CP4101 B.Comp. Dissertation

Final Year Project Report

Multi-Modal Entity Resolution

By

Chen Xihao

Department of Computer Science

School of Computing

National University of Singapore

AY 2022/2023

CP4101 B.Comp. Dissertation

Final Year Project Report

Multi-Modal Entity Resolution

By

Chen Xihao

Department of Computer Science

School of Computing

National University of Singapore

AY 2022/2023

Project No: H064740
Advisor: Professor Tan Kian-Lee
Evaluator: Assistant Professor Djordje Jevdjic
Deliverables:
Report: 1 Volume

Abstract

Entity resolution is a fundamental problem in data processing, querying and information retrieval. The problem is amplified in the age of big data, where we are constantly seeking out more efficient and precise methods to retrieve relevant documents. In this project, I explore and present a resolution scheme for multi-modality linkage using novel deep-learning methods in representation learning.

Many existing entity resolution methods, whether in single or multi-modal settings, assume that records come in the form of some low-level representations, e.g., the text being represented as an index vector from a bag-of-words model, images being represented as a correlogram or a histogram. In real life, however, data are usually stored in their raw form, e.g., text as string, and image as RGBA pixel values. As a result, in production pipelines, these existing solutions will either store extra information alongside the raw data or transform raw data into low-level representations on the fly. Hence, this project aims to develop a system that can learn latent representations from raw data, and then perform entity resolution on the learned representations.

We present a new architecture for multi-modal entity resolution - the TransforMMER. The model differentiates from previous solutions by making use of transformers to take in data from various modalities in their raw form and encode them into a common representation space. The model is trained using a joint framework as a generative-adversarial network to ensure modality invariance in the representation produced. Using empirical results on a new dataset, we show that the TransforMMER model outperforms various existing solutions for multi-modal entity resolution.

This report details my literature study, proposed model, experiments, results, discussions and future work.

Subject Descriptors:

H.3.3: Information Search and Retrieval

I.2.7: Natural Language Processing

I.4.0: Image Processing and Computer Vision (General)

Keywords:

Entity resolution, information retrieval, multi-modality, deep learning.

Implementation Software and Hardware:

Python, PyTorch, Sci-kit Learn, SoC Compute Cluster, NVIDIA Tesla A100 GPU, NVIDIA Tesla V100 GPU

Acknowledgement

I would like to express my sincerest gratitude to my advisor, Professor Tan Kian Lee for his mentorship and guidance guidance throughout this journey. I could not have asked for a more supportive and dedicated advisor, and I am truly blessed to have had the privilege of working under his guidance.

Table of Contents

1 Introduction

1.1 Problem and Motivation

In 2021, an estimated staggering 2.1 quintillion bytes of data is generated each day. With the vast availability of data, there is an evermore increasing reliance on systems and algorithms to store and operate on this big data. However, such data come in various forms, and one prominent problem is the discrepancy between how the same idea is represented by different pieces of data. This disparity exists naturally in real life. For example, "SoC", "Com" and "School of Computing" are various ways to refer to the Computing faculty at NUS. Similarly, I can be referred to as "Chen Xihao", "Chen, X.", "Xihao Chen" or simply "Xihao", depending on the scenario. As humans, we learn the names of surrounding objects and individuals and we can understand that an entity may have different names. However, how can a computer system learn this?

1.1.1 Entity Resolution

Entity resolution (ER), or record linkage, refers to the process of resolving whether multiple records, potentially from different sources, refer to the same real-world entity. Formally, a record is a piece of information that represents some real-world object.

ER has been extensively studied with various methods and implementations. In recent years, deep learning (DL) algorithms have been used to extract inherent features from records. Such DL-based solutions have advanced ER systems drastically with improved accuracy in linkage. However, despite the advances in DL methods, there remain problems, and above all, the poor ability to generalise. This inability occurs across various areas, including datasets, domains, and modalities.

1.1.2 Multi-Modality

In addition to the examples above, records can exist in forms other than text (K. Wang et al. 2016). In general, a record can be a piece of text, an image, a video, an audio recording or other forms of multimedia. In multi-modal problems, we aim to perform resolution with and using records from one or more of these forms.

Furthermore, multi-modal entity resolution problems exist in numerous variants. For example, one variant of such a problem is to match a text record to other text records, but each text record is linked to and substantiated with an image record. In this scenario, solutions measure the similarity between text records, with the help of the similarity between image records.

Modality gaps are critical problems in multi-modal problems. Many solutions have attempted to reduce these gaps and biases to achieve modality invariance. Our work incorporates existing methods in modality invariance to achieve better performance. This will be further explained in subsection 3.2

1.2 Contributions

In this project, we focus on incorporating novel representation learning methods for entity resolution to allow for feature extraction from raw inputs. Such a model sets us apart from most existing works, which use pre-encoded low-level features, such as colour histograms to represent images, or index vectors after passing text through a bag-of-words model. These existing methods have one major drawback - they require either 1) the data to be converted into low-level features on the fly before being fed into the model, or 2) store these low-level features alongside the raw data. In either case, there is the need for extra computation or storage overhead, which can be undesirable in production pipelines. Furthermore, due

to this pre-processing, a certain amount of information is lost, which may affect the performance of the model. For example, the positions of words in a sentence are crucial in understanding the meaning of a sentence, but this information is lost in a bag-of-words model.

We propose a novel solution, TransforMMER, incorporating state-of-the-art transformers to learn representations from raw multi-modal data, capturing the maximum amount of information possible.

We show empirically the effectiveness of TransforMMER for the problem of image-text entity resolution and the significant improvements over prior solutions. Our model achieves higher mean average precision (MAP) in all four types of queries: image-to-image, text-to-text, image-to-text and text-to-image. Notably, TransforMMER achieves stellar results for text-to-text queries (nearly 100% MAP), image-to-text queries (around 60% MAP and nearly 100% MAP@5) and text-to-image queries (around 60% MAP and more than 70% MAP).

Furthermore, we evaluate TransforMMER's capabilities in learning with minimal samples. We show that it achieves near-optimal results for some queries, despite being trained on few (less than 200) samples per concept. This is a testament to the model's ability to perform continual learning.

2 Literature Review

2.1 Entity Resolution

2.1.1 Problem Definition

Entity resolution is a family problem concerned with matching records of the same entities (Papadakis et al. 2021). Formally, given two datasets $A = \{a_1, \ldots, a_m\}$ and $B = \{b_1, \ldots, b_n\}$, we want to find $M \subseteq A \times B$ where $(a, b) \in M$ is a pair of records that represent the same entity. There are various variants of the ER problem, for example:

- comparing records within the same dataset, to find duplicates,
- comparing records from different sources, and
- comparing records of different modalities, and various others.

In this project, we explore multi-modal entity resolution (MMER) with text and images. To simplify the problem, we assert that the images and text records share exactly the same common set of concepts. If time permits, we will explore and experiment across sets of concepts.

2.2 Deep Learning for Entity Resolution

2.2.1 Overview of Deep Learning

Deep learning is a technique in machine learning that learns the inherent patterns that exist in data. With advances in DL models, they have been extensively studied for ER use cases.

DL models mostly exist in the form of artificial neural networks, or NN for short. A neural network approximates some objective function $f(X)$ using parameters P. Given

some model architecture D, the approximated function is in the form of $\hat{f}_D(X \mid \theta)$, where θ are the parameters (weights) to the neural network to be learned from training data.

Neural networks typically consist of the following components: an input layer, hidden layer(s) and an output layer. In DL, NNs have more than one hidden layer, which allows the model to capture some hierarchical representation of the input.

Since the project focuses on image and text ER, we have first reviewed the DL methods for each modality individually.

Text. In text processing tasks, recurrent neural networks (RNNs) were pioneers in dealing with inputs with dynamic lengths, such as those in sentences. However, to address the problem of vanishing gradient due to large numbers of layers of the models, Long Short-Term Memory (LSTM) models were developed (Hochreiter et al. 1997). These models tap on an attention mechanism to control the passing of information, boosting those considered "useful" and shrinking others.

In recent years, a new mechanism in self-attention has emerged, which forms the foundation of the transformer family of models (Vaswani et al. 2017). Self-attention is a mechanism that allows each position in a sequence to relate to other positions in the same sequence. Subsequently, models such as ELMo (Peters et al. 2018) and BERT (Devlin et al. 2019) have emerged to extract features from text sequences.

Since its development, BERT has been the go-to choice for deep learning methods for text-based ER (Li et al. 2020; Chatterjee et al. 2022).

Image. In image processing tasks, such as computer vision, the traditional method of convolutional neural networks (CNNs) can be used to detect edges and shapes that exist in an image. Notably, in recent years, CNN models such as ResNeXt50 (Xie et al. 2017)

and InceptionResNetV2 (Szegedy et al. 2016), which incorporate residual layers on top of convolutions, have shown significant performance in classification tasks.

With the advances in transformer models, self-attention has also been ported to image processing, creating models such as the Image Transformer (Parmar et al. 2018) and Vision Transformer (Dosovitskiy et al. 2021).

2.2.2 Deep Learning in Multi-Modal Entity Resolution

Deep learning methods in the multi-modal context can be categorised roughly into two types:

1. real-valued representation learning, and
2. binary representation learning.

In general, these methods use some form of representation learning to extract features.

Given the context of ER, records of the same entity, despite having different modalities, are expected to share similar features, and hence common representations that are similar to each other. Hence, most models share a similar general pipeline shown in Figure 1.

