.25 / 78.20	81.86 / 0.07 / 81.98	66.39 / 0.20 / 66.56	66.60 / 0.21 / 66.74	66.39 / 0.20 / 66.56	66.82 / 0.24 / 66.95
		K=3	79.41 / 0.25 / 79.81	80.55 / 0.22 / 80.90	79.41 / 0.25 / 79.81	83.67 / 0.11 / 83.85	69.36 / 0.20 / 69.69	69.20 / 0.19 / 69.45	69.36 / 0.20 / 69.69	69.06 / 0.20 / 69.29
		K=4	79.65 / 0.07 / 79.68	80.75 / 0.07 / 80.79	79.65 / 0.07 / 79.68	83.68 / 0.13 / 83.79	70.25 / 0.38 / 70.51	69.84 / 0.39 / 70.17	70.25 / 0.38 / 70.51	69.53 / 0.41 / 69.89
		K=5	**82.00** / 0.06 / 82.10	**82.89** / 0.06 / 83.00	**82.00** / 0.06 / 82.10	**85.29** / 0.12 / 85.43	**73.38** / 0.15 / 73.48	**72.72** / 0.16 / 72.84	**73.38** / 0.15 / 73.48	**72.35** / 0.18 / 72.48
	CLIP-Image	K=0	90.79 / 0.31 / 91.12	90.97 / 0.29 / 91.27	90.79 / 0.31 / 91.12	91.34 / 0.26 / 91.59	84.48 / 0.39 / 85.17	84.55 / 0.37 / 85.20	84.48 / 0.39 / 85.17	84.63 / 0.34 / 85.22
		K=1	91.04 / 0.20 / 91.25	91.23 / 0.19 / 91.43	91.04 / 0.20 / 91.25	91.67 / 0.15 / 91.81	86.79 / 0.32 / 87.15	86.82 / 0.31 / 87.17	86.79 / 0.32 / 87.15	86.85 / 0.30 / 87.19
		K=2	92.38 / 0.20 / 92.73	92.46 / 0.20 / 92.80	92.38 / 0.20 / 92.73	92.59 / 0.19 / 92.91	**88.37** / 0.09 / 88.47	**88.37** / 0.09 / 88.47	88.34 / 0.09 / 88.47	**88.37** / 0.09 / 88.47
		K=3	91.79 / 0.13 / 91.92	91.92 / 0.13 / 92.05	91.79 / 0.13 / 91.92	92.19 / 0.11 / 92.32	88.04 / 0.09 / 88.14	88.14 / 0.08 / 88.23	88.04 / 0.09 / 88.14	88.30 / 0.08 / 88.38
		K=4	**92.47** / 0.18 / 92.76	**92.52** / 0.07 / 92.58	**92.41** / 0.07 / 92.46	**92.80** / 0.05 / 92.84	87.77 / 0.22 / 88.14	87.81 / 0.21 / 88.16	87.77 / 0.22 / 88.14	87.85 / 0.21 / 88.18
		K=5	91.63 / 0.11 / 91.79	91.80 / 0.11 / 91.96	91.63 / 0.11 / 91.79	92.16 / 0.11 / 92.33	**88.37** / 0.19 / 88.63	88.32 / 0.17 / 88.56	**88.37** / 0.19 / 88.63	88.29 / 0.17 / 88.52

Table 3: Results for IxRUMEN and IxROOT9 based on visual unimodality represented either by VGG19 or CLIP embeddings.

representations outperforms overall results compared to VGG19 encodings for a great margin, the impact of the patch-based strategy is more mitigated. For the IxRUMEN dataset, it is clear that the use of patches seems to be counter-productive, while the contrary is true for the IxROOT9 dataset, with best values obtained for higher values of K. Nevertheless, by looking closely at the results for IxRUMEN, we can observe that values for $K = 5$ are close to the ones of $K = 0$, although intermediate values of K show lower performances, but with a small increasing tendency along with higher values of K. We can then hypothesize that if higher values of K had been tested, CLIP representations would benefit more from the patch-based strategy.

