Open-Learning Framework for Multi-modal Information Retrieval with Weakly Supervised Joint Embedding

KMA Solaiman, Bharat Bhargava

Purdue University, West Lafayette, IN, USA
ksolaima@purdue.edu, bbshail@purdue.edu

Abstract

Data scientists often need to find relevant data from various multimedia sources (e.g., organizational databases, data lakes) by submitting query of any media type. Current cross-media retrieval models finds a common representation across multiple media by jointly modeling the encoded features from different modalities. However, supervised and semi-supervised models require a large number of annotated data. In practice, fine-tuning a retrieval model with class label dependence is not only expensive and time-consuming, but often impossible. Moreover, semantic gap between the low level data features and high level human comprehensible features hinder the understanding of the cross-media retrieval results.

We present *WeS-JEm*, a weakly supervised open-learning framework for jointly learning data representations from all modalities in a shared low dimensional vector space, by exploring the structural components of the data samples. The framework characterizes and formulates responses to different novelties encountered during multimodal retrieval from unknown application domains, user requirements, or temporal changes. *WeS-JEm* follows a three-step process: (1) Different modalities of data are translated to textual descriptions. (2) Weak similarity labels are generated among data samples by comparing topics and different structural elements (entities, relationships, and events) of the text. (3) Vector representations are learned for the data samples in the joint embedding space by exploring the relationships among the topics and structural elements. We address the supervision bottleneck problem, and show that topics and structural features can be used as a weak-supervision source, as well as provide a better semantic representation for retrieval of similar multi-modal data. Initial experiments are conducted using documents and videos as multi-modal sources, and topic as weak labels. In comparison to unsupervised methods, LSI and LDA, our model showed promising performance to capture the similarities in the low dimensional space.

Introduction

Finding relevant data from large data sources is the prerequisite for any data analysis task. Current data discovery systems require human hours to sift through the large influx of multi-media data (e.g., text, image, video, audio, and 3-D model) for data preparation task. Multi-modal information retrieval takes queries in one modality to retrieve relevant data from other modalities, augmenting information from a single source with information from other sources. Cross-modal retrieval results can be improved if context is introduced in learning the relevance. For example, uploading the image of a person in google search returns images of similar cloths that the person is wearing. User experience would improve if the search results include the images of similar cloths, videos of people wearing similar cloths, or places where these kind of cloth can be purchased.

Previous works on multi-modal information retrieval have followed the idea of projecting modality-specific features from different modalities into a shared embedding space. (Rasiwasia et al. 2010; Andrew et al. 2013; Wang et al. 2015; Kan, Shan, and Chen 2016; Peng et al. 2017; Zhang et al. 2017; Zhai, Peng, and Xiao 2013) focuses on correlation learning to learn the projection function, using both pairwise information and class labels. (Feng, Wang, and Li 2014; Hu et al. 2019) uses auto-encoders to find correlations. Metric learning methods (Faghri et al. 2017; Xu et al. 2019; Wei et al. 2020; Sah, Gopalakrishnan, and Ptucha 2020) learn a distance function over data objects based on a loss function. Attention mechanism (Sah, Gopalakrishnan, and Ptucha 2020), (Luo et al. 2020; Wang et al. 2017) proposes pre-training models for better generalization. While the aforementioned learning methods exhibit good performance on benchmark datasets, they suffer from the lack of labeled training samples for data discovery in practice. In open world environment, test data distribution is almost always different from the training data distribution. Current works do not focus on the noise in the input data, or the data relevance change over time. Many of the learning methods focus only on uni-modal or bi-modal retrieval, and cannot generate results for queries of all media types. Moreover, most of the above models suffer from the lack of explainable reasoning on how two different multimedia data are similar.