Leveraging this common representation, models in (1) extract features of the records into a vector representation. They are then projected into the same feature space, where similarity can be measured directly through a distance-neighbouring algorithm. A major drawback in real-valued representations is in the slow execution of matching, where distance neighbouring algorithms take at least linear time to compute.

Models in (2) build on this idea by speeding up the matching process. Speed-up is commonly achieved through hashing. Extracted features are hashed into some bucket, where each bucket represents some concept. Records that hash to the same bucket are inferred to belong to the same entity.

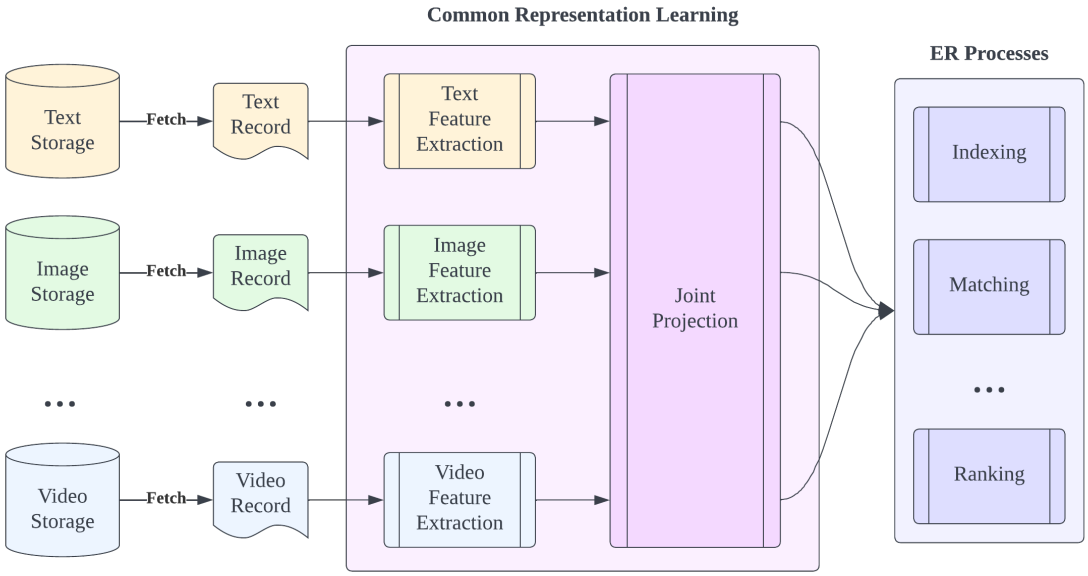
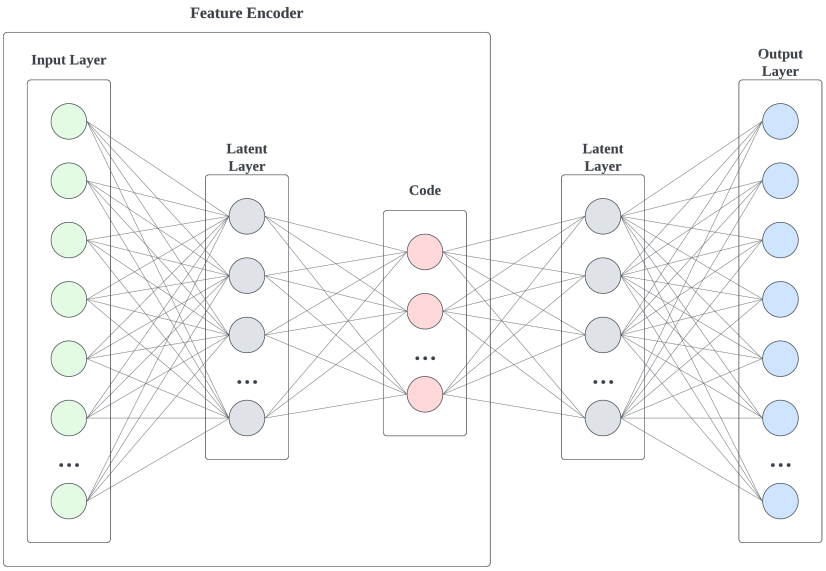


Figure 1: General architecture of multi-modal entity resolution solutions.

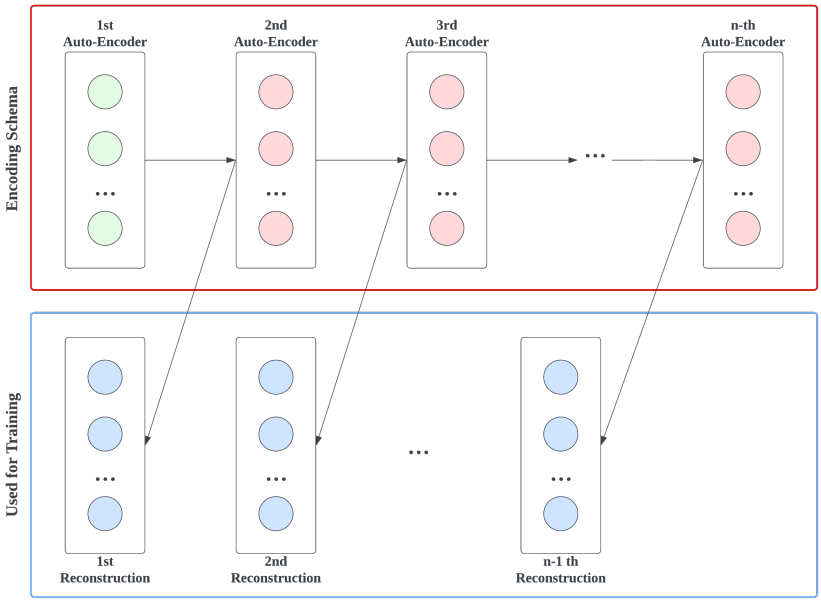
Models mostly differ in how the common representation is extracted and how similarity is calculated. Here, we review notable deep learning models for multi-modal entity resolution:

Multi-Modal Stacked Auto-Encoders. W. Wang et al. 2014 observed that prior research such as Cross View Hash (CVH) (Kumar et al. 2011), Multi-modal Latent Binary Embedding (MLBE) (Zhen et al. 2012) and Latent Semantic Cross-Modal Ranking (LSCMR) (Lu et al. 2013) require large amounts of prior knowledge on the training dataset, increasing the time taken to preprocess. This includes manual inspection of the dataset, performing manual transformations to the data, etc. This was necessary to build encoders to extract meaningful features.

Hence, W. Wang et al. 2014 proposed a new way to create such feature extractors, through the use of stacked auto-encoders (SAE).



(a) Structure of Auto-Encoders.



(b) Structure of Stacked Auto-Encoders for a single modality.

Figure 2: SAE Architecture

In each auto-encoder, the input is passed through a multi-layer neural network, consisting of mainly two segments - an encoder and a decoder, as seen in Figure 2a. The network takes in an input, encodes it, and then reconstructs the original input from the encoded features. During inference, the encoder part of the network is extracted to be used as the feature encoder.

The autoencoders are concatenated into a pipeline to create the stacked autoencoders. The output of the i^{th} autoencoder is used as the input to the $(i+1)^{\text{th}}$ autoencoder. In Figure 2b, the top row (red box) represents the encoding schema of an SAE. This is similar to a deep neural network. The bottom row (blue box) is used to reconstruct the input during the training stage, where reconstruction loss is calculated. The architecture attempts to address the issue of vanishing gradients during backward propagation, where the losses are not propagated up the layers effectively.

An SAE is initialised for each modality. Each SAE is first trained independently on data from its modality to learn the different concepts of its own modality. Then, to ensure they extract features into a common representation space, the SAEs from different modalities are trained jointly to capture similarities between documents of the same concept but of different modalities. This is done through a joint loss function:

$$\mathcal{L}\left(X^0, Y^0\right) = \alpha \mathcal{L}_r^I\left(X^0, X^{2h}\right) + \beta \mathcal{L}_r^T\left(Y^0, Y^{2h}\right) + \mathcal{L}_d\left(X^h, Y^h\right) + \xi(\theta), \tag{1}$$

where X^0 and Y^0 are the matrices for images and text records, X^{2h} and Y^{2h} are the matrices for the reconstructed images and records by the MSAE model, $\mathcal{L}_r^I$ and $\mathcal{L}_r^T$ are reconstruction loss errors from image and text SAEs, and $\mathcal{L}_d$ is a distance function between the latent features of images and text records. A regularisation function ξ is imposed on the parameters θ of the network.

Each SAE directly produces a vector representing the input data. Then, each vector is

indexed by a distance algorithm, and the index is stored locally. The MSAE model supports both binary representations and real-valued representations, where the prior achieves fast queries at a loss of accuracy, whereas the latter does the opposite.

MSAE proves effective in training a model from scratch. Importantly, it requires little to no prior knowledge of the dataset, as the auto-encoders can learn the underlying semantic values themselves. MSAE, as compared to a previous method of CVH, showed improvements in the MAP metric in all image-to-image, text-to-text, image-to-text and text-to-image queries. Furthermore, given the modular implementation, the model may be easily extended to other modalities.

However, the MSAE is not without its drawbacks. Importantly, auto-encoders themselves require inputs in 1-dimensional vectors, which are not the natural ways in which records are stored. Hence, the model requires some preprocessing steps to transform these records into 1D vectors, such as histograms or correlograms for images, which leads to a loss of information. While low-level features in the form of 1D vectors are provided by the datasets, they are not the ideal representation of native records.