It is clear that relying exclusively on visual information does not compete with text-based strategies, especially when the datasets include abstract words, i.e. IxRUMEN. Indeed, the results differences between the textual modality and the visual modality range from 7.79 points in F1 score for synonymy in IxRUMEN in favor of the textual modality to 1.53 points improvements of the textual modality in terms of F1 score for co-hyponymy in IxROOT9. This situation was expected due to the uncontrolled process of gathering extra visual information with search engine queries. Nevertheless,

			Synonym v/s Random				Hypernym v/s Random			
		Patch Size	Accuracy (avg/stddev/max)	F1 Score (avg/stddev/max)	Precision (avg/stddev/max)	Recall (avg/stddev/max)	Accuracy (avg/stddev/max)	F1 Score (avg/stddev/max)	Precision (avg/stddev/max)	Recall (avg/stddev/max)
RUMEN	GLOVE+CLIP-Img	K=0	82.97 / 0.09 / 83.07	82.96 / 0.09 / 83.06	82.97 / 0.09 / 83.07	83.00 / 0.11 / 83.14	81.32 / 0.60 / 81.95	81.25 / 0.60 / 81.87	81.32 / 0.60 / 81.95	81.28 / 0.61 / 81.90
		K=1	**86.62** / 0.19 / 86.86	**86.61** / 0.19 / 86.85	**86.62** / 0.19 / 86.86	**86.72** / 0.18 / 86.96	82.77 / 0.21 / 83.06	82.62 / 0.21 / 82.89	82.77 / 0.21 / 83.06	82.83 / 0.23 / 83.18
		K=2	86.44 / 0.07 / 86.50	86.42 / 0.07 / 86.49	86.44 / 0.07 / 86.50	86.57 / 0.08 / 86.64	83.54 / 0.20 / 83.76	83.37 / 0.19 / 83.60	83.54 / 0.20 / 83.76	83.68 / 0.23 / 83.92
		K=3	85.09 / 0.21 / 85.25	85.06 / 0.21 / 85.23	85.09 / 0.21 / 85.25	85.26 / 0.22 / 85.44	83.48 / 0.16 / 83.76	83.30 / 0.16 / 83.57	83.48 / 0.16 / 83.76	83.62 / 0.23 / 83.99
		K=4	85.51 / 0.14 / 85.67	85.49 / 0.15 / 85.65	85.51 / 0.14 / 85.67	85.68 / 0.13 / 85.81	**84.07** / 0.08 / 84.17	**83.93** / 0.08 / 84.03	**84.07** / 0.08 / 84.17	**84.17** / 0.10 / 84.28
		K=5	85.50 / 0.14 / 85.72	85.48 / 0.14 / 85.70	85.50 / 0.14 / 85.72	85.68 / 0.14 / 85.90	82.89 / 0.12 / 83.06	82.73 / 0.11 / 82.90	82.90 / 0.11 / 83.06	82.97 / 0.13 / 83.16
			Co-hyponym v/s Random				Hypernym v/s Random			
ROOT9	GLOVE+CLIP-Img	K=0	94.51 / 0.31 / 97.89	94.54 / 0.30 / 94.91	94.51 / 0.31 / 94.89	94.57 / 0.28 / 94.93	87.64 / 0.26 / 87.97	87.62 / 0.30 / 88.00	87.64 / 0.26 / 87.97	87.61 / 0.33 / 88.04
		K=1	94.80 / 0.20 / 95.15	94.85 / 0.21 / 95.20	94.80 / 0.20 / 95.15	94.91 / 0.22 / 95.29	92.26 / 0.20 / 92.42	92.25 / 0.20 / 92.41	92.26 / 0.20 / 92.42	92.24 / 0.20 / 92.40
		K=2	**95.07** / 0.15 / 95.29	**95.11** / 0.15 / 95.33	**95.07** / 0.15 / 95.29	**95.19** / 0.14 / 95.40	**92.95** / 0.22 / 93.25	**92.92** / 0.22 / 93.22	**92.95** / 0.22 / 93.25	**92.91** / 0.22 / 93.20
		K=3	94.86 / 0.11 / 95.02	94.91 / 0.11 / 95.06	94.86 / 0.11 / 95.02	95.01 / 0.10 / 95.14	92.75 / 0.17 / 92.92	92.73 / 0.17 / 92.91	92.75 / 0.17 / 92.92	92.72 / 0.17 / 92.90
		K=4	94.89 / 0.10 / 95.02	94.94 / 0.09 / 95.07	94.89 / 0.10 / 95.02	95.07 / 0.06 / 95.17	92.62 / 0.07 / 92.75	92.63 / 0.08 / 92.77	92.62 / 0.07 / 92.75	92.66 / 0.08 / 92.81
		K=5	94.99 / 0.11 / 95.15	95.04 / 0.12 / 95.21	94.99 / 0.11 / 94.89	95.14 / 0.13 / 95.32	92.69 / 0.15 / 92.92	92.70 / 0.15 / 92.93	92.64 / 0.16 / 92.92	92.73 / 0.16 / 92.96

Table 4: Results for IxRUMEN and IxROOT9 for early fusion with the attention fusion network.

