Multimedia Semantic Integrity Assessment Using Joint Embedding Of Images And Text

Ayush Jaiswal*
USC Information Sciences Institute
Marina del Rey, CA, USA
ajaiswal@isi.edu

Ekraam Sabir*
USC Information Sciences Institute
Marina del Rey, CA, USA
esabir@isi.edu

Wael AbdAlmageed
USC Information Sciences Institute
Marina del Rey, CA, USA
wamageed@isi.edu

Premkumar Natarajan
USC Information Sciences Institute
Marina del Rey, CA, USA
pnataraj@isi.edu

ABSTRACT

Real-world multimedia data is often composed of multiple modalities such as an image or a video with associated text (e.g., captions, user comments, etc.) and metadata. Such multimodal data *packages* are prone to manipulations, where a subset of these modalities can be altered to misrepresent or repurpose data packages, with possible malicious intent. It is therefore important to develop methods to assess or verify the integrity of these multimedia packages. Using computer vision and natural language processing methods to directly compare the image (or video) and the associated caption to verify the integrity of a media package is only possible for a limited set of objects and scenes. In this paper we present a novel deep-learning-based approach that uses a reference set of multimedia packages to assess the semantic integrity of multimedia packages containing images and captions. We construct a joint embedding of images and captions with deep multimodal representation learning on the reference dataset in a framework that also provides image-caption consistency scores (ICCSs). The integrity of query media packages is assessed as the *inlierness* of the query ICCSs with respect to the reference dataset. We present the MultimodAl Information Manipulation dataset (MAIM), a new dataset of media packages from Flickr, which we are making available to the research community. We use both the newly created dataset as well as Flickr30K and MS COCO datasets to quantitatively evaluate our proposed approach. The reference dataset does not contain *unmanipulated* versions of tampered query packages. Our method is able to achieve F_1 scores of 0.75, 0.89 and 0.94 on MAIM, Flickr30K and MS COCO, respectively, for detecting semantically incoherent media packages.

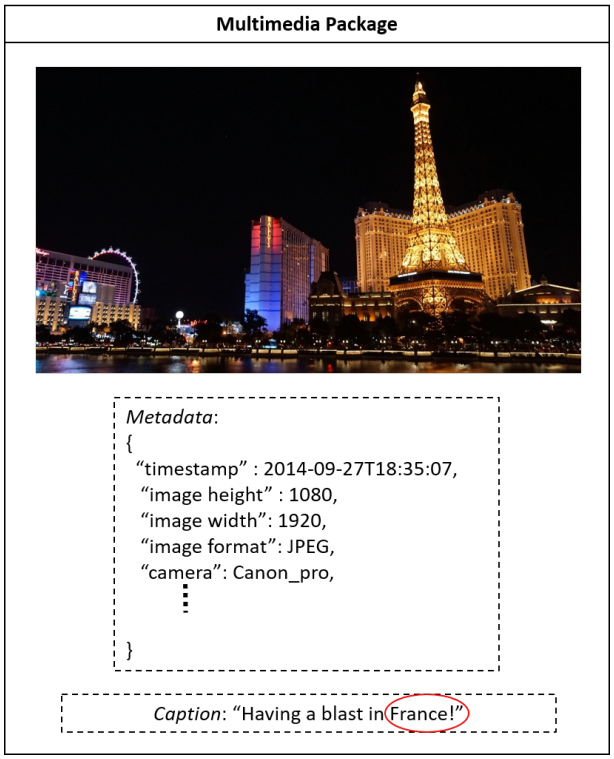

Figure 1: Multimodal information manipulation example. This is a photograph of the Eiffel Tower in Las Vegas, but the caption says France.

*Ayush Jaiswal and Ekraam Sabir contributed equally to the work in this paper.

MM '17, October 23–27, 2017, Mountain View, CA, USA

ACM ISBN 978-1-4503-4906-2/17/10…$15.00
https://doi.org/10.1145/3123266.3123385

CCS CONCEPTS

• **Computing methodologies** → *image representations*; *computer vision*; *natural language processing*; *unsupervised learning*; *neural networks*;