Multi-Modal Semantic Auto-Encoder. Wu et al. 2019 developed a model similar to W. Wang et al. 2014's MSAE. Notably, Wu et al. 2019's model makes use of a conditional principle label space transformation (CPLST) process, which dynamically adjusts the feature space during training, dropping "redundant" features and reducing dimensionality (Z. Lin et al. 2014).[1]

Such a method allowed for a simpler model with only two auto-encoders, one for images and one for text, as compared to the stacked auto-encoders used by W. Wang et al. 2014's MSAE. Wu et al. 2019's model achieved only marginally better MAP performance on the

[1] Similar in idea to principle component analysis in statistical regression problems.

WIKI dataset. However, the simplicity of the model, which leads to shorter training times, is appreciated.

Despite the ability for auto-encoders to capture both semantic features and original information of the record, we see that the common need for preprocessing of data still exists, as discussed previously.

Multi-Modal Adversarial Network. Instead of auto-encoders, generative adversarial networks (GANs) have become popular methods for entity resolution. For example, the Entity resolution Generative Adversarial Network (ErGAN) (Shao et al. 2020) was introduced for text-based entity resolution, and the multi-modal Adversarial Network (MAN) (Hu et al. 2019) was proposed, which extended GANs to multi-modal entity resolution for more than just images and text records.

The MAN model consists of a series of generator modules, one for each modality, and one discriminator module. Each generator module encodes its input into a common representation space, where the outputs are used to compute intra- and inter-modality distance. Each generator attempts to "fool" the discriminator while the discriminator attempts to classify the representation to its corresponding modality.[2]

An emphasis in the discriminator to classify a representation to its modality trains the generators to produce features that are modality-invariant, allowing records of the same concept, but from different modalities, to be mapped to similar neighbourhoods in the common representation subspace.

However, this inadvertently introduces a potential for cheating. As the discriminator only tries to classify the modality of a representation, but not its concept, the generators can potentially, learn to output some fixed representation, regardless of the input.

[2]Note: classify the representation's modality, not its concept.

To resolve this, Hu et al. 2019 propose a novel multi-modal discriminant analysis (MDA) module, which forces the generated representations from different concepts to be distinctively discriminative, i.e., they are well separated in common representation space. Inherently, this mechanism prevents the set of generators from "cheating", allowing them to produce representations that are discriminative and modality-invariant.

Multi-Modal Entity Resolution with DeepMatcher. DeepMatcher provided a robust approach to text-based entity resolution (Mudgal et al. 2018). The study revealed advantages in DL approaches for non-structured, text-based or dirty records. It also uncovered DL's weaknesses in producing meaningful matches with structured records. Wilke et al. 2021 extend the DeepMatcher model to a multi-modal context to support the matching of images through the use of existing, well-performing CNN models such as ResNet50. The paper points out the need for an attribute summariser to handle texts of different lengths and suggested the use of RNNs for such a purpose. Similarly, if multiple images are used to represent one record, such a summariser is also required. However, the work assumes that each record consists of a single image and hence did not provide any implementation of an image summariser.

Notably, the model was implemented on a dataset of product images and descriptions, where both descriptive text and product images can be noisy with unnecessary information. The work further extended Mudgal et al. 2018's claims of DL being robust to noisy data in a multi-modal context. However, given the advances and maturity in sequence-to-sequence text models, such as BERT, which could produce text embeddings independent of input lengths, it was unclear why Wilke et al. 2021 chose to use the traditional word embeddings, which were input-length dependent and required the use of an RNN summariser.

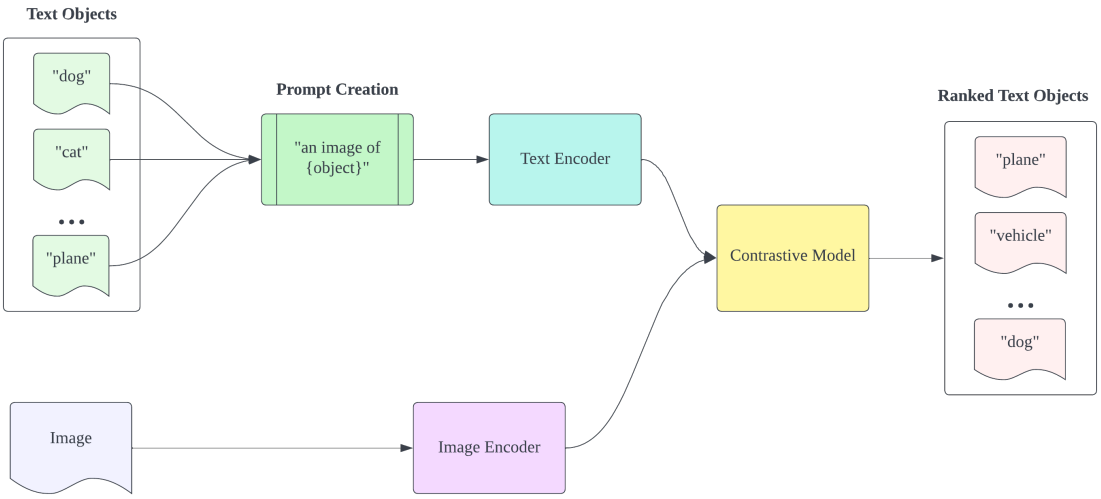


Figure 3: Architecture of the CLIP model (Radford et al. 2021).

2.3 Other Useful Deep Learning Frameworks

Another piece of work that may prove to be useful is Radford et al. 2021's CLIP model, which learns visual concepts from natural language supervision, shown in Figure 3.

CLIP is capable of classifying an image into one of the various text prompts and is built on various works in zero-shot knowledge transfer, natural language supervision and multi-modal learning.

Instead of directly optimising on classification benchmarks, CLIP can be instructed via natural language prompts in the form of "a photo of a furry brown cat", allowing the model to perform a greater variety of classification tasks. Natural language supervision with text obtained from internet crawls allowed for the model to be trained without human supervision. That is, the training process did not require manual annotation and pairing of images and text records.

The model consists mainly of a text encoder, an image encoder (using either a ResNet

or a Vision Transformer model) and a projection component for ranking and classification. Similar to DL solutions with real-valued representations in multi-modal ER, the encoders in CLIP output a common representation, where similarity is computed. The encoders are trained jointly with pairs of images and text records using the *Information Noise-Contrastive Estimation* (InfoNCE) loss function (Oord et al. 2019).

While Radford et al. 2021's CLIP model is used in classification and cannot be directly used for ER, it showed that such a joint loss function allows encoders to not only produce representations with semantic features but also learn the inherent relationships between concepts. For example, the image encoder may have only been exposed to images of dogs with non-white fur during training, but it is capable of encoding and drawing similarities to the phrase "dog with white fur". Such capabilities may prove to be useful in entity resolution tasks.

3 Methodology

3.1 Use of Transformers

Recently, transformers have emerged as competitive alternatives to traditional recurrent neural networks (RNN), for text modelling, and convolutional neural networks (CNN), for image modelling. Crucially, the use of transformers allows us to use the documents in their raw format as inputs.

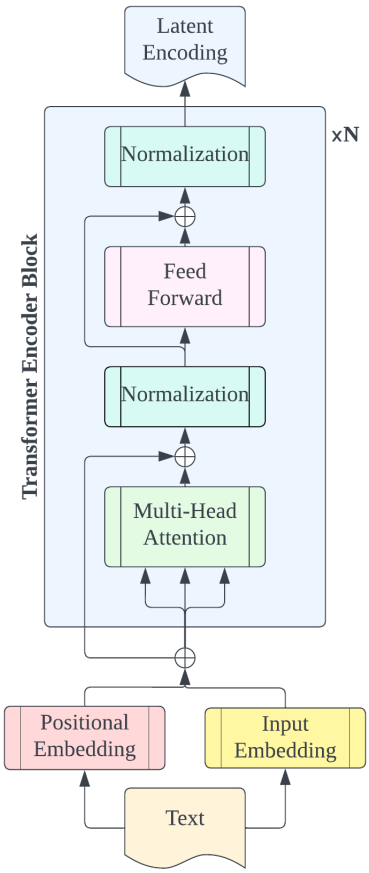
Transformers use a self-attention mechanism to process input sequences and make predictions. Transformers contain two modules, the encoder and the decoder. Here, we make use of the encoder module to produce latent representations of the input document.

Self-attention. The self-attention mechanism allows the transformer network to focus on different parts of the input sequence at each time step, rather than processing the input in the order of the sequence. In language modelling tasks, when processing a token of the sequence, the transformer pays "more attention" to certain tokens (appearing on either side of the current token) that it deems relevant. Suppose we have the sentence "A girl wearing a red shirt petting a cat.". When the transformer is processing the token "red", it pays more attention to the token "shirt" than "cat", since the word "red" is used to describe the word "shirt" but not "cat". Such a mechanism has a significant advantage over long-short term memory networks and gated RNNs, which only references tokens that have appeared *before* the current one. The attention mechanism allows the model to capture dependencies between distant parts of a sequence.