We propose a **We**akly **S**upervised **J**oint **Em**bedding model (WeS-JEm) for multi-modal information retrieval. Our model adopts the metric learning approach (Bellet, Habrard, and Sebban 2013) as the backbone, and proposes to build a data information network as the weak signal generator. It has four components, including a translation module, a weak label generator module, a data information network, and multi-task learning. In detail, we first generate a dense

video caption from the videos using a proposal and captioning module. Then we learn the weak labels using existing single modal encoders and text feature extractors. The separate stream design allows scalability to very large datasets in retrieval tasks. Finally, we create a data information network among all different modalities in terms of their similar features. A multi-task joint objective performs on the network, which aims to learn better representation for each data sample while maintaining multiple degrees of similarity among them. The objective function aims to maintain the inter-connection between different data modalities based on their structural features. Even in the absence of the weak labels, the objectives can be adapted to be trained in an unsupervised setting. The translation module and the weak label generation module are independent of the embedding architecture and can be replaced.

Furthermore, we discuss and formalize novelties in multimodal retrieval task in terms of data shift. As part of our framework, we add a novelty detection and characterization criterion. Finally, we design a pre-training strategy for handling out-of-distribution inputs. It has three parts in our setting. We pretrain the video encoder separately on video captioning task. The weak label generation does not need any pre-training. Then the graph object representations will be pre-trained under the relationship objectives in the final stage. Our contributions are summarized as follows:

1. We propose a multi-modal joint embedding model which is pre-trained with weak supervision. The model can use existing video and image translation models along with the text feature extractors. WeS-JEm can be applied to any application domain for cross modal retrieval.
2. The multi-task joint objective function is built upon a data information network based on how different data samples interact with each other via their structural features.
3. We characterize and formulate novelties in multimodal information retrieval. Proposed framework includes a novelty detection and response module for open-world learning. To the best of our knowledge, this is the first framework that formalizes novelty for cross-modal retrieval task.
4. WeS-JEm has the flexibility to take into account any user provided features and similarity labels during the joint multi-task training. This allows it to be adapted by application domains which already have extracted features. Preliminary experiments demonstrate our model's effectiveness on retrieval and similarity evaluation tasks.

Related Works

Correlation Learning. Traditional cross-modal retrieval models focus on *correlation learning* to project data instances into a latent common subspace. (Rasiwasia et al. 2010) implements linear projection using canonical correlation analysis to optimize only the pairwise information. (Zhang et al. 2017) learns the common features using class labels as a linkage to model correlations. Joint representation learning (Zhai, Peng, and Xiao 2013) constructs graphs to jointly model the correlation and semantic information with sparse and graph regularization. For non-linear projection, deep Canonical Correlation Analysis (DCCA) (Andrew et al. 2013) uses modality specific subnetworks. (Wang et al. 2015) extends DCCA with an auto-encoder regularization term. Multi-view Deep Network (Kan, Shan, and Chen 2016) uses a view-specific and a common sub-network to learn the common space. (Peng et al. 2017) overcome using only shallow networks for common stage with hierarchical networks.

Metric Learning. (Liong et al. 2016) proposes a deep coupled metric learning approach with two hierarchical nonlinear transformations. (Frome et al. 2013) used a hinge rank loss as objective function to map visual and semantic features into the shared space. (Faghri et al. 2017) minimized the loss function using hard negatives with a variant triplet sampling. (Xu et al. 2019) introduced an additional regularization in the loss function with a modality classifier as part of the adversarial learning. (Wei et al. 2020) enables different weighting on positive and negative pairs with an universal weighting framework and a polynomial loss function.

With recent advancements in encoder-decoder networks (Devlin et al. 2018; Ji et al. 2012), (Luo et al. 2020; Wang et al. 2017) provides a solution of pre-training the model on a large scale dataset. (Feng, Wang, and Li 2014) used correspondence autoencoders to find correlations between images and text. (Hu et al. 2019) removes the dependency of jointly learning from all modalities by predefining a common subspace. (Wang et al. 2021) avoids explicitly learning a common space by integrating relation learning. (Solaiman and Bhargava 2021) computes modality specific similarities with neural tensor networks. (Sah, Gopalakrishnan, and Ptucha 2020) uses attention mechanism to align multimodal embeddings learned through a multimodal metric loss function. (Boult et al. 2021) describes a unified framework for formal theories of novelty in learning algorithms, which is applied towards different domains, including multi-agent game, and open world image recognition. (Liu et al. 2021) discusses a self initiated open world learning agent with the example of a conversational bot in a hotel. (Langley 2020) discusses characterization and changes of environments in which a radically autonomous physical agent can operate.