KEYWORDS

semantic integrity assemessment; multimedia data; multimodal semantic integrity

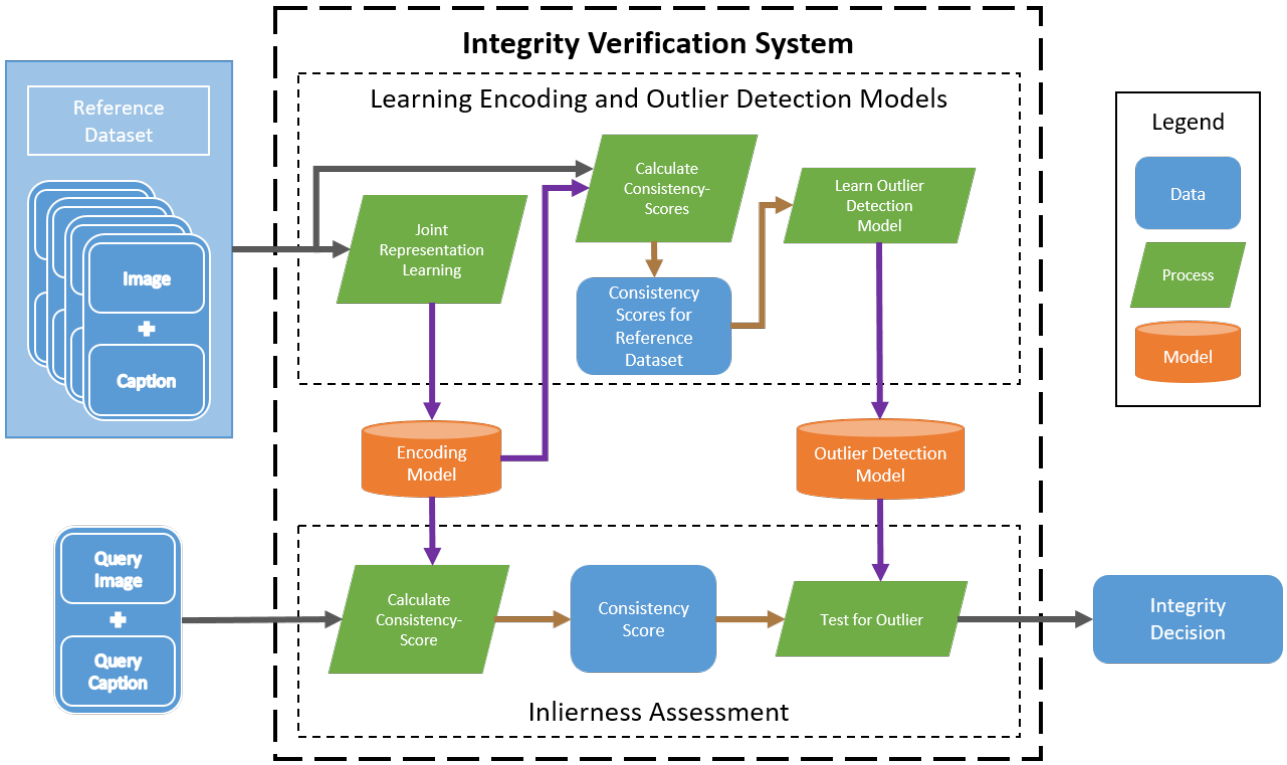

Figure 2: Package Integrity Assessment System

1 INTRODUCTION

In real life, data often presents itself with multiple modalities, where information about an entity or an event is incompletely captured by each modality separately. For example, a caption associated with the image of a person might provide information such as the name of the person and the location where the picture was taken, while other metadata might provide the date and time at which the image was taken. Independent existence of each modality makes multimedia data packages vulnerable to tampering, where the data in a subset of modalities of a multimedia package can be modified to misrepresent or repurpose the multimedia package. Such tampering, with possible malicious intent, can be misleading, if not dangerous. The location information, for example, in the aforementioned caption could be modified without an easy way to detect such tampering. Fig. 1 demonstrates an example of information manipulation where a photograph of the Paris casino in Las Vegas, Nevada (which includes a half-scale replica of the Eiffel Tower) is repurposed as a photograph of the real Eiffel Tower in Paris, France. Nevertheless, if the image has visual cues, such as a landmark, a person *familiar* with the location can easily detect such a manipulation. However, this is a challenging multimedia analysis task, especially with the subtlety of data manipulation, the absence of clear visual cues (e.g., Eiffel Tower) and the proliferation of multimedia content from mobile devices and digital cameras.