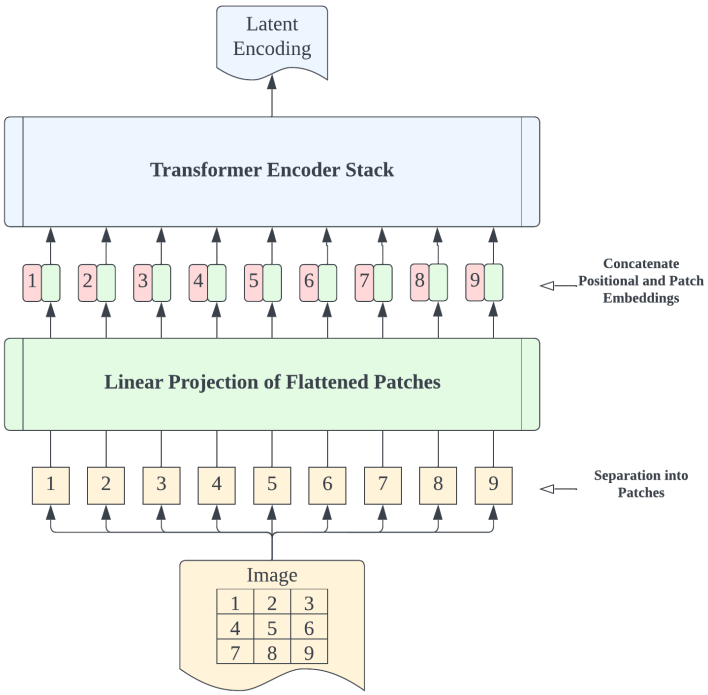
Architecture. Figure 4a shows the architecture of a text transformer. The encoder contains N transformer blocks (in blue) that are stacked together. Each token in the text

document is first passed through a text embedding model and concatenated with its positional embedding (which captures the location of the token in the sequence), before being passed into the transformer stack. Figure 4b shows the architecture of an image transformer. The encoder is similar to that of its text counterpart. Since images are multi-dimensional (at least 2D) objects, each image is first segmented into "patches", 9 in this scenario. Each patch is then flattened and fed through a linear projection to produce patch embeddings. Then, two vectors, one for positional embeddings and another for patch embeddings, are concatenated and used as input for the transformer stack.[3]

[3]Positional embeddings contain simply positional information. For texts, this may be a part-of-speech tag (e.g. noun, verb, etc.). For images, this denotes where the patch belongs in the original image.



(a) Text transformer encoder.



(b) Image transformer encoder.

Figure 4: Architecture of transformer model encoders. Note the $\oplus$ sign represents vector concatenation.

3.2 Modality Invariance in Multi-Modal Systems

Modality invariance refers to the ability of a multi-modal model, system or pipeline to learn representations that are modality-independent.

Various works have been done to evaluate the importance of modality invariance and ways to reduce modality bias.

As each modality requires its own encoder to extract features from, when trained disjointly, the latent representations are modality-dependent. For example, two records of the same concept, but from different modalities, may be encoded into drastically different representations. This is known as the modality gap (Hazarika et al. 2020).

Furthermore, as records from each modality contain a different amount of information, one modality may "overpower" another in contributing to the learning of a system. More generally, training can be dominated by one or more modalities, resulting in sub-optimal performance during inference. For example, a modality may contribute more to the learning of a system due to a higher correlation between the inputs and the labels. This is known as the modality bias (Guo et al. 2023).

Modality gaps and modality biases are more pronounced when the sub-models for each modality are trained independently, such as in W. Wang et al. 2014.

On the contrary, works such as Hu et al. 2019 and Yuan et al. 2021 showed that a combined training framework for the various components can effectively reduce modality gaps. Ideally, two encoded representations, regardless of modality, should be "close by" in the latent representation space if they represent the same concept, and "far away" otherwise. The distance or similarity measure between any two representations should only be affected by the entity they represent, but not by the modality of the input. This is known as modality invariance, as illustrated in Figure 5.

3.3 Proposed Architecture: TransforMMER

With this in mind, we propose the **TransforMMER** family of models, which incorporates the following philosophies:

1. The model should only require raw data as input. Learning and making inferences on

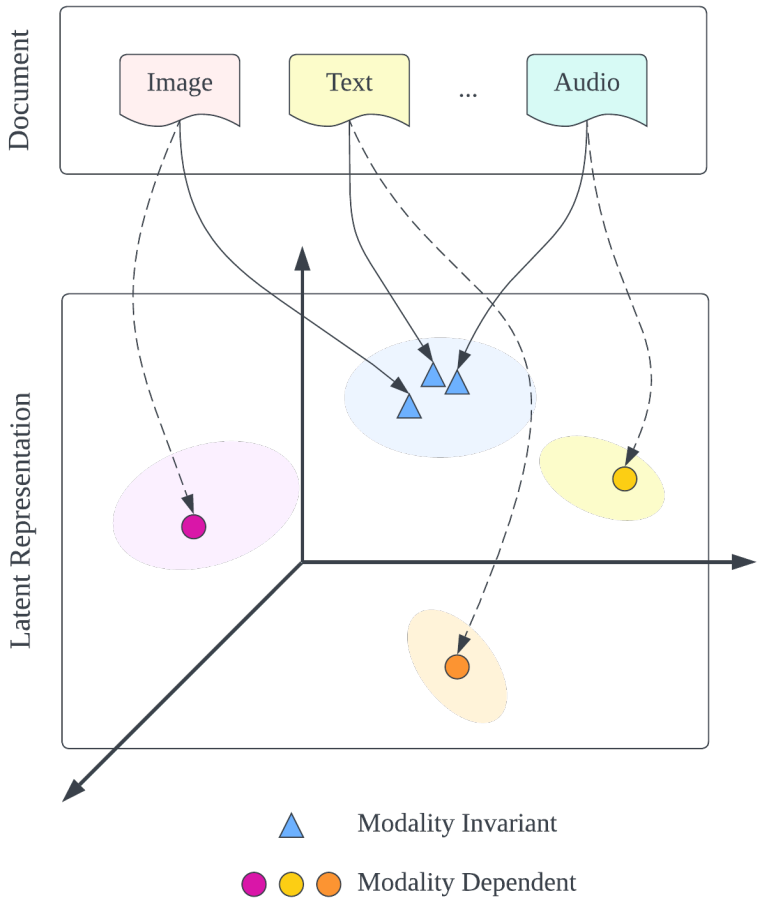


Figure 5: Modality invariance vs modality dependent representations for documents of the same concept.

raw data bypasses the need for any pre-processing into or storage of low-level features, bringing significant benefits to production systems. Furthermore, learning and making inferences on low-level features, such as index vectors from bag-of-words text models, are not extensible. For example, a text encoder trained on a dataset with a vocabulary size of 100 needs to be retrained when new records containing words that are out of vocabulary appear. Learning and making inferences on raw data is achieved through the use of transformers, introduced in subsection 3.1.

2. The model should be modal-invariant, where the latent representations produced by two modalities should only differ by the concept of the inputs, not by the modality, as discussed in subsection 3.2.

TransforMMER. The proposed model, TransforMMER is based on a joint generative-adversarial network architecture, similar to the works by Hu et al. 2019 and Shao et al. 2020. However, the differentiating factor is that the TransforMMER models incorporate state-of-the-art transformer architectures in the feature encoding layer, which allows the models to take in raw data directly.

The model comprises three main sections:

- Transformer encoders: take in a raw record and produce its latent representation. There is one encoder for each modality.
- Predictor: takes in a latent representation and predicts the concept (or concepts) of the input.
- Modal discriminator: takes in a latent representation and predicts which modality it has come from.

Two resolution scenarios are considered:

1. Multi-nomial (also known as multi-class or multi-output) retrieval: each record belongs to **at least** one of K concepts. The predictor outputs a vector of length K, where a "1" in the i^{th} position indicates that the record is predicted to represent the i^{th} concept. The multi-nomial variant of the architecture is illustrated in Figure 6a.

2. Multi-label retrieval: each record belongs to **exactly** 1 of K concepts. The predictor outputs a single value representing a class [1, K]. The multi-label variant of the architecture is illustrated in Figure 6b.

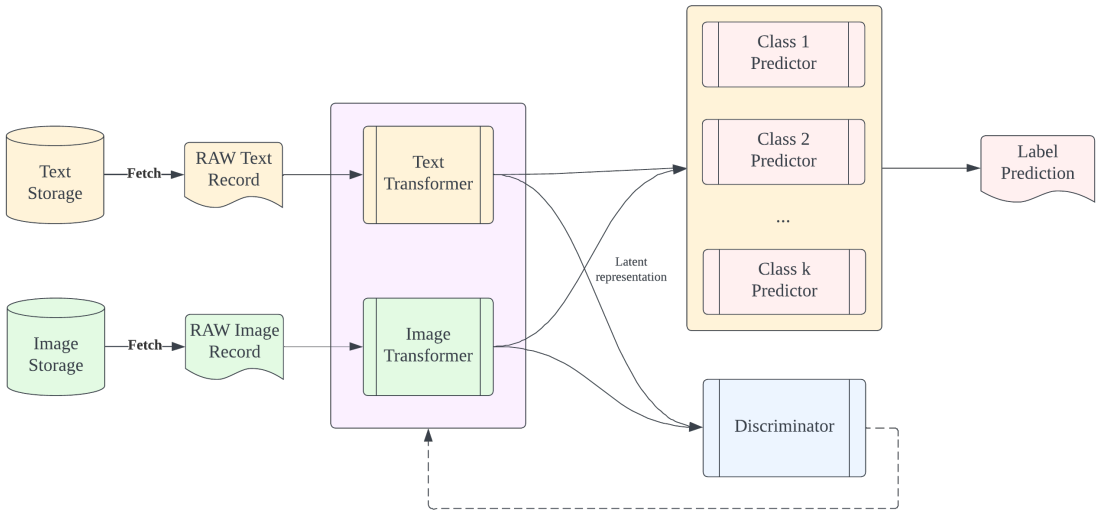
3.4 Joint Training Scheme

The model is trained jointly. The loss function is a weighted sum of the generative loss from either modality and the adversarial loss from the discriminator.