Most of these models require annotated labels specifying which data samples belong to the same category. The novelty frameworks does not explain information retrieval as a domain. WeS-JEm has close resemblance to metric learning and intermediate fusion approaches. Our approach differs from existing works in terms of representation learning methodology and independent module flexibility. Like (Song and Soleymani 2019), we also use modality specific encoders for translation module. The main difference in our proposed metric learning approach lies in building the data information network, and using the structural features as weak labels. Our method does not require annotated labels and we choose the positive and negative pairs such that similarity in structure is maintained. This work has the capability to take both the data instance and data features as query. WeS-JEm is capable of not only encoding the ontological information, but also pairwise and semantic information, if available.

Discussed existing works on retrieval task assume similar training and testing data distribution, and do not reflect on the novelties encountered during test. Existing works on novelty theories have proposed well-established frameworks, but to the best of our knowledge, this is the first work to formalize novelties in a domain with heterogeneous training instances and user-system interdependence.

Methodology

Problem Formulation and Overview

Multi-modal information retrieval is defined as retrieving the results of all modalities by submitting a query of any modality. Existing works tackle the problem in two phases: cross-media feature learning, and similarity measurement. The main contribution of this work is in cross-media feature learning. Formally, we consider the problem of information retrieval from a dataset $\mathcal{D}$ with a collection of data from m modalities. We denote the j-th sample of the i-th modality as $\mathbf{x}_j^i$. Modalities can include text documents, tweets, video snippets, images, and others.

The main goal of this work is to jointly learn high-quality vector representations for individual data samples from unlabeled multi-modal data set. We design our embedding function $\mathcal{F}$ to map the multi-modal data samples into a low dimensional vector space, such that multiple degrees of similarity are preserved in the embedding space. During inference, the similarity between two projected data samples $sim(\mathcal{F}(\mathbf{x}_p^v), \mathcal{F}(\mathbf{x}_q^t))$ will be measured in the joint space, using the existing methods such as the Cosine or the Euclidian distance.

Our main insight is that representing data in terms of different structural features through which different modalities of data can be similar, can provide us with a source of weak supervision for cross-modal retrieval. Our motivation comes from how structural representation of a raw unstructured text allows readers to infer better knowledge. Structural representation of a document entails topics, entities, events, and relationships in the document. Let $\mathcal{A} = \{A_1, A_2, \ldots, A_n\}$ be a set of corresponding features from each data sample. The k-th value of a feature p is denoted by A_p^k. Features consist of topics, metadata, and mid-level structural units (entities, events, relationships etc.) of a data sample that can infer further higher order structures from them in a bottom-up manner. The goal of using topics and structural units as features is to infer an explainable understanding of how different data samples are similar (or, dissimilar). A data sample $\mathbf{x}_j^i$ is a combination of any subset of $\mathcal{A}$. Features are generated automatically in two steps - 1) a textual description of each data sample is generated from any modality; 2) topics, entities, and events are extracted from the textual descriptions and are considered as weak labels for two reasons. *First,* the quality of the extracted structural units rely on the choice of the extraction models, and can be noisy. *Second,* output generated from the modality specific textual descriptors can be ambiguous and noisy.

For our approach, first, we utilize the existing neural network approaches to find a translation from different modalities of data to a textual representation. Then, we create a data information network by connecting data samples to their features via their interactions. Finally, we construct a structure-infused textual representation, by jointly embedding in a single space the data samples, the features in which these data samples are similar, and the similarity labels associated with them. We define a multi-task learning objective capturing the interaction information, by aligning the representation of the data samples, defined by their textual content, with the representation of structural features, based on their on their common relations. Moreover, we formalize novelties or data shift that occurs during test time for retrieval task. We also characterize novelties and include appropriate response for different types of novelties.