Verification of the integrity of information contained in any kind of data requires the existence of some form of prior knowledge. In the previous example, this knowledge is represented by a person's *familiarity* with the location. Human beings use their knowledge, learned over time, or external sources such as encyclopedias, as knowledge bases (KBs). Motivated by this important observation, multimedia analysis algorithms could also take advantage of a KB to automatically assess the integrity of multimedia packages. A KB can be either implicit (such as a trained scene understanding and/or recognition model) or explicit (such as a database of known facts). In this paper we explore the use of a reference dataset (RD) of multimedia packages to assess the integrity of query packages. The RD is assumed to *not* include other copies of the query image. Otherwise, existing image retrieval methods would suffice to verify the multimedia package integrity.

While information manipulation detection is a broad problem, in this paper we focus on verifying the *semantic* integrity of multimedia packages. We define multimedia semantic integrity as the semantic consistency of information across all the modalities of a multimedia package.

We present a novel framework to solve a limited version of the multimedia information integrity assessment problem, where we consider each data package to contain only one image and an accompanying caption. Data packages in the reference dataset are used to train deep multimodal representation learning models (DMRLMs). The learned DMRLMs are then used to assess the integrity of query packages by calculating image-caption consistency scores (ICCSs) and employing outlier detection models (ODMs) to find their *inlierness* with respect to the RD. We evaluate the proposed method on two publicly available datasets—Flickr30K [22] and MS

COCO [9], as well as on the MultimodAl Information Manipulation (MAIM) dataset that we created from image and caption pairs downloaded from Flickr, which we make publicly available.

To the best of our knowledge, ours is the first work to formally define the larger problem and provide an approach to solve it. Our work is significantly different from past work on robust hashing and watermarking [1, 17, 19, 21] as those methods focus on the prevention of information manipulation while ours focuses on detection at a semantic level. The remainder of this paper is organized as follows. Section 2 reviews related work. In Section 3 we describe the proposed method for assessing the semantic integrity of multimedia packages. In Section 4 we discuss existing public datasets as well as the new MAIM dataset. Experimental results and analysis are presented in Section 5. Finally, in Section 6 we conclude the paper and introduce directions for future research.

2 RELATED WORK

Content integrity of multimedia data has been tackled in the past from the perspective of manipulation prevention using watermarking, authentication and hashing [1, 17, 19, 21]. Most of such work is aimed at detecting tampered data, especially images, and approaches to recover the original data. This group of work focuses on the integrity of data with one modality. Our work is different in that it focuses on assessing the integrity across modalities of a multimodal package at a semantic level. The larger problem that we have defined above assumes that images are not tampered with but might be repurposed, thus creating fake data packages with inconsistent information.

Our methods in this paper are based on recent work in deep multimodal represenation learning [14, 18] and semantic retrieval of images and captions involving image-caption ranking [4, 7, 8, 10, 20]. Deep representation learning performs very well at learning highly non-linear latent representations of data when large volumes of data are available. Autoencoders [5] are a popular framework for unsupervised representation learning. Ngiam et al. [14] showed how MAEs can be used to learn joint representations of data with multiple modalities. Vukotić et al. [18] developed the BiDNN model that learns cross-modal mappings and joint representations of multimodal data.

Semantic retrieval of images from captions and vice versa has gained traction in recent years. Several methods have been developed that map images and captions to a common feature space so that their similarity, such as the cosine similarity, can be used to rank the affinity of image-caption pairs and return the top-K candidates [4, 7, 8, 10, 20]. Hodosh et al. [7] use Kernel Canonical Correlation Analysis [2] to map image features and caption features to a common induced space. Wetson et al. [20] provide a method to simultaneously learn linear mappings from image and caption features to a common space with the objective of learning to associate images with correct captions. Frome et al. [4] present a deep visual-semantic embedding (DeViSE) model that learns to map image features to the space of caption features by optimizing a loss function that maximizes the cosine similarity of image-caption pairs while minimizing that of images and randomly picked text. The neural language model of Kiros et al. [8] is based on a similar framework but learns to map caption features to the space of image features instead. In experiments, they showed that their model's performance is much better at the task of image-caption ranking compared to DeViSE.

Kiros et al. [8] also compared their model to the multimodal recurrent neural network model of Mao et al. [12] that automatically generates captions for images. This class of methods, based on caption generation, does not explicitly give a score for image-caption affinity and relies on perplexity when used for image-caption ranking or retrieval.

While previous work offers a way to rank image-caption pairs based on a measure of similarity, it does not provide a way to check the consistency of information between images and associated captions with respect to a reference dataset. Our work in this paper provides this novel contribution towards the larger goal of assessing the integrity of multimodal data packages.