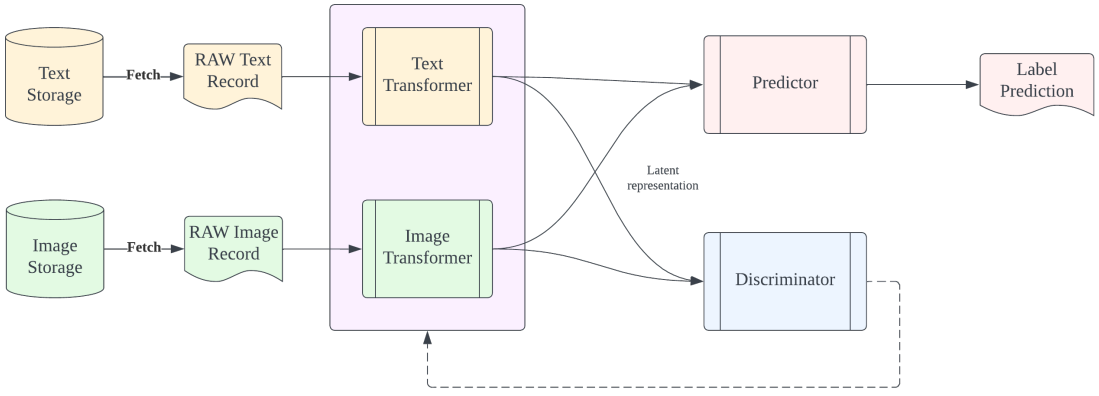
We train the model based on the following joint loss function:

$$\begin{aligned}\mathcal{L}\left(X_I, Y_I, X_T, Y_T\right) = \alpha_I \mathcal{L}_g\left(\mathcal{G}(X_I), Y_I\right) + \alpha_T \mathcal{L}_g\left(\mathcal{G}(X_T), Y_T\right) \\ + \beta\left[\mathcal{L}_d\left(\mathcal{D}\left(\mathcal{G}(X_I)\right), Z_I\right) + \mathcal{L}_d\left(\mathcal{D}\left(\mathcal{G}(X_T)\right), Z_T\right)\right],\end{aligned} \tag{2}$$

where X, Y are the matrix representations of the input and labels respectively, and Z is an indicator variable for the modality, with subscripts $_{I,T}$ representing image and text respectively. $\mathcal{G}, \mathcal{D}$ are the generator (encoder) and discriminator models, $\mathcal{L}_g, \mathcal{L}_d$ are the generative and discriminator losses respectively. $\alpha_I, \alpha_T, \beta > 0$ are hyper-parameter coefficients.



(a) Multi-nomial or multi-class retrieval.



(b) Multi-label retrieval.

Figure 6: Proposed architecture of the TransforMMER models.

4 Experiments

4.1 Datasets

We survey the dataset Coco-2017 (T.-Y. Lin et al. 2015)[4]. Details of the dataset are shown in Table 1.

Dataset	**Modality**	**Size**	**Concepts** (K)	**Features**	
				Image	Text
Coco-2017	Image + Text	117,266	80	Raw + CH	Raw + BOW
Coco-2017-A	Image + Text	53,151	79	Raw + CH	Raw + BOW

Table 1: Details of the Coco-2017 dataset.

The Coco-2017 dataset is a multi-nomial large-scale dataset developed for object detection and captioning. However, as it contains images, texts and topics, it is also suitable for multi-modal entity resolution. A major benefit of this dataset is that it contains images and texts in their raw formats. It contains 117,266 images from 80 concepts, each of which is associated with a caption. The distribution of the top 20 concepts in the dataset is shown in Figure 7. Note that as the dataset is multi-nomial, where each record belongs to at least one concept, the sum of the frequencies exceeds the total size of the concept.

We observe that the distribution of concepts is heavily skewed. A significant portion of the data (64,115 to be exact) belong to the "person" concept. The presence of non-uniform distribution in the training data is a major source of over-fitting in deep learning models. Hence we created an augmented dataset, marked "Coco-2017-A", to exclude any

[4]The citation is for MS-COCO family of datasets, there wasn't any new papers published for the updated Coco-2017 dataset

data belonging to the "person" concept.

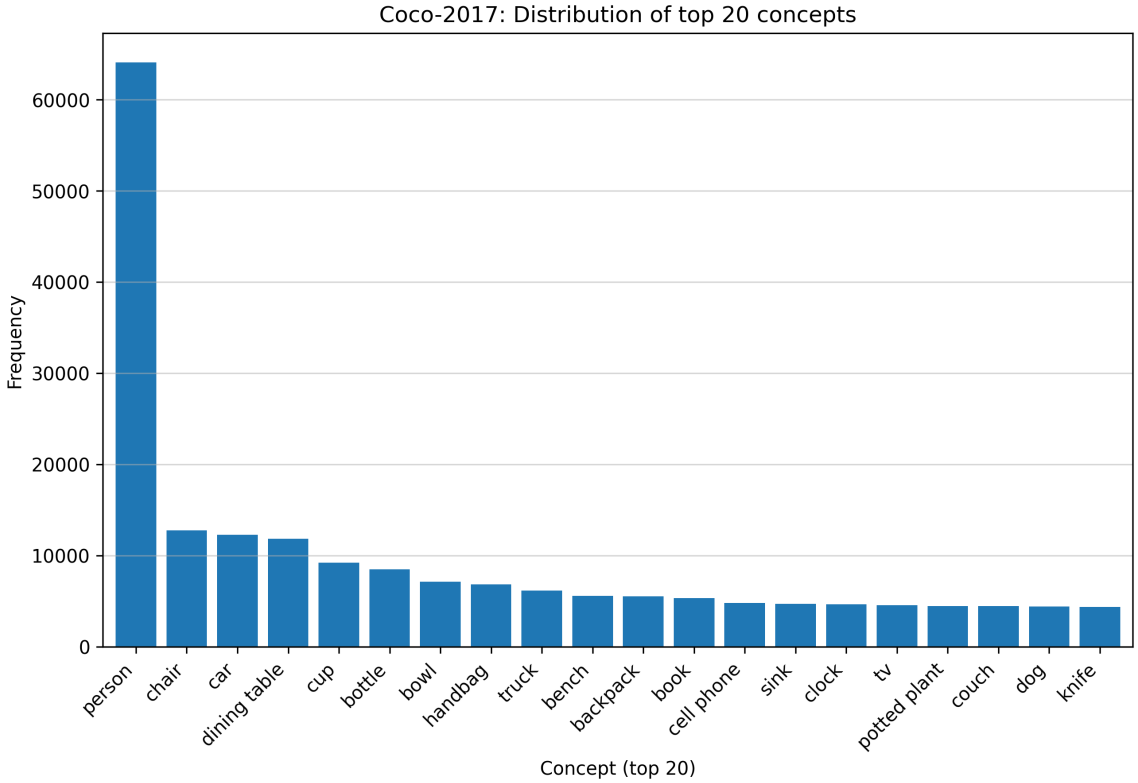


Figure 7: Distribution of concepts in the Coco-2017 dataset.

Note that since existing methods MSAE and MAN take in low-level features as inputs instead of raw data, we have performed an extra step of pre-processing text and images in Coco-2017 prior to training and evaluation. Pre-processing is performed in a similar way as in the literature, producing colour histograms (denoted "CH") from images and Bag-of-Words indexes (denoted "BOW") from texts.

4.2 Evaluation Criteria

To evaluate and compare the performance of models, we need to make use of various metrics to describe the outcome of inferences made. The end goal is to develop a model that infers as many correct matches as possible while reducing irrelevant matches.

Entity resolution can be viewed abstractly as a classification task, to classify records into their inherent classes. We make use of a classification metric, precision, as a basis for evaluating the performance of the models. Precision is defined as:

$$\text{Precision} = \frac{|\ \{\text{relevant documents}\} \cap \{\text{retrieved documents}\}\ |}{|\ \text{retrieved documents}\}\ |} \tag{3}$$

For entity resolution with more than 2 concepts, precision can be adapted to the use case of multi-class classification through macro and micro averaging. **Average precision** (AP) for a single query is defined as:

$$\text{AP} = \frac{1}{T} \sum_{r=1}^{R} P(r)\delta(r), \tag{4}$$

where R is the total number of records retrieved, $T \leq R$ is the number of relevant records retrieved, $P(r)$ is the precision for the top r records and $\delta(r)$ is an indicator variable, where $\delta(r) = 1$ if the rth retrieved document belongs to the same concept as the query, else $\delta(r) = 0$ (Rasiwasia et al. 2010).

Furthermore, with Q queries, the **mean average precision** (MAP) is the mean of the APs across all queries:

$$\text{MAP} = \frac{1}{Q} \sum_{i=1}^{Q} \text{AP}_i. \tag{5}$$

Lastly, we denote MAP@k to be the mean of the AP over the top k documents retrieved for each query.

4.3 Experimentation

We performed two types of experiments: prolonged training and few-shot learning.