Translation Module

Videos $\rightarrow$ ***Textual Description.*** To extract an initial representation of videos, we resort to dense video captioning (DVC). We will use a version of dense video captioning described in (Xu et al. 2018). DVC localizes distinct events in video streams and generates a description for that. As a feature extraction stage, it uses 3D convolutional network (C3D) to encode all incoming frames. For identifying the event boundaries, maintaining the temporal information is important and (Xu et al. 2018) preserves this using convolution and pooling in spatiotemporal space. Using the features from the first stage, in the proposal network, variable-length temporal event proposals are generated and in the final captioning module, they generate a caption for those proposals.

After we have a caption for the whole video, we create the information network from the textual descriptions of all the data samples.

Information Network across Data Samples

There are multiple direct relationships among the data samples and their features. Features can have semantic relationships between them. We define a simple data information graph $G = \{V, E\}$ consisting of several different types of vertices and edges, as follows -

- Let $A_T \subset V$ denote the set of the *topics*.
- Let $A_n \subset V$ denote the set of the *named entities*. A_n is derived from a knowledge base such as, NELL (Mitchell et al. 2015), YAGO (Tanon, Weikum, and Suchanek 2020), Wikidata.
- Let $A_{event} \subset V$ denote the set of the *events*. A_{event} is a sentence describing certain real world events.
- Let $x \subset V$ denote the set of the *data samples*. x has a *data modality* attribute.
- Let $A_{sim} \subset V$ denote the set of *user defined similarity labels*.

The graph vertices are connected via a set of edges described hierarchically, as follows:

- $E_{xA_T} \subset E$: All data samples are connected to their corresponding topics. Note that a data sample can be connected to more than one topic.
- $E_{xA_{event}} \subset E$: All data samples are connected to the events that it describes.

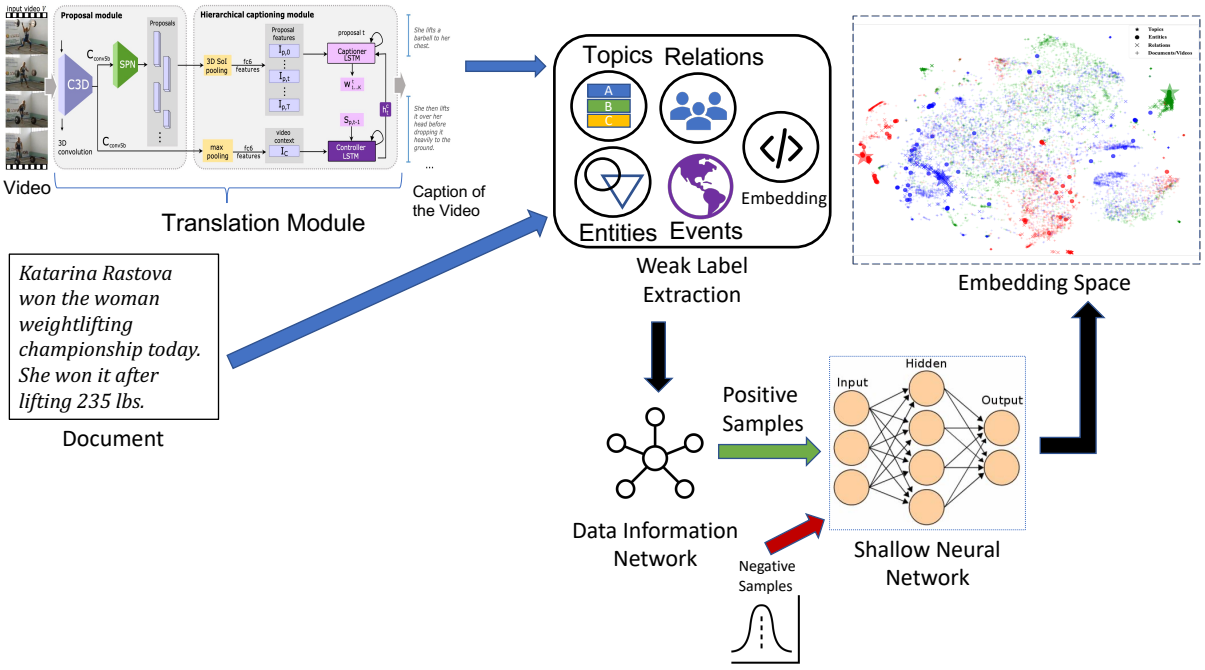

Figure 1: Architecture for Weakly Supervised Joint Embedding for Multi-modal Information Retrieval. Translation module is a two-level dense video captioning model from (Xu et al. 2018).