3 SEMANTIC INTEGRITY ASSESSMENT

One approach to information integrity assessment of a data object is to compare it against an existing knowledge-base (KB), with the assumption that such a KB exists. This KB can be explicit (such as a database of facts) or implicit (such as a learned statistical inference model). We use the observation that human beings verify the information integrity of pieces of data using world knowledge learned over time and external sources, such as an encyclopedia, to develop machine learning models that mimic world knowledge, and then use these models to assess the integrity of query data packages.

In order to verify the integrity of a query multimedia package that contains an image and an associated caption, we assume the existence of a reference set of similar media packages. This set, which we call the reference dataset (RD), serves as the KB to compare query packages against to measure their integrity. More specifically, we train an outlier detection model (ODM) on image-caption consistency scores (ICCSs) from packages in RD and use it to calculate the *inlierness* of query packages. We employ deep multimodal representation learning models (DMRLMs) for jointly encoding images and corresponding captions, inspired by their success as reflected in recent literature, and use them to calculate ICCSs (depending on the DMRLM used). Fig. 2 explains our complete integrity assessment system.

In this work we use a multimodal autoencoder (MAE) [14], a bidirectional (symmetrical) deep neural network (BiDNN) [18] or the unified visual semantic neural language model (VSM) [8] as the embedding model. VGG19 [16] image features are given as inputs to all these models, along with either average word2vec [13] embeddings (MAE and BiDNN) or one-hot encodings of words in captions (VSM). The ODMs that we work with are the one-class support vector machine (OCSVM) [15] and isolation forest (iForest) [11]. We discuss the aforementioned DMRLMs, with their associated ICCSs, and ODMs in detail in the following subsections.

3.1 Deep Multimodal Representation Learning

3.1.1 Multimodal Autoencoder. An autoencoder is a neural network that learns to reconstruct its input [5]. Autoencoders are typically used to learn low-dimensional representations of data. The network architecture is designed such that the input goes

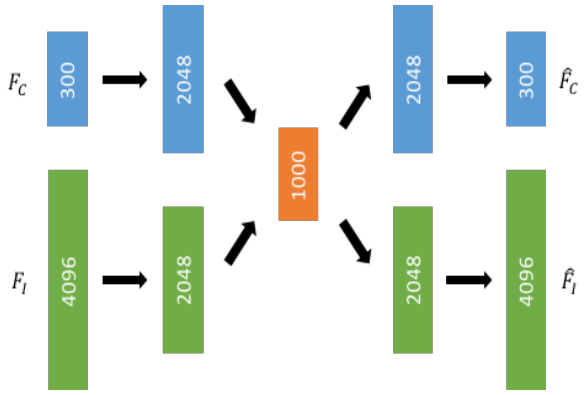

Figure 3: Our Multimodal Autoencoder Architecture. F_I and F_C are image and caption features respectively, while $\widehat{F}_I$ and $\widehat{F}_C$ are their reconstructed versions.

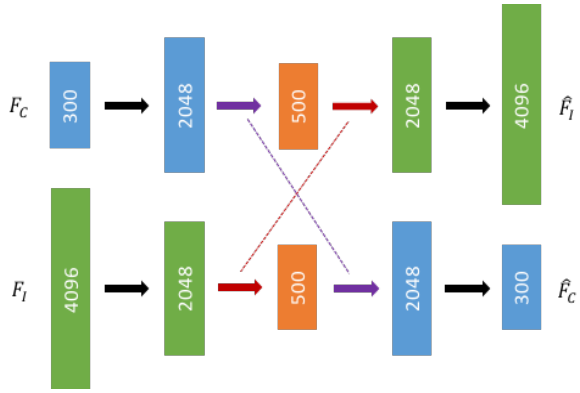

Figure 4: Our BiDNN Architecture. F_I and F_C are image and caption features respectively, while $\widehat{F}_I$ and $\widehat{F}_C$ are their reconstructed versions. Colored arrows with dotted connections reflect weight tying.