Prolonged training. Similar to prior works, we train the proposed model for a prolonged period of time to near-optimality. The models are fine-tuned to perform entity resolution on the new datasets.

Few-shot learning. Given that the sub-models used are pre-trained, we evaluate their few-shot capabilities in knowledge transfer and ability to learn with limited data. We train the model with a certain number of samples per concept, denoted $\mathcal{S}$, and evaluate its performance (MAP score) on an independent test set. Following the convention of few-shot learning, in the ith iteration, we select $\mathcal{S}_i = 2^i$ samples to created dataset $\mathcal{D}_i$, where $i = 1, 2, \ldots$ Furthermore, note that the samples are cumulative, formally, $\mathcal{D}_i \subset \mathcal{D}_{i+1}$.

4.3.1 Evaluation

Queries. Four types of queries are evaluated:

- $\mathcal{Q}_{I \to I}$: image to image queries,
- $\mathcal{Q}_{T \to T}$: text to text queries,
- $\mathcal{Q}_{I \to T}$: image to text queries,
- $\mathcal{Q}_{T \to I}$: text to image queries.

Metrics. Two criteria are presented: mean average precision (MAP) and MAP@5, the mean of the AP over the top 5 documents retrieved for each query, both as defined in subsection 4.2.

4.3.2 Other Considerations

Dataset. Due to the proposed model taking raw inputs, we are limited in the datasets to perform experimentation on. Hence, the Coco-2017 data is used in our evaluations. We compare our proposed model, trained on raw features, to existing models after transforming the dataset into the same low-level features in the literature.

Transformers. In our experiments, we rely on the following pre-trained transformers: CLIP image model, BEiT image model, CLIP text model and RoBERTa text model. It is worth noting that both the CLIP and the BEiT image models use augmented versions of the vision transformer (ViT) model. CLIP models were developed through an image-text contrastive pre-training framework, and the BEiT image model was pre-trained through a self-supervised framework. The use of pre-trained transformers serves as a proof-of-concept, and their combinations are evaluated to determine a baseline. In fact, these modules can be swapped for other transformers, which could achieve better results in certain use cases.

5 Results and Discussions

We present the results of TransforMMER models and compare them to existing models in Table 2. The models were trained and evaluated on both the *original* Coco-2017 dataset and an augmented version of it (Coco-2017-A), which excludes any data points associated with the "person" concept, as previously described in subsection 4.1. Results obtained on the original dataset are presented in Table 2a, and those from the augmented dataset are presented in Table 2b.

5.1 Strengths

5.1.1 Comparison with Existing Solutions

We present the results of training the model on the full dataset. In the "Model" column, the TransforMMER models are labelled as "Ours (⟨Img⟩ + ⟨Txt⟩)", where ⟨Img⟩ and ⟨Txt⟩ represent the configuration of the respective underlying image and text transformer models.

Overall results. Overall, it is clear that the TransforMMER models outperform existing solutions in all 4 types of queries. Moreover, the improvement is significant in text-to-text, image-to-text and text-to-image queries. Our results demonstrate that the use of pre-trained transformers to encode raw data can significantly enhance retrieval results. These exemplary results serve as proof of concept for the effectiveness of this approach in multi-modal entity resolution.

Robustness to imbalanced data. Furthermore, results indicate that TransforMMER models are robust to imbalances in the statistical distribution of the training data. Previous

Model	$\mathcal{Q}_{I \to I}$		$\mathcal{Q}_{T \to T}$		$\mathcal{Q}_{I \to T}$		$\mathcal{Q}_{T \to I}$	
	MAP	MAP@5	MAP	MAP@5	MAP	MAP@5	MAP	MAP@5
MSAE	0.301	0.464	0.337	0.478	0.298	0.382	0.303	0.376
MAN	0.448	0.545	0.591	0.638	0.476	0.638	0.566	0.584
Ours (CLIP+CLIP)	0.488	0.677	0.996	**0.999**	0.579	**0.999**	0.578	0.701
Ours (CLIP+RoBERTa)	0.481	0.752	0.926	0.984	0.607	0.984	0.603	0.784
Ours (BEiT+CLIP)	**0.565**	**0.778**	**0.997**	**0.999**	**0.675**	**0.999**	**0.653**	**0.804**
Ours (BEiT+RoBERTa)	0.516	0.756	0.901	0.962	0.624	0.971	0.620	0.796

(a) MAP and MAP@5 scores for on the Coco-2017 dataset.

Model	$\mathcal{Q}_{I \to I}$		$\mathcal{Q}_{T \to T}$		$\mathcal{Q}_{I \to T}$		$\mathcal{Q}_{T \to I}$	
	MAP	MAP@5	MAP	MAP@5	MAP	MAP@5	MAP	MAP@5
MSAE	0.389	0.486	0.421	0.499	0.325	0.407	0.343	0.391
MAN	0.456	0.577	0.591	0.668	0.587	0.680	0.566	0.598
Ours (CLIP+CLIP)	0.493	0.715	0.996	0.999	0.609	0.999	0.609	0.732
Ours (CLIP+RoBERTa)	0.469	0.703	0.823	0.957	0.530	0.929	0.524	0.724
Ours (BEiT+CLIP)	0.514	0.744	**0.999**	**1.000**	**0.655**	**1.000**	**0.655**	0.790
Ours (BEiT+RoBERTa)	**0.515**	**0.758**	0.943	0.985	0.643	0.985	0.638	**0.798**

(b) MAP and MAP@5 scores for on the *augmented* Coco-2017-A dataset.

Table 2: Comparison of results on the Coco-2017 datasets.

methods, in contrast, produce highly varying results when dominant concepts exist in the dataset. When moving from the augmented dataset (balanced) to the original dataset (imbalanced), across all 4 queries, the MAP results of the MSAE and MAN models reduce by a maximum of 0.09 and 0.11 respectively. However, the MAP results of the TransforMMER models decreased by at most 0.04.

Furthermore, the MAP results of both MAN and MSAE are consistently lower for the imbalanced dataset. Such a trend is not observed for the TransforMMER models. In fact, in some queries, the MAP results on the imbalanced dataset are higher.

This robustness is critical for real-world applications, as it is unrealistic to assume that training data has uniformly distributed concepts. Instead, the distribution likely follows an exponential decay pattern, with a few concepts dominating the dataset and many concepts having limited data points.

5.1.2 Few-Shot Learning Capabilities

After performing training on the entire dataset, we selected the best-performing combination of models (BEiT image encoder + CLIP text encoder) and evaluate its few-shot learning performance. Few-shot learning is performed on the *original* Coco-2017 dataset. The results are tabulated in Table 3. We plot the trends in Figure 8 and document the time taken to train (in 1000 seconds) in Figure 9.

With limited data, the TransforMMER (BEiT + CLIP) model did not perform as well compared to when it was trained for a prolonged period. This is to be expected. However, we observe an upsurge in MAP results during training within the first 32 samples per class. Furthermore, from Figure 9, we observe a linear trend between $\mathcal{S}$ and the average time to train.

Such capabilities are beneficial in various ways (non-exhaustive):

- Improved efficiency: with few-shot learning, the model can learn from a small sample of data, reducing reliance on large-sized, labelled datasets and required time-to-train. This significantly improves the efficiency of the training process, saving both resources and time.

- Improved flexibility: the TransforMMER model can be fine-tuned quickly. For example, when we need to perform entity resolution on a new concept which has not been seen before. The increased flexibility allows for faster prototyping.

S	$\mathcal{Q}_{I \to I}$	$\mathcal{Q}_{T \to T}$	$\mathcal{Q}_{I \to T}$	$\mathcal{Q}_{T \to I}$	Time (sec)
1	0.20	0.41	0.39	0.37	5
2	0.29	0.46	0.42	0.39	535
4	0.36	0.53	0.45	0.44	1,729
8	0.39	0.53	0.49	0.49	3,048
16	0.43	0.55	0.53	0.53	5,841
32	0.46	0.54	0.53	0.54	11,615
64	0.44	0.56	0.52	0.52	21,762
128	0.47	0.58	0.56	0.54	61,167

Table 3: MAP and Average Time-to-Train in seconds for few-shot learning on the Coco-2017 dataset, **S** represents the no. of samples per concept.