			Synonym v/s Random				Hypernym v/s Random			
		Patch Size	Accuracy (avg/stddev/max)	F1 Score (avg/stddev/max)	Precision (avg/stddev/max)	Recall (avg/stddev/max)	Accuracy (avg/stddev/max)	F1 Score (avg/stddev/max)	Precision (avg/stddev/max)	Recall (avg/stddev/max)
RUMEN	GLOVE+CLIP-Img	K=0	81.67 / 0.19 / 81.98	81.67 / 0.19 / 81.98	81.67 / 0.19 / 81.98	81.68/ 0.20/81.99	78.76 / 0.21 / 78.96	78.71 / 0.23 / 78.95	78.76 / 0.21 / 78.96	78.71 / 0.23 / 78.95
		K=1	85.15 / 0.20 / 85.46	85.15 / 0.20 / 85.46	85.15 / 0.20 / 85.46	85.16/ 0.20/85.48	81.83 / 0.56 / 82.36	81.74 / 0.57 / 82.27	81.83 / 0.56 / 82.36	81.79 / 0.57 / 82.32
		K=2	85.25 / 0.41 / 85.88	85.25 / 0.41 / 85.87	85.25 / 0.41 / 85.88	85.27 / 0.42 / 85.91	82.55 / 0.27 / 83.00	82.42 / 0.28 / 82.88	82.55 / 0.27 / 83.00	82.56 / 0.28 / 83.02
		K=3	84.76 / 0.19 / 84.94	84.76 / 0.19 / 84.94	84.76 / 0.19 / 84.94	84.76/ 0.19/84.94	82.64 / 0.26 / 82.94	82.46 / 0.26 / 82.77	82.64 / 0.26 / 82.94	82.75 / 0.27 / 83.07
		K=4	85.27 / 0.28 / 85.62	85.27 / 0.28 / 85.62	85.27 / 0.28 / 85.62	85.27/ 0.28/85.63	**82.86** / 0.30 / 83.18	**82.77** / 0.32 / 83.16	**82.86** / 0.30 / 83.18	**82.85** / 0.27 / 83.14
		K=5	**85.49** / 0.34 / 85.83	**85.49** / 0.34 / 85.82	**85.49** / 0.34 / 85.83	**85.50**/ 0.34/85.83	82.17 / 0.66 / 82.65	82.01 / 0.70 / 82.53	82.17 / 0.66 / 82.65	82.23 / 0.64 / 82.68
			Co-hyponym v/s Random				Hypernym v/s Random			
ROOT9	GLOVE+CLIP-Img	K=0	93.73 / 0.20 / 93.94	93.78 / 0.18 / 93.97	93.73 / 0.20 / 93.94	93.87/ 0.14/94.01	87.38 / 0.83 / 88.14	87.42 / 0.81 / 88.16	87.38 / 0.83 / 88.14	87.48 / 0.80 / 88.18
		K=1	94.51 / 0.36 / 94.89	94.54 / 0.36 / 94.91	94.51 / 0.36 / 94.89	94.59/ 0.34/94.93	91.23 / 0.18 / 91.49	91.20 / 0.17 / 91.40	91.23 / 0.18 / 91.43	91.18 / 0.17 / 91.38
		K=2	**95.18** / 0.15 / 95.42	**95.20** / 0.13 / 95.41	**95.20** / 0.15 / 95.42	**95.24** / 0.11 / 95.41	92.82 / 0.22 / 93.08	92.76 / 0.22 / 93.03	92.82 / 0.22 / 93.08	92.75 / 0.22 / 93.02
		K=3	94.89 / 0.33 / 95.15	94.90 / 0.33 / 95.17	94.89 / 0.33 / 95.15	94.92/ 0.34/95.20	**93.31** / 0.19 / 93.57	**93.24** / 0.18 / 93.50	**93.31** / 0.19 / 93.57	**93.26** / 0.19 / 93.53
		K=4	94.24 / 0.36 / 94.75	94.29 / 0.38 / 94.80	94.24 / 0.36 / 94.75	94.35/ 0.39/94.87	92.89 / 0.30 / 93.25	92.88 / 0.29 / 93.24	92.89 / 0.30 / 93.25	92.88 / 0.29 / 93.23
		K=5	94.51 / 0.15 / 94.75	94.54 / 0.16 / 94.80	94.51 / 0.15 / 94.75	94.59/ 0.18/94.87	92.72 / 0.74 / 93.41	92.70 / 0.70 / 93.36	92.72 / 0.74 / 93.41	92.70 / 0.70 / 93.36