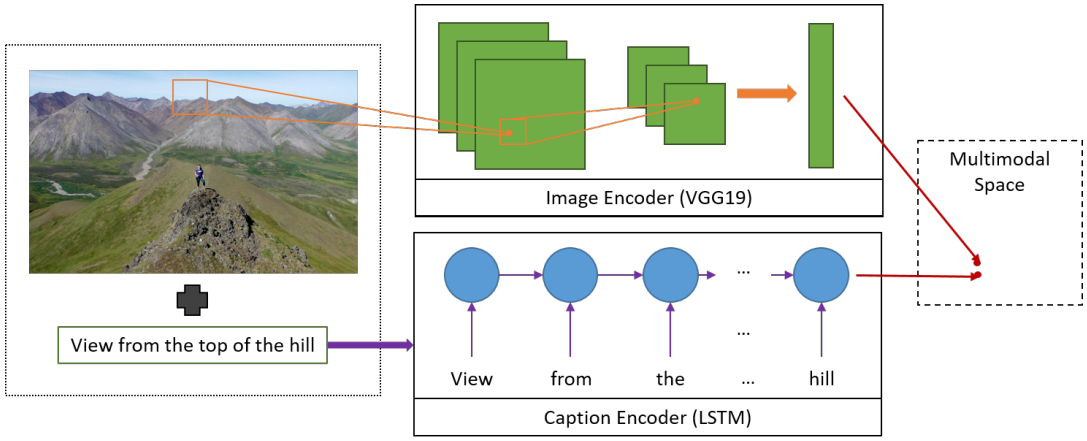

Figure 5: The VSM Architecture of Kiros et al. [8]

through a series of layers with decreasing dimensionality to produce an encoding, which is then transformed through layers of increasing dimensionality to finally reconstruct the input. Ngiam et al. [14] showed how an autoencoder network can be used to learn representations over multiple modalities. We train an MAE on the image-caption pairs in RD to learn their shared representation. Fig. 3 shows our MAE architecture, inspired by the bimodal deep autoencoder of Ngiam et al. [14]. The image and caption features are passed through a series of unimodal layers before combining them in the shared representation layer. The decoder module of the MAE is a mirror image of its encoder. For MAE, we use reconstruction loss as the ICCS.

3.1.2 Bidirectional (Symmetrical) Deep Neural Network. A BiDNN is composed of two unimodal autoencoders with tied weights for the middle representation layers [18]. The network is trained to simultaneously reconstruct each modality from the other, learning cross-modal mappings as well as a joint representation. Fig. 4 shows our BiDNN architecture and illustrates the tied weights for a better understanding. Our formulation of the joint representation is the same as Vukotić et al. [18], i.e., the concatenation of the activations of the two representation layers. We used the BiDNN package made available by Vukotić et al. [18][1] to implement our model. Reconstruction loss also serves as the ICCS in the case of BiDNN.

3.1.3 Unified Visual Semantic Neural Language Model. Kiros et al. [8] introduced the unified visual semantic neural language model (VSM) that learns representations of captions in the embedding space of images, where image embeddings are first calculated using a deep neural network such as VGG19 [16]. The VSM is trained to optimize a contrastive loss, which aims to maximize the cosine similarity between the representation of an image and the learned encoding of its caption while minimizing that between the image and captions not related to it. Fig. 5 shows the structure of the VSM. The network uses long short-term memory (LSTM) units [6] to encode variable-length captions. We used the VSM package made