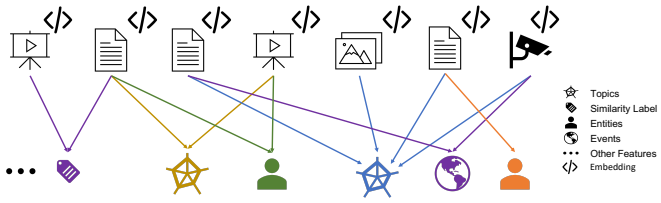

Figure 2: Data Information Network

- $E_{xA_n} \subset E$: All data samples are connected to the entities it describes. Note that an entity may be described by many different data samples.
- $E_{xA_{sim}} \subset E$: If user has defined a similarity between two data samples, both of them are connected to that similarity label.

In addition to the relations expressed in the graph, all the graph nodes are also associated with textual content. Let A_{text} denote the set of the *textual representations* of the data samples. Topics, entities, similarity label, and events also have their own text representation.

Multi-Task Learning

After we have defined the information graph, we need to design the embedding function to map the graph objects into a low dimensional vector space, such that the graph relationships are preserved in the embedding space. In the embedding space, the relations originally defined over the vertices in the information graph are expressed as a similarity score between the vectors representing these vertices. The relationships between the features themselves are also expressed as a similarity score between the vectors representing them. To force these relationship constraints on the data samples, we consider this as a multi-task learning problem, over all the relations in the graph. Jointly learning over all the relationships allows the weak labels to propagate through elements and enforce multiple degrees of similarities. For example, if a document and a video, or two documents have same topic, they should have similar embedding. In parallel, if the two data samples are discussing about the same event, they should have similar embedding. Our embedding function should jointly reflect these similarities.

For each individual graph relation, R, we can define the learning objective as follows:

$$L_R = \sum_i L(o_i, s_i^p, s_i^n) \tag{1}$$

$$\begin{aligned} L(o_i, s_i^p, s_i^n) = \; & y \log sim(o_i, s_i^p)) + \\ & (1-y)\log(1 - sim(o_i, s_i^n))) \end{aligned} \tag{2}$$

where $sim(o_i, s_i^p) = \sigma(e_{o_i} \cdot e_{s_i^p})$;
$sim(o_i, s_i^n) = \sigma(e_{o_i} \cdot e_{s_i^n})$

In equation 1, for each object, o_i in the graph participating in relation R, s_i^p and s_i^n refers to positive and negative examples, respectively. e_{o_i} refers to the vector embedding of the graph object o_i, and y is the label. The objective of the model is to maximize the similarity with a positive example and minimize the similarity with a negative example.

So, $y = 1$ for (o_i, s_i^p) pairs and $y = 0$ for (o_i, s_i^n) pairs since they have been sampled from the noise distribution.

Next, we introduce different learning objectives associated with different relations.

Features to Features (A_TA_T/A_nA_n/$A_{event}A_{event}$): These objective functions place the same type of features with similar context together in the embedding space. The similarity in context refers to similar word or sentence embedding. If the topics, named entities, or events have embedding value within a certain threshold, they are considered similar.

Data Sample to Data Sample (x^Dx^V/x^Dx^D/x^Vx^V): Currently, in this work, data samples refer to videos (x^V) and documents (x^D). This objective function maximizes the similarity of the data samples pair (x_i, x_j) with the motivation that data objects discussing about *similar events between similar entities on similar topics* should be semantically similar. We select the positive pairs for respective objective function in following ways:

1. $xx - topics$. If data samples are annotated and *topic* annotations are available, we pair the data samples which belong to the same group.
2. $xx - events - entities$. For named entities and events, we can consider them together to select the pairs since entities separately does not contribute towards two document or videos being similar. So, data samples discussing about *common events* between a threshold number of *common entities* belong to the same pair.
3. $xx - label$. Two data samples with a user-provided positive similarity between them should have similar embedding.
4. $xx - embedding$. If initially two data samples have text embedding representation ($a_{text} \subset A_{text}$) within a certain threshold, they are placed in the same embedding space. The threshold is determined empirically for each application domain.