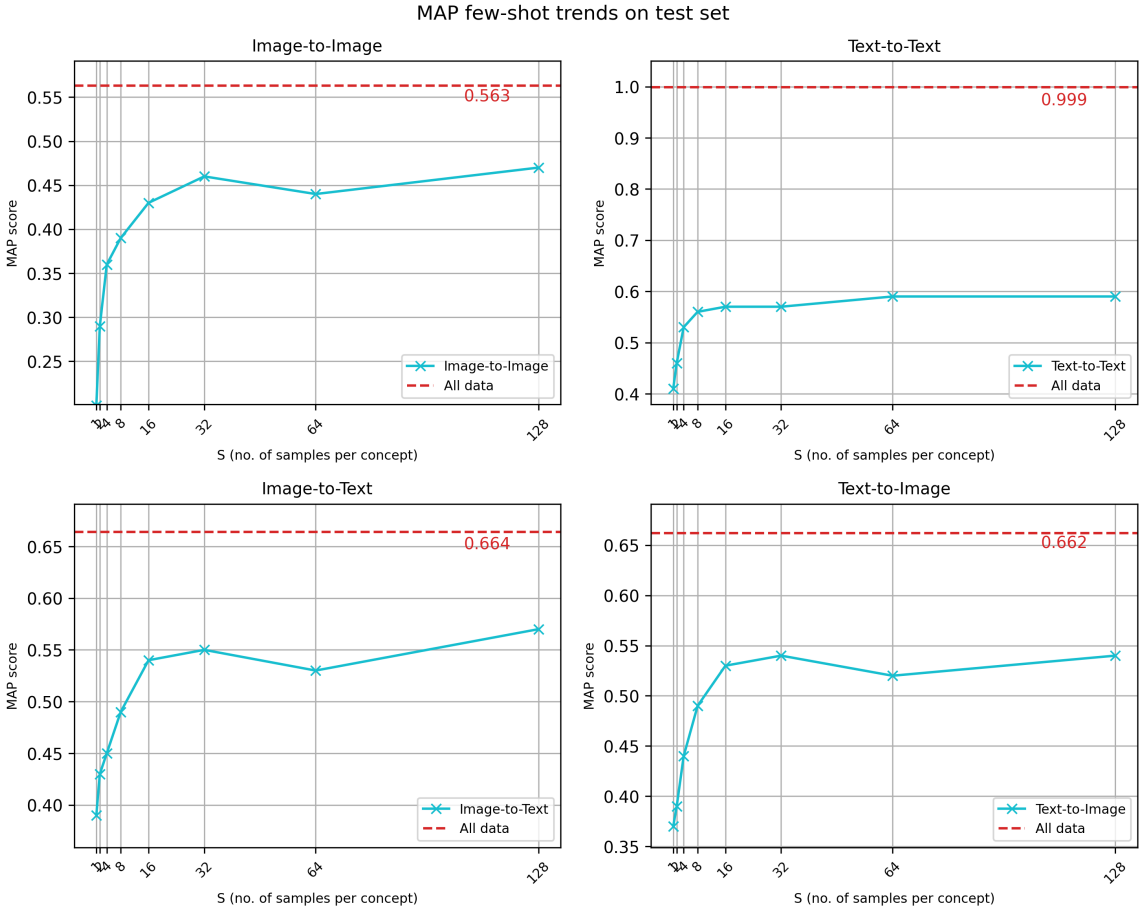


Figure 8: Trend of MAP with few-shot learning.

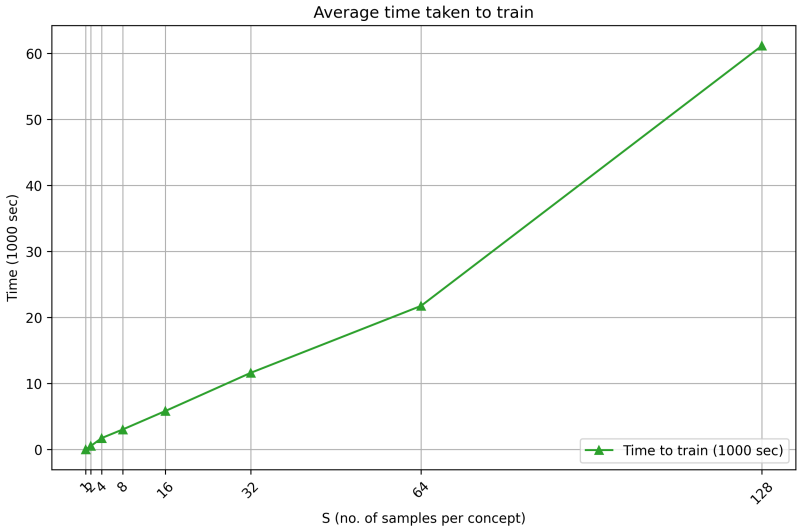


Figure 9: Average time taken to train with few-shot learning

5.2 Shortcomings

Although the overall results produced by the TransforMMER models have beaten previous solutions, they are not without weaknesses. Notably, it performs extremely poorly in some concepts, such as "backpack" and "knife" in image-related queries. Upon inspection of the images, we noticed that the objects representing these concepts are obscured in the images in the dataset. Examples include the images Figure 10a and Figure 10b, which are labelled as "backpack" and the images Figure 10c and Figure 10d, which are labelled as "knife". It is obvious the objects related to the concepts are difficult to spot by humans, let alone by an algorithm. The model's poor performance on these concepts can be attributed to the fact that, in most instances, the concept-relevant object only takes up a small portion of the image.

Furthermore, similar to Figure 10a, many images in the *original* Coco-2017 dataset which belongs to multiple concepts are accompanied by a person as the main subject. As a

result, the model overfits onto the person but not the backpack (or other concepts present in the image), which is a serious source of bias during training. This is a significant challenge posed in multi-nomial systems (where each data point contains multiple concepts).

While it is not difficult to verify whether a dataset has a dominating concept or to quantify the "quality" of a data point [5], blindly removing concepts or data points during data preparation may trivialise the problem at hand. By removing the concept "person", there is a major shift in the distribution in the Coco-2017 dataset. However, we could afford to do so, as the dataset is large and there were 80 concepts in total (79 after augmenting), as detailed in Table 1. Nonetheless, such a privilege could not be generalised to other datasets.

5.3 Limitations of Project

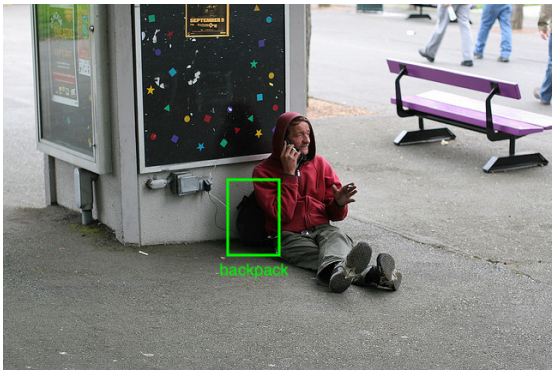
Our project is not without its limitations.

Firstly, TransforMMER models heavily rely on pre-trained generative models, which already performed well in sequence-to-sequence content generation. Hence, they are expected to perform decently for other purposes as well.

This is in contrast to existing solutions, which have to be trained from some random initialisation. Furthermore, these models have enormous sizes (BEiT has 307 million parameters, RoBERTa has 123 million parameters), making it unfeasible to train the entirety of the models when under the constraints of computational resources. For instance, the CLIP model with ViT was trained on 256 NVIDIA Tesla V100 GPUs over 12 days. In contrast, we only have access to 1 A100 or V100 GPU with a maximum runtime of 72 hours. Given the limitations, in this project, we selectively retrain only the downstream layers of the models. Despite so, exemplary results have been shown through empirical experiments. It

[5] For example, we can quantify the proportion each concept takes up in an image, through some object-detection algorithm.

(a) Sample image of a backpack.



(b) Sample image of a backpack.

(c) Sample image of a knife.

(d) Sample image of a knife.

Figure 10: Samples of images from concepts with poor performance.

is worth noting that if we had more resources at our disposal, perhaps better results could have been obtained for image-to-image queries.

Lastly, as most prior solutions take various low-level features as input, there is an absence of multi-modal datasets with raw features for the use of entity resolution. Aside from raw features, datasets also need to include the concept(s) that each record belongs to. We have come across another dataset with both images and text existing in the raw form, however, lacks any labelling of the concept. Performing Latent Dirichlet Allocation (LDA) topic modelling on the text descriptions resulted in nearly 1,300 unique concepts for 12,000 documents, with each record belonging to only 1 concept. Worryingly, many concepts contained only 1 record, making training and evaluations outstandingly difficult, if not impossible[6]. Hence, it was discarded, leaving us with only 1 usable dataset for experimentation.

5.4 Runtime Environment

To obtain the results, each configuration of the TransforMMER model is trained for 72 hours for at least 20 epochs. However, it is worth noting that the MAP metrics have stabilised within the first 10 epochs. Training could have been cut short without noticeable compromises to the results.

The training is accelerated on a single NVIDIA Tesla A100 or V100 GPU.

[6] Can't split a concept into train and test sets if there's only 1 data point.

6 Future Works

Our proposed TransforMMER models heavily rely on various pre-trained transformer architectures as the backbone.

The use of transformers and attention mechanisms is effective in many NLP tasks, such as in machine translation (cite), summarisation (cite), question-answering (cite), etc. Large language models (LLMs) such as GPT-3 have been extensively studied.

In contrast, vision transformers have yet to be studied as deeply as their language counterparts. In this project, we have shown that vision transformers can be used to learn representations from raw images, and then perform entity resolution on the learned representations. However, there are still many aspects that can be explored further.

6.1 Context-based Resolution

Current entity resolution methods are single-query focused and produce "one size fits all" solutions. As a result, the burden is on the user to perform subsequent queries to retrieve more documents, e.g. by creating more specific queries. Hence, it is likely that queries performed in a short period are related. Therefore, an area of research is in temporal context-related resolution, where the model takes into account previous queries to predict the underlying context for the subsequent queries. Future systems should incorporate a certain degree of "preemption" to rank retrieved documents that are relevant to the search context.

6.2 Logical Relationship Reasoning

Logical reasoning. An area of research is in the development of models that can perform not only entity resolution but also to perform relationship reasoning. For example, in

developing a system that can retrieve segments of videos relevant to the prompt "a girl in a red shirt petting a cat", the model needs to not only identify the entities (girl, red shirt and cat) but also deduce the "petting" relationship between the different entities.