[1] https://github.com/v-v/BiDNN

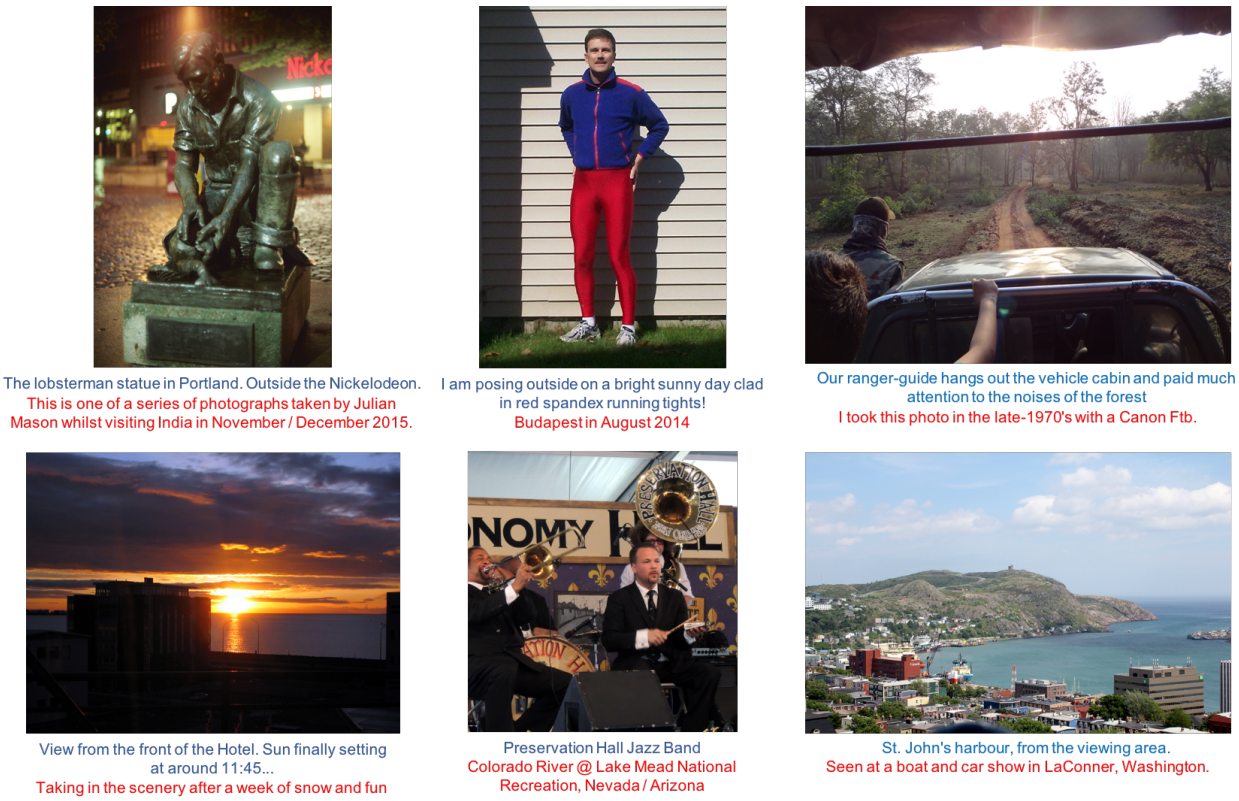

Figure 6: Image-Caption Data package examples from MAIM dataset. The blue captions are the original ones that came with the image while the red ones are their manipulated versions.

available by Kiros et al. [8] [2] and trained one model on each RD. Cosine similarity becomes the natural choice of ICCS in the case of VSM.

3.2 Outlier Detection

3.2.1 One-Class Support Vector Machine. The OCSVM is an unsupervised outlier detection model trained only on positive examples [15]. It learns a decision function based on the distribution of the training data in its original or kernel space to classify the complete training dataset as the positive class, and everything else in the high-dimensional space as the negative class. This model is then used to predict whether a new data point is an inlier or an outlier with respect to the training data. This formulation of OCSVM fits very well with our approach of assessing the semantic integrity of a data package with respect to an RD (by using an OCSVM trained on the RD).

3.2.2 Isolation Forest. An isolation forest (iForest) is a collection of decision trees that isolate a data point through recursive partitioning of random subsets of its features [11]. It works by first randomly selecting a feature of a data point and then finding a random split-value between its minimum and maximum values. This is then repeated recursively on the new splits. The recursive partitioning of a tree stops when a node contains only the provided data point. Under this setting, the average number of splits required (across all trees in the forest) to isolate a point gives an indication of its outlierness. The smaller the number, the higher the confidence that the point is an outlier; it is easier to isolate outliers as they lie in relatively low-density regions with respect to inliers (RD).

4 DATA

We provide a quantitative evaluation of the performance of our method on three datasets: Flickr30K [22], MS COCO [9] and a dataset that we created from images, captions and other metadata downloaded from Flickr (MAIM). While Flickr30K and MS COCO datasets contain objective captions which describe the contents of images, MAIM contains subjective captions, which do not necessarily do so and sometimes contain related information that might not be obvious from the image. Fig. 6 shows some examples from the MAIM dataset.

We use the training, validation and testing subsets of Flickr30K and MS COCO as made available by Kiros et al. [8][3], which makes sure that there is no overlap of images among the subsets. This is necessary because each image in these datasets has five captions (giving five image-caption pairs). There are 158, 915 and 423, 915 image-caption pairs in Flickr30K and MS COCO respectively, in total. Our dataset (MAIM) has 239, 968 image-caption pairs with

[2]https://github.com/ryankiros/visual-semantic-embedding

[3]https://github.com/ryankiros/visual-semantic-embedding

Table 1: Evaluation Results on Flickr30K

		Deep Multimodal Representation Learning Model					
		MAE		BiDNN		VSM	
		F_1-tampered	F_1-clean	F_1-tampered	F_1-clean	F_1-tampered	F_1-clean
ODM	One-class SVM	0.48	0.50	0.47	0.67	**0.89**	**0.88**
	Isolation Forest	0.49	0.50	0.63	0.62	**0.81**	**0.7**