Table 5: Results for IxRUMEN and IxROOT9 for hybrid fusion based on the CentralNet fusion network.

unimodal visual results tend to show that visual information provides some useful information that can be used in a multimodal decision process. This is confirmed by the values of the error analysis proposed in Table 6, that evidence the complementarity between both modalities. Indeed, although there is a great deal of learning instances that are correctly classified by the textual modality and misclassified by the visual modality, the opposite is also true for a non negligible set of learning instances. It is also interesting to note that the patch-based strategy for the visual modality with CLIP embeddings shows the worst complementary performance with the textual modality, as the highest percentage of correct guesses made by the visual modality, that are misclassified by the textual modality is evidenced for $K = 0$. This is in line with the results discussed in the above paragraph about the impact of the patch-based strategy on the visual modality, when encoded with multimodal representations.

As a consequence of these preliminary results, we propose to combine the textual unimodal models based on GloVE with the visual unimodal models based on CLIP to apply early fusion and hybrid fusion techniques.

5.2 Multimodal Models

Results for multimodal models are given in Table 4 for the early fusion with the attention fusion network (AFN), and in Table 5 for the hybrid fusion with the CentralNet fusion network (CFN). The first conclusion to be drawn is that both fusion techniques allow to achieve higher results in terms of F1 score for all experimental

		Synonym v/s Random				Hypernym v/s Random			
	Patch Size	VC,TC	VC,TI	VI,TC	VI,TI	VC,TC	VC,TI	VI,TC	VI,TI
RUMEN	K=0	1259	221	289	157	1110	185	245	166
	K=1	1295	157	326	148	1149	109	254	192
	K=2	1297	156	338	135	1151	134	248	163
	K=3	1302	142	329	153	1153	136	262	155
	K=4	1289	158	344	135	1153	138	250	165
	K=5	1320	155	314	137	1159	139	239	169
		Co-hyponym v/s Random				Hypernym v/s Random			
ROOT9	K=0	639	33	47	24	462	53	57	35
	K=1	667	11	30	35	509	18	44	36
	K=2	668	19	30	26	511	25	47	24
	K=3	663	22	32	26	504	22	59	22
	K=4	639	13	58	33	510	20	52	25
	K=5	663	17	36	27	513	23	45	26

Table 6: Error Analysis on IxRUMEN and IxROOT9 for the GloVE textual model and the CLIP visual model. VC (resp. TC) stands for visual (resp. textual) correct predictions, and VI (resp. TI) stands for visual (resp. textual) incorrect guesses.

configurations when compared to the best unimodal model (here textual modality encoded with GloVE). In particular, 1.71 point improvement is obtained for synonymy on IxRUMEN with AFN, 1.44 point for hypernymy on IxRUMEN with AFN, 1.16 point for co-hyponymy on IxROOT9 with CFN, and 0.60 point for hypernymy on IxROOT9 with CFN. Interestingly, the early fusion provides better results for IxRUMEN, while the hybrid fusion presents stronger results for IxROOT9. For instance, for IxRUMEN, the difference between AFN and CFN ranges from 1.16 points for hypernymy to 1.12 points for synonymy, while for IxROOT9, the difference is smaller to the advantage of CFN, with values ranging from 0.09

point for co-hyponymy to 0.32 point for hypernymy. Note that these results are obtained for similar values of K, to the exception of synonymy for IxRUMEN, where the AFN provides best results for $K = 1$, while the CFN achieves highest performance for $K = 5$. Nevertheless, when closely looking at the results, it is clear that the difference between the AFN and the CFN is marginal for IxROOT9, while it is clearly in favour of the AFN for IxRUMEN. Note that the best configuration of CFN is presented here, which freezes the unimodal results. Indeed, lower experimental results were obtained for the all-trainable architecture proposed in [58].

The patch-based strategy is also beneficial for the multimodal models. Indeed, all result values for $K > 0$ steadily exceed the figures obtained for $K = 0$, in all experimental setups, i.e. for all datasets, lexico-semantic relations and fusion techniques. Note that within this paper, we propose to use the same number of K for both modalities. This can be an obstacle for further improvements as it has been shown in section 5.1 that the textual modality and the visual modality behave differently with respect to patch size[6]. The other particularity of the multimodal models is that they tend to produce higher results for less number of patches for the symmetric relations (i.e. synonymy and co-hyponymy). As such, they rely on less information for each modality, but take advantage of the diversity of the representations. In particular, for synonymy in IxRUMEN, best results are obtained for $K = 1$, while the best unimodal model provides highest results for $K = 5$. For co-hyponymy in IxROOT9, highest results are evidenced for $K = 2$, while the best unimodal model relies on $K = 5$ to achieve the maximum performance. Note that this situation does not hold for asymmetric relations (i.e. hypernymy), as similar values of K are needed to reach highest results.