Inverted search. Our TransforMMER models have shown successes in retrieving a set of documents $\mathcal{D}$ relevant to a given query $\mathcal{Q}$. When given a query "yellow-coloured cats", the system could retrieve the set of images featuring said cats. However, it prompted whether the inverse would be possible: given a set of documents, $\mathcal{D}$, where some relation R exists in every $d \in \mathcal{D}$, could we find a query $\mathcal{Q}$ to capture R such that when fed into an ER system, the original set $\mathcal{D}$ would be retrieved? For example, given a set of images of yellowed-coloured objects, the system should produce an output describing the common feature, say "yellow-coloured objects". Such a system would be beneficial in finding and visualising similarities and concepts that co-exist in some document set $\mathcal{D}$.

References

Chatterjee, Shubham and Laura Dietz (July 6, 2022). "BERT-ER: Query-specific BERT Entity Representations for Entity Ranking". In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. Madrid Spain: ACM, pp. 1466–1477.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (May 24, 2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv: 1810.04805[cs].

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby (June 3, 2021). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. arXiv: 2010.11929[cs].

Guo, Yangyang, Liqiang Nie, Harry Cheng, Zhiyong Cheng, Mohan Kankanhalli, and Alberto Del Bimbo (Aug. 31, 2023). "On Modality Bias Recognition and Reduction". In: *ACM Transactions on Multimedia Computing, Communications, and Applications* 19.3, pp. 1–22.

Hazarika, Devamanyu, Roger Zimmermann, and Soujanya Poria (Oct. 12, 2020). "MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis". In: *Proceedings of the 28th ACM International Conference on Multimedia*. MM '20: The 28th ACM International Conference on Multimedia. Seattle WA USA: ACM, pp. 1122–1131.

Hochreiter, Sepp and Jürgen Schmidhuber (Nov. 1, 1997). "Long Short-Term Memory". In: *Neural Computation* 9.8, pp. 1735–1780.

Hu, Peng, Dezhong Peng, Xu Wang, and Yong Xiang (Sept. 2019). "Multimodal adversarial network for cross-modal retrieval". In: *Knowledge-Based Systems* 180, pp. 38–50.

Kumar, Shaishav and Raghavendra Udupa (July 2011). "Learning Hash Functions for Cross-View Similarity Search". In: *Proceedings of the Twenty-Second international joint conference on Artificial*. International Joint conference on Artificial 2011. Vol. 2. Barcelona, Catalonia, Spain: AAAI Press, pp. 1360–1365.

Li, Yuliang, Jinfeng Li, Yoshihiko Suhara, AnHai Doan, and Wang-Chiew Tan (Sept. 2020). "Deep Entity Matching with Pre-Trained Language Models". In: *Proceedings of the VLDB Endowment* 14.1, pp. 50–60. arXiv: 2004.00584[cs].

Lin, Tsung-Yi, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár (Feb. 20, 2015). *Microsoft COCO: Common Objects in Context*. arXiv: 1405.0312[cs].

Lin, Zijia, Guiguang Ding, Mingqing Hu, and Jianmin Wang (June 2014). "Multi-label Classification via Feature-aware Implicit Label Space Encoding". In: *Multi-label Classification via Feature-aware Implicit Label Space Encoding*. 31st International Conference on Machine Learning, ICML 2014. Vol. 2, p. 9.

Lu, Xinyan, Fei Wu, Siliang Tang, Zhongfei Zhang, Xiaofei He, and Yueting Zhuang (July 28, 2013). "A low rank structural large margin method for cross-modal ranking". In: *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. SIGIR '13: The 36th International ACM SIGIR conference on research and development in Information Retrieval. Dublin Ireland: ACM, pp. 433–442.

Mudgal, Sidharth, Han Li, Theodoros Rekatsinas, AnHai Doan, Youngchoon Park, Ganesh Krishnan, Rohit Deep, Esteban Arcaute, and Vijay Raghavendra (May 27, 2018). "Deep Learning for Entity Matching: A Design Space Exploration". In: *Proceedings of the 2018 International Conference on Management of Data*. SIGMOD/PODS '18: International Conference on Management of Data. Houston TX USA: ACM, pp. 19–34.

Oord, Aaron van den, Yazhe Li, and Oriol Vinyals (Jan. 22, 2019). *Representation Learning with Contrastive Predictive Coding*. arXiv: 1807.03748[cs,stat].

Papadakis, George, Ekaterini Ioannou, Emanouil Thanos, and Themis Palpanas (2021). *The Four Generations of Entity Resolution*. Synthesis Lectures on Data Management. Cham: Springer International Publishing.

Parmar, Niki, Ashish Vaswani, Jakob Uszkoreit, Łukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran (June 15, 2018). *Image Transformer*. arXiv: 1802.05751[cs].

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer (Mar. 22, 2018). *Deep contextualized word representations*. arXiv: 1802.05365[cs].

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever (Feb. 26, 2021). *Learning Transferable Visual Models From Natural Language Supervision*. arXiv: 2103.00020[cs].

Rasiwasia, Nikhil, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert R.G. Lanckriet, Roger Levy, and Nuno Vasconcelos (2010). "A new approach to cross-modal multimedia retrieval". In: *Proceedings of the international conference on Multimedia - MM '10*. International conference on Multimedia. Firenze, Italy: ACM Press, p. 251.

Shao, Jingyu, Qing Wang, Asiri Wijesinghe, and Erhard Rahm (Dec. 17, 2020). *ErGAN: Generative Adversarial Networks for Entity Resolution*. arXiv: 2012.10004[cs].

Szegedy, Christian, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi (Aug. 23, 2016). *Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning*. arXiv: 1602.07261[cs].

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin (Dec. 5, 2017). *Attention Is All You Need*. arXiv: 1706.03762[cs].

Wang, Kaiye, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang (July 21, 2016). *A Comprehensive Survey on Cross-modal Retrieval*. arXiv: 1607.06215[cs].

Wang, Wei, Beng Chin Ooi, Xiaoyan Yang, Dongxiang Zhang, and Yueting Zhuang (Apr. 2014). "Effective multi-modal retrieval based on stacked auto-encoders". In: *Proceedings of the VLDB Endowment* 7.8, pp. 649–660.

Wilke, Moritz and Erhard Rahm (May 25, 2021). "Towards Multi-modal Entity Resolution for Product Matching". In: *32nd GI-Workshop on Foundations of Databases*, p. 5.

Wu, Yiling, Shuhui Wang, and Qingming Huang (Feb. 2019). "Multi-modal semantic autoencoder for cross-modal retrieval". In: *Neurocomputing* 331, pp. 165–175.

Xie, Saining, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He (Apr. 10, 2017). *Aggregated Residual Transformations for Deep Neural Networks*. arXiv: 1611.05431[cs].

Yuan, Xin, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta (June 2021). "Multimodal Contrastive Training for Visual Representation Learning". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA: IEEE, pp. 6991–7000.

Zhen, Yi and Dit-Yan Yeung (2012). "A probabilistic model for multimodal hash function learning". In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12*. the 18th ACM SIGKDD international conference. Beijing, China: ACM Press, p. 940.

Appendix A

Preliminary Experiments

This section extends the preliminary experiments, submitted in my interim report.

Table A.1 compares the two datasets surveyed as part of my preliminary experiments. The Coco-2017 dataset, introduced in subsection 4.1, is included for a full comparison.

Dataset	**Modality**	**Size**	**Concepts** (K)	**Features**	
				Image	Text
NUS-WIDE	Image + Tags	186,577	81	Histogram, Correlogram, etc.	Tag occurrences
WIKI	Image + Paragraphs	2,866	10	SIFT + BOW	LDA
Coco-2017	Image + Text	117,266	80	Raw	Raw
Coco-2017*	Image + Text	53,51	79	Raw	Raw

Table A.1: Comparison between all datasets.

The WIKI Image-Text dataset is generated from Wikipedia's featured articles, consisting of 2,866 pairs of images and their captions (in paragraphs). The images are represented as features after being passed through Scale Invariant Feature Transform (SIFT) or Bag of Visual Words (BOW). Paragraphs are represented as features after being transformed with

a Latent Dirichlet Allocation (LDA) model. Each pair belongs to one or more of the 10 semantic classes.

The NUS-WIDE dataset contains 186,577 labelled images. Each image is represented using its low-level features, such as 64-D colour histogram, 144-D colour correlogram, etc. Each image is associated with tags, where each tag represents some word. Each image belongs to one or more of the 81 concepts.

The above datasets are commonly used for training and evaluating existing multi-modal entity resolution solutions. However, as they only contain low-level features but not the raw documents themselves, they are not compatible with our TransforMMER scheme.

Table A.2 documents the results of the performance of the Multiple-Stacked Auto Encoders model (MSAE) and the Multi-modal Adversarial Network model (MAN) on the two datasets.

Dataset	$\mathcal{Q}_{I \to I}$	$\mathcal{Q}_{T \to T}$	$\mathcal{Q}_{I \to T}$	$\mathcal{Q}_{T \to I}$
MSAE	0.342	0.383	0.392	0.327
MAN	0.345	0.401	0.418	0.410

(a) Comparison of different MSAE and MAN on the NUS-WIDE dataset.

Dataset	$\mathcal{Q}_{I \to I}$	$\mathcal{Q}_{T \to T}$	$\mathcal{Q}_{I \to T}$	$\mathcal{Q}_{T \to I}$
MSAE	0.101	0.124	0.130	0.129
MAN	0.141	0.312	0.134	0.132

(b) Comparison of different MSAE and MAN on the WIKI dataset.

Table A.2: Preliminary experiments for MSAE and MAN models