Table 2: Evaluation Results on MS COCO

		Deep Multimodal Representation Learning Model					
		MAE		BiDNN		VSM	
		F_1-tampered	F_1-clean	F_1-tampered	F_1-clean	F_1-tampered	F_1-clean
ODM	One-class SVM	0.53	0.46	0.68	0.55	**0.94**	**0.94**
	Isolation Forest	0.5	0.48	0.76	0.77	**0.94**	**0.94**

Table 3: Evaluation Results on MAIM

		Deep Multimodal Representation Learning Model					
		MAE		BiDNN		VSM	
		F_1-tampered	F_1-clean	F_1-tampered	F_1-clean	F_1-tampered	F_1-clean
ODM	One-class SVM	0.49	0.49	0.46	0.5	**0.75**	**0.76**
	Isolation Forest	0.56	0.42	0.52	0.52	**0.75**	**0.77**

exactly one caption for each unique image. We randomly split MAIM into training, validation and testing subsets and treat the training subset of each dataset as the RD in our framework. We replace the captions of half of the validation and testing images with captions of other images to create manipulated image-caption pairs for evaluation.

MAIM also has metadata for each package but we do not use them in our experiments in this work. This metadata includes location where the image was taken, time and date when the image was taken and information associated with the device used to capture the image.

5 ANALYSIS

The inlier/outlier decisions of the ODMs in our system serve as the prediction of semantic information manipulation in query packages. We use F_1 scores as our evaluation metrics. Tables 1, 2 and 3 summarize the results of our experiments on all combinations of DMRLMs and ODMs that we use in this work, on Flickr30K, MS COCO and MAIM respectively. We treat tampered packages as the positive class when calculating F_1-tampered and as the negative class for F_1-clean. The F_1-tampered and F_1-clean scores are coupled, i.e., every pair is from the same trained model.

We see that VSM consistently performs better than MAE and BiDNN in all our experiments on both metrics, with MAE consistently performing the worst. This gives us some key insights into the working of these DMRLMs. Even though MAE can compress multimodal data with low reconstruction error, it does not learn semantic associations between images and captions very well. The BiDNN model is trained to learn cross-modal mappings between images and captions, which forces it to learn those semantic associations. This explains why it works better than MAE at this task. The VSM model is trained to map captions to the representation space of images. The learning objective explicitly requires it to learn semantic relationships between the two modalities so that it can map captions consistent with an image close to it while inconsistent ones are mapped far from it. This makes VSM the strongest of the three models.

We also see that the F_1 scores of VSM on MS COCO are significantly better than those on the other datasets. This is expected and explained by the process through which the captions in the dataset were collected. Chen et al. [3] used Amazon's Mechanical Turk[4] to gather objective captions with strong guidelines for their quality and content. The numbers are higher simply due to the better quality of captions and their objective content. This indicates that our method is better suited for objective captions.

[4]https://www.mturk.com/mturk/welcome

6 CONCLUSIONS AND FUTURE WORK

Real-world multimedia data is often multimodal, consisting of images, videos, captions and other metadata. While multiple modalities present additional sources of information, it also makes such data packages vulnerable to tampering, where a subset of modalities might be manipulated, with possible malicious intent. In this paper we formally defined this problem and provided a method to solve a limited version of it (where each package has an image and a caption) as a first step towards the larger goal. Our method combines deep multimodal representation learning with outlier detection methods to assess whether a caption is consistent with the image in its package. We introduced the MultimodAl Information Manipulation dataset (MAIM) that we created for the larger problem, containing images, captions and various metadata, which we make available to the research community.[5] We presented a quantitative evaluation of our method on Flickr30K and MS COCO datasets, containing objective captions, and on the MAIM dataset, containing subjective captions. Our method was able to achieve F_1 scores of 0.75, 0.89 and 0.94 on MAIM, Flickr30K and MS COCO, respectively, for detecting semantically incoherent media packages.

In our work we used the general formulation of MAE and BiDNN, providing these models VGG19 image features and aggregated word2vec caption features as inputs. It is possible that an end-to-end model with raw images and captions as inputs and a combination of convolution and recurrent layers might perform better. Similarly, training the image encoder of VSM jointly with the caption encoder might further boost its performance. We intend to explore these issues in future work. Our future work will also incorporate metadata and assess the integrity of entire packages. It is easy to see that our framework can be extended to include more modalities such as audio and video. We leave this to future work. This is in accordance with the larger goal of semantic multimodal information integrity assessment.