5.3 Qualitative Analysis

In order to better understand the quantitative results, we provide a qualitative analysis between unimodal models and multimodal models, by looking at specific successful and unsuccessful cases. In Table 7, we first show learning examples that have been correctly identified by the multimodal fusion model and misclassified by both the unimodal models, and where a specific lexico-semantic relation holds (i.e. synonymy, co-hyponymy, hypernymy). These examples show that when the set of images of both words are closely related in terms of visual content and the respective words non polysemous, positive decisions can be made by the multimodal architecture. Note that in this study, we refer to the multimodal model with the attention fusion network, and both GloVE and CLIP-Image unimodal models.

Word pair	Dataset	Relation	Unimodal
(labour, toil)	RUMEN	synonym	random
(rub, snag)	RUMEN	synonym	random
(walk, paseo)	RUMEN	hypernym	random
(rebate, discount)	RUMEN	hypernym	random
(bowl, tumbler)	ROOT9	co-hyponym	random
(falcon, crow)	ROOT9	hypernym	random

Table 7: Pairs identified by multimodal fusion but misclassified by unimodal models, where a semantic relation holds.

[6]This line of work remains for future work.

However, this situation is relatively rare in the IxRUMEN dataset, while more frequent in the IxROOT9 dataset. But, visual information can also help in disambiguating wrong guesses from the unimodal models for the random relation. Indeed, unimodal models show a high rate of false positives that the multimodal model is capable of handling, as shown in Table 8. Note that most of the result improvements by the multimodal architecture come from this situation. In this case, while the unimodal models infer a lexico-semantic relation, the multimodal model correctly classifies the learning input as random. This situation stands if visual contents are clearly unrelated and word pairs non polysemous.

Word pair	Dataset	Relation	Unimodal
(trafficker, trading)	RUMEN	random	synonym
(esr, keyboard)	RUMEN	random	synonym
(jog, trot)	RUMEN	random	hypernym
(spouse, mate)	RUMEN	random	hypernym
(bettle, ant)	ROOT9	random	hypernym
(flute, saxophone)	ROOT9	random	hypernym

Table 8: Pairs identified by multimodal fusion but misclassified by unimodal models, where a random relation holds.

Finally, some good predictions made by the multimodal model are difficult to interpret based on the associated multimodal information, as illustrated in Table 9. This clearly shows that the proposed model is still subject to deep improvements, especially when the word pair is polysemous and when the quality of the visual information is not controlled, or difficult to retrieve in the case of abstract words.

Word pair	Dataset	Relation	Unimodal
(chalk, trash)	RUMEN	synonym	random
(chest, bureau)	RUMEN	synonym	random
(slob, pig)	RUMEN	synonym	random
(cardholder, clef)	RUMEN	hypernym	random
(bite, snack)	RUMEN	hypernym	random
(bang, fringe)	RUMEN	hypernym	random

Table 9: Pairs identified by multimodal fusion but misclassified by unimodal models, where a semantic relation holds, but interpretation is hard.

6 CONCLUSION

In this paper, we propose the first attempt to deal with the identification of lexico-semantic relations based on multimodal information, thus following the semiotic textology linguistic theory. For that purpose, we build the IxROOT9 and IxRUMEN datasets, the multimodal versions of the gold standards RUMEN and ROOT9, as well as we gather the necessary visual information to apply the augmentation data paradigm. To take advantage of the multimodal information, we implement two fusion techniques (early and hybrid), and extend the patch-based strategy to visual information. Experimental results demonstrate that introducing visual information can reliably supplement the missing semantics of textual information. In particular, improvements are observed that range from 1.71 point to 0.60 point in terms of F1 score depending on the dataset and the lexico-semantic relation. Nevertheless, improvements are still limited, essentially due to the automatic selection process of images, which cannot guarantee the quality of the visual information, as well as the inability to connect abstract words to reliable visual information.

REFERENCES

[1] Houssam Akhmouch, Gaël Dias, and Jose G. Moreno. 2021. Understanding Feature Focus in Multitask Settings for Lexico-semantic Relation Identification. In *Findings of the Association for Computational Linguistics (ACL/IJCNLP)*. ACL, Thailand, 2762–2772.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *IEEE International Conference on Computer Vision (ICCV)*. 2425–2433.