Data Samples to Features (xA_T/xA_{event}/xA_n): This objective tends to maximize the similarity of data samples to their features. For example, if we only consider topics as the only feature, we want to place the data samples belonging to a certain topic closer to that topic. In the embedding space, the data sample vectors should be closer to the topic embedding vectors.

Joint Objective Function. Finally, we combine the loss functions of all the learning objectives to define our joint embedding loss function. The set of possible learning objectives, $O = \{A_TA_T, A_nA_n, A_{event}A_{event}, x^Dx^V, x^Dx^D, x^Vx^V, xA_T, xA_{event}, xA_n\}$ is expandable as we consider more features in future. We experimented with different combinations of these objective functions. So, the combined loss function is

$$L_{total} = \sum_{i \in O_s, O_s \subset O} \lambda_i L_i \tag{3}$$

Here, O_s refers to the selected objective functions and λ_i refers to the weight applied to the objective function i. For our experiments we set the value of λ_i to 1 for all the objectives.

Initial Representation of Graph Elements: For all objects in the graph, the initial representation is chosen from different representations for text. We experimented with different initial representations for text, and then use a hidden layer to map the initial representations in the joint embedding space. This linear layer filters out the important features from the initial representation for the joint embedding. For an initial representation, t of a text, the hidden layer computes it's embedding e as follow.

$$e = f(Wt + b) \tag{4}$$

Reasoning Over the Data Information Network

Our end goal is to use the vector representations of the data samples and features to extract all relevant data samples from the database given a particular data sample. The relevance can be defined directly over the embedding space, by comparing the similarity of the vectors representing respective data samples. We can calculate a relevance score by taking the graph structure into account, by exploiting interdependencies among features.

Weak Supervised Baseline. To use the information graph that we built, we use the information from graph directly without any learning. This is achieved by counting the paths from one data sample to a given data sample or a given feature. Let $P(a, b)$ define the set of paths from given data sample a to another data sample b. Each path is associated with a weight w. Weights are assigned to each path considering the features that exists in the path. Initially, we can consider all weights to 1. But in reality some degrees of similarity have higher precedence. For example, a user defined similarity should have the highest priority. We hypothesize the following feature order to assign the weights based on the priority assigned by domain experts -

$$A_{sim} > A_{text} > A_T > A_{event}, A_n | A_u \tag{5}$$

So a path with $E_{xA_{sim}}$ has a higher weight than a path with $E_{xA_{text}}$. Given the graph G, edges connect a data sample to its features, and then features to other data samples having the same features. The relevance score between a and b is then defined as:

$$Rel(a, b) = \frac{\sum_{i \in P(a,b)} w_i}{\sum_{b \in B} \sum_{i \in P(a,b)} w_i} \tag{6}$$

where B is the set of all the data samples in the database.

In case we need to find all the data samples given a feature, we can retrieve them directly from the graph structure. Let $N_p(f, b)$ define the number of paths from given feature f to a data sample b. The relevance score between f and b is then defined as:

$$Rel(f, b) = I * N_p(f, b) \tag{7}$$

where I is an indicator variable. I = 0, if there is no path between f and b, otherwise I = 1.

Similarity Based Score. Given a data sample, or a feature a and their embedding e_a the relevance score with other data sample b with embedding e_b is:

$$Rel(a, b) = sim(e_a, e_b) \tag{8}$$

where $sim()$ is the cosine distance between the vectors representing the given data sample, or feature and the other data sample.