ACKNOWLEDGMENTS

This work is based on research sponsored by the Defense Advanced Research Projects Agency under agreement number FA8750-16-2-0204. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Defense Advanced Research Projects Agency or the U.S. Government.

REFERENCES

[1] Sufyan Ababneh, Rashid Ansari, and Ashfaq Khokhar. 2008. Scalable multimedia-content integrity verification with robust hashing. In *Electro/Information Technology, 2008. EIT 2008. IEEE International Conference on*. IEEE, 263–266. http://ieeexplore.ieee.org/abstract/document/4554310/

[2] Francis R. Bach and Michael I. Jordan. 2002. Kernel independent component analysis. *Journal of machine learning research* 3, Jul (2002), 1–48. http://www.jmlr.org/papers/v3/bach02a

[3] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr DollÃąr, and C. Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325* (2015). https://arxiv.org/abs/1504.00325

[4] Andrea Frome, Greg S. Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, and others. 2013. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*. 2121–2129. http://papers.nips.cc/paper/5204-devise-a-deep-visual-semantic-embedding-model

[5] Geoffrey E. Hinton and Ruslan R. Salakhutdinov. 2006. Reducing the dimensionality of data with neural networks. *science* 313, 5786 (2006), 504–507. http://science.sciencemag.org/content/313/5786/504.short

[6] Sepp Hochreiter and JÃijrgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780. http://www.mitpressjournals.org/doi/abs/10.1162/neco.1997.9.8.1735

[7] Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research* 47 (2013), 853–899. http://www.jair.org/papers/paper3994.html

[8] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539* (2014). https://arxiv.org/abs/1411.2539

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr DollÃąr, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*. Springer, 740–755. http://link.springer.com/chapter/10.1007/978-3-319-10602-1_48

[10] Xiao Lin and Devi Parikh. 2016. Leveraging visual question answering for image-caption ranking. In *European Conference on Computer Vision*. Springer, 261–277. http://link.springer.com/chapter/10.1007/978-3-319-46475-6_17

[11] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2008. Isolation forest. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*. IEEE, 413–422. http://ieeexplore.ieee.org/abstract/document/4781136/

[12] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L. Yuille. 2014. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090* (2014). https://arxiv.org/abs/1410.1090

[13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[14] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. 2011. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*. 689–696. http://machinelearning.wustl.edu/mlpapers/paper_files/ICML2011Ngiam_399.pdf

[15] Bernhard SchÃűlkopf, Robert C. Williamson, Alexander J. Smola, John Shawe-Taylor, John C. Platt, and others. 1999. Support vector method for novelty detection.. In *NIPS*, Vol. 12. 582–588. https://papers.nips.cc/paper/1723-support-vector-method-for-novelty-detection.pdf

[16] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014). https://arxiv.org/abs/1409.1556

[17] Rui Sun and Wenjun Zeng. 2014. Secure and robust image hashing via compressive sensing. *Multimedia tools and applications* 70, 3 (2014), 1651–1665. http://link.springer.com/article/10.1007/s11042-012-1188-8

[18] Vedran VukotiÄĞ, Christian Raymond, and Guillaume Gravier. 2016. Bidirectional joint representation learning with symmetrical deep neural networks for multimodal and crossmodal applications. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. ACM, 343–346. http://dl.acm.org/citation.cfm?id=2912064

[19] Xiaofeng Wang, Kemu Pang, Xiaorui Zhou, Yang Zhou, Lu Li, and Jianru Xue. 2015. A visual model-based perceptual image hash for content authentication. *IEEE Transactions on Information Forensics and Security* 10, 7 (2015), 1336–1349. http://ieeexplore.ieee.org/abstract/document/7050251/

[20] Jason Weston, Samy Bengio, and Nicolas Usunier. 2010. Large scale image annotation: learning to rank with joint word-image embeddings. *Machine learning* 81, 1 (2010), 21–35. http://www.springerlink.com/index/Y277128518468756.pdf

[21] Cai-Ping Yan, Chi-Man Pun, and Xiao-Chen Yuan. 2016. Multi-scale image hashing using adaptive local feature extraction for robust tampering detection. *Signal Processing* 121 (2016), 1–16. http://www.sciencedirect.com/science/article/pii/S0165168415003709

[22] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2 (2014), 67–78. https://www.transacl.org/ojs/index.php/tacl/article/view/229

[5] The code and data (MAIM) used in this work are available on request through email.