[3] Mohammed Attia, Suraj Maharjan, Younes Samih, Laura Kallmeyer, and Thamar Solorio. 2016. CogALex-V Shared Task: GHHH - Detecting Semantic Relations via Word Embeddings. In *Workshop on Cognitive Aspects of the Lexicon*. 86–91.

[4] Georgios Balikas, Gaël Dias, Rumen Moraliyski, Houssam Akhmouch, and Massih-Reza Amini. 2019. Learning Lexical-Semantic Relations Using Intuitive Cognitive Links. In *41st European Conference on Information Retrieval (ECIR)*. 3–18.

[5] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* 41, 2 (2018), 423–443.

[6] Nesrine Bannour, Gaël Dias, Youssef Chahir, and Houssam Akhmouch. 2020. Patch-Based Identification of Lexical Semantic Relations. In *42nd European Conference on Information Retrieval (ECIR)*. 126–140.

[7] Marco Baroni, Raffaella Bernardi, Ngoc-Quynh Do, and Chung chieh Shan. 2012. Entailment Above the Word Level in Distributional Semantics. In *13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. 23–32.

[8] Zied Bouraoui, Jose Camacho-Collados, and Steven Schockaert. 2020. Inducing relational knowledge from BERT. In *34th AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 34. 7456–7463.

[9] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: UNiversal Image-TExt Representation Learning. In *European Conference on Computer Vision (ECCV)*. 104–120.

[10] Fabio Crestani, Martin Braschler, Jacques Savoy, Andreas Rauber, Henning Müller, David E Losada, G Heinatz Bürki, Linda Cappellato, and Nicola Ferro. 2019. Experimental IR Meets Multilinguality, Multimodality, and Interaction. In *10th International Conference of the CLEF Association (CLEF)*, Vol. 11696. Springer.

[11] Sebastien Delecraz, Leonor Becerra-Bonache, Benoît Favre, Alexis Nasr, and Frédéric Béchet. 2021. Multimodal Machine Learning for Natural Language Processing: Disambiguating Prepositional Phrase Attachments with Images. *Neural Processing Letters* 53, 5 (2021), 3095–3121.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 248–255.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. 4171–4186.

[14] Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. 2017. Learning to Paraphrase for Question Answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 886–897.

[15] R. Fu, J. Guo, B. Qin, W. Che, H. Wang, and T. Liu. 2015. Learning Semantic Hierarchies: A Continuous Vector Space Approach. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23, 3 (2015), 461–471.

[16] Mahak Gambhir and Vishal Gupta. 2017. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review* 47, 1 (2017), 1–66.

[17] Benito García-Valero. 2020. The Legacy of János S. Petőfi. Text Linguistics, Literary Theory and Semiotics. *Journal of Literary Semantics* 49, 1 (2020), 61–64.

[18] Goran Glavas and Ivan Vulic. 2019. Generalized Tuning of Distributional Word Vectors for Monolingual and Cross-Lingual Lexical Entailment. In *57th Conference of the Association for Computational Linguistics (ACL)*. 4824–4830.

[19] Marti Hearst. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. In *14th Conference on Computational Linguistics (COLING)*. 539–545.

[20] Judith Holler and Stephen C Levinson. 2019. Multimodal language processing in human communication. *Trends in Cognitive Sciences* 23, 8 (2019), 639–652.

[21] MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. A comprehensive survey of deep learning for image captioning. *Comput. Surveys* 51, 6 (2019), 1–36.

[22] Glyn W Humphreys and Jie Sui. 2016. Attentional control and the self: the Self-Attention Network (SAN). *Cognitive neuroscience* 7, 1-4 (2016), 5–17.

[23] Sergio Jimenez, Fabio A. Gonzalez, Alexander Gelbukh, and George Duenas. 2019. Word2set: WordNet-Based Word Representation Rivaling Neural Word Embedding for Lexical Similarity and Sentiment Analysis. *IEEE Computational Intelligence Magazine* 14, 2 (2019), 41–53.

[24] Aishwarya Kamath, Jonas Pfeiffer, Edoardo Maria Ponti, Goran Glavaš, and Ivan Vulić. 2019. Specializing Distributional Vectors of All Words for Lexical Entailment. In *4th Workshop on Representation Learning for NLP (RepL4NLP)*. 72–83.

[25] Neha Kathuria, Kanika Mittal, and Anusha Chhabra. 2017. A Comprehensive Survey on Query Expansion Techniques, their Issues and Challenges. *International Journal of Computer Applications* 168, 12 (2017).

[26] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations (ICLR)*, Yoshua Bengio and Yann LeCun (Eds.).

[27] Zornitsa Kozareva and Eduard Hovy. 2010. A Semi-supervised Method to Learn and Construct Taxonomies Using the Web. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1110–1118.

[28] Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. Combining Language and Vision with a Multimodal Skip-gram Model. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL/HTL)*. 153–163.

[29] Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan. 2015. Do supervised distributional methods really learn lexical inference relations?. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. 970–976.

[30] Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan. 2015. Do Supervised Distributional Methods Really Learn Lexical Inference Relations?. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. 970–976.

[31] Lizi Liao, Yunshan Ma, Xiangnan He, Richang Hong, and Tat-seng Chua. 2018. Knowledge-aware multimodal dialogue systems. In *26th ACM International Conference on Multimedia (MM)*. 801–809.

[32] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. *CoRR* abs/1405.0312 (2014).

[33] Fenglin Liu, Xian Wu, Shen Ge, Xuancheng Ren, Wei Fan, Xu Sun, and Yuexian Zou. 2021. DiMBERT: Learning Vision-Language Grounded Representations with Disentangled Multimodal-Attention. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 16, 1 (2021), 1–19.

[34] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems (NeurIPS)* 32 (2019).

[35] Catherine Marechal, Dariusz Mikolajewski, Krzysztof Tyburek, Piotr Prokopowicz, Lamine Bougueroua, Corinne Ancourt, and Katarzyna Wegrzyn-Wolska. 2019. Survey on AI-Based Multimodal Methods for Emotion Detection. *High-performance modelling and simulation for big data applications* 11400 (2019), 307–324.

[36] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography* 3, 4 (1 January 1990), 235–244.

[37] Kim Anh Nguyen, Maximilian Köper, Sabine Schulte im Walde, and Ngoc Thang Vu. 2017. Hierarchical Embeddings for Hypernymy Detection and Directionality. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 233–243.

[38] Kim Anh Nguyen, Sabine Schulte im Walde, and Ngoc Thang Vu. 2017. Distinguishing Antonyms and Synonyms in a Pattern-based Neural Network. In *15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. 76–85.

[39] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. *The PageRank Citation Ranking: Bringing Order to the Web.* Technical Report 1999-66. Stanford InfoLab.

[40] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Conference on Empirical Methods on Natural Language Processing (EMNLP)*. 1532–1543.

[41] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Mazumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Multi-level multiple attentions for contextual multimodal sentiment analysis. In *IEEE International Conference on Data Mining (ICDM)*. 1033–1038.

[42] James Pustejovsky, Eben Holderness, Jingxuan Tu, Parker Glenn, Kyeongmin Rim, Kelley Lynch, and Richard Brutti. 2021. Designing Multimodal Datasets for NLP Challenges. *CoRR* abs/2105.05999 (2021). arXiv:2105.05999

[43] Syed Arbaaz Qureshi, Sriparna Saha, Mohammed Hasanuzzaman, and Gaël Dias. 2019. Multitask Representation Learning for Multimodal Estimation of Depression Level. *IEEE Intelligent Systems* 34, 5 (2019), 45–52.

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. *CoRR* abs/2103.00020 (2021).

[46] Marek Rei, Daniela Gerz, and Ivan Vulić. 2018. Scoring Lexical Entailment with a Supervised Directional Similarity Network. In *56th Annual Meeting of the Association for Computational Linguistics (ACL)*. 638–643.

[47] Stephen Roller, Katrin Erk, and Gemma Boleda. 2014. Inclusive yet Selective: Supervised Distributional Hypernymy Detection. In *25th International Conference on Computational Linguistics (COLING)*. 1025–1036.

[48] Stephen Roller, Douwe Kiela, and Maximilian Nickel. 2018. Hearst Patterns Revisited: Automatic Hypernym Detection from Large Text Corpora. In *56th Annual Meeting of the Association for Computational Linguistics (ACL)*. 358–363.
[49] Enrico Santus, Alessandro Lenci, Tin-Shing Chiu, Qin Lu, and Chu-Ren Huang. 2016. Nine Features in a Random Forest to Learn Taxonomical Semantic Relations. In *10th International Conference on Language Resources and Evaluation (LREC)*. 4557–4564.
[50] Enrico Santus, Vered Shwartz, and Dominik Schlechtweg. 2017. Hypernyms under Siege: Linguistically-motivated Artillery for Hypernymy Detection. In *15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. 65–75.
[51] Xindi Shang, Zehuan Yuan, Anran Wang, and Changhu Wang. 2021. Multimodal Video Summarization via Time-Aware Transformers. In *29th ACM International Conference on Multimedia (MM)*. 1756–1765.
[52] Vered Shwartz, Yoav Goldberg, and Ido Dagan. 2016. Improving Hypernymy Detection with an Integrated Path-based and Distributional Method. In *54th Annual Meeting of the Association for Computational Linguistics (ACL)*. 2389–2398.
[53] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
[54] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *3rd International Conference on Learning Representations (ICLR)*.
[55] Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2004. Learning Syntactic Patterns for Automatic Hypernym Discovery. In *17th International Conference on Neural Information Processing Systems (NeurIPS)*. 1297–1304.
[56] Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. 2017. A survey of multimodal sentiment analysis. *Image and Vision Computing* 65 (2017), 3–14.
[57] Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. 2018. CentralNet: A Multilayer Approach for Multimodal Fusion. In *European Conference on Computer Vision (ECCV)*. 575–589.
[58] Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. 2018. Centralnet: a multilayer approach for multimodal fusion. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. 0–0.
[59] Tu Vu and Vered Shwartz. 2018. Integrating Multiplicative Features into Supervised Distributional Methods for Lexical Entailment. In *7th Joint Conference on Lexical and Computational Semantics (*SEM)*. 160–166.
[60] Ivan Vulić and Nikola Mrkšić. 2018. Specialising Word Vectors for Lexical Entailment. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. 1134–1145.
[61] Ivan Vulic and Nikola Mrksic. 2018. Specialising Word Vectors for Lexical Entailment. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. 1134–1145.
[62] Ivan Vulic, Nikola Mrksic, Roi Reichart, Diarmuid Ó Séaghdha, Steve J. Young, and Anna Korhonen. 2017. Morph-fitting: Fine-Tuning Word Vector Spaces with Simple Language-Specific Rules. In *55th Annual Meeting of the Association for Computational Linguistics (ACL)*. 56–68.
[63] Ekaterina Vylomova, Laura Rimell, Trevor Cohn, and Timothy Baldwin. 2016. Take and Took, Gaggle and Goose, Book and Read: Evaluating the Utility of Vector Differences for Lexical Relation Learning. In *54th Annual Meeting of the Association for Computational Linguistics (ACL)*. 1671–1682.
[64] Chengyu Wang and Xiaofeng He. 2020. BiRRE: Learning Bidirectional Residual Relation Embeddings for Supervised Hypernymy Detection. In *58th Annual Meeting of the Association for Computational Linguistics (ACL)*. 3630–3640.
[65] Julie Weeds, Daoud Clarke, Jeremy Reffin, David J. Weir, and Bill Keller. 2014. Learning to Distinguish Hypernyms and Co-Hyponyms. In *5th International Conference on Computational Linguistics (COLING)*. 2249–2259.
[66] Xiaoshi Wu, Hadar Averbuch-Elor, Jin Sun, and Noah Snavely. 2021. Towers of babel: Combining images, language, and 3D geometry for learning multimodal vision. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. 428–437.
[67] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. 2020. Contrastive Learning of Medical Visual Representations from Paired Images and Text. *CoRR* abs/2010.00747 (2020). arXiv:2010.00747
[68] Changmeng Zheng, Junhao Feng, Ze Fu, Yi Cai, Qing Li, and Tao Wang. 2021. Multimodal Relation Extraction with Efficient Graph Alignment. In *29th ACM International Conference on Multimedia (MM)*. 5298–5306.

A EXPERIMENTAL SETUPS

Within this first attempt to combine visual and textual information for the identification of lexico-semantic relations, only the highest ranked image for each word has been taken into account, the two less ranked images being withdrawn from the process[7]. In order to train each model, a random split of 90% training and 10% validation instances is built from the original training set. Note that at validation, lexical split is not performed. All models are run 5 times for patch size ranging from 0 (no augmentation) to 5 (5 extra words form the patch), to produce average performance results with corresponding standard deviation values and maximum performance scores. All models are trained with a batch size of 32 for up to 200 epochs with early stopping (patience = 10). Adam optimizer [26] is used with a learning rate = 10^{-5}, β_1 = 0.9 and β_2 = 0.999. With respect to encodings, GloVE embeddings are of size 300, CLIP embeddings are 512-dimensional vectors and VGG19 encodings are of size 4096.

[7] The use of this extra information remains for future work.