Formalization of Novelties

Data Shift in Multimodal Data Retrieval Task

Existing works in multi-modal information retrieval defines it in different ways. In *supervised setting*, following our previous notations, Let the training data be $\mathcal{D}_{tr} = \{(\mathbf{x}_j^i, y_j^i)\}_{j=1}^{n_i}$, where n_i is the number of samples in i-th modality, $\mathbf{x}_j \in X$ is a training example following the training distribution $P_{tr}(\mathbf{x})$. $y_j \in Y_{tr}$ is the corresponding class label of $\mathbf{x}_j$ and Y_{tr} is the set of all class labels that appear in $\mathcal{D}_{tr}$. Each modality have their own training distribution $P_{tr}(\mathbf{x}^i)$, but for simplicity purpose, we are going to denote training distribution as only $P_{tr}(\mathbf{x})$. For *unsupervised setting*, y_j is absent in $\mathcal{D}_{tr}$. In our *weakly supervised setting*, class labels are still absent, but the extracted features act as weak labels and amplifies similarity signal among data samples through the network structure. The retrieval task refers to estimating probability of a data sample being relevant to a query given the data sample and a query sample, $P(R|x_p, x_q)$, where $x_p, x_q \in X$, and R is the corresponding relevance label.

Following the discussion in (Liu et al. 2021) and (Moreno-Torres et al. 2012), we define the three main types of data shift that can happen during testing for Multimodal Data Retrieval Task.

Covariate shift refers to the distribution change of the input variable x between training and test phases, i.e., $P_{tr}(R|x_p, x_q) = P_{te}(R|x_p, x_q)$ and $P_{tr}(\mathbf{x}) \neq P_{te}(\mathbf{x})$. This can refer to change in application domain while still dealing with the same modalities in P_{tr}. This also can occur if user starts to phrase their queries differently.

Prior probability shift refers to the distribution change of the class variable y, or the relevance variable R, or the weak feature variables $\mathcal{A}$, i.e., in our framework, $P_{tr}(x_p|R, x_q) = P_{te}(x_p|R, x_q)$ and ($P_{tr}(R) \neq P_{te}(R)$ or $P_{tr}(A) \neq P_{te}(A)$). This includes not having extracted weak features from a data sample during testing.

Concept drift refers to the change in the posterior probability distribution between training and test phases, i.e., $P_{tr}(R|x_p, x_q) \neq P_{te}(R|x_p, x_q)$ and $P_{tr}(\mathbf{x}) = P_{te}(\mathbf{x})$. This can be a temporal effect or user requirement change over time.

Besides the three types of data shifts, multimodal retrieval faces one other type of change during testing, i.e., data samples that do not belong to the modalities that the framework can handle. These are *novelty* or *novel instances*. This is closely related to covariate shift and some framework may handle novel instances as part of a known class.

Novelty Detection

We use the *data information network* to detect the changes between pre-novelty and post-novelty environments. During inference with a novelty introduction, after the translation and weak feature extraction, we have the post-novelty graph. We can use existing node discovery techniques to detect change from the training time information network. In case of a *novel modality*, our proposed framework would either identify the new weak features (different from training time), or tackle the new modality as part of the training distribution. In the later case, we may see a decline in the model performance.

Definition 1 ***(Novel Instance).*** A test instance x is novel if $G(V_{P_{tr}+\mathbf{x}}, E)$ is different from $G(V_{P_{tr}}, E)$. This can be explained as having a knowledge base for the weak features during training time ($\mathcal{A}_{tr}$). If during inference, we discover weak features that are absent in $\mathcal{A}_{tr}$, we consider the instance as novel.

Novelty Characterization and Response

Definition 2 ***(Characterization of Novelty).*** Characterization of Novelty is the description of the novelty, according to which appropriate course of actions are taken to respond to the novelty. We characterize novelty based on the data shift variations -

1. Covariate shift with change in application domain with the modalities for which translation module is available *(covar-1)*.
2. Prior probability shift with novel weak features *(prior-1)*.
3. Prior probability shift with no weak features *(prior-2)*.
4. Prior probability shift with novel relevance label *(prior-3)*.
5. Temporal concept drift with previously relevant data being non-relevant *(concept-1)*.
6. Covariate shift with new modality introduction *(covar-2)*.

Novelty Response. For a generalized response, we propose to build a pre-trained retrieval model from WeS-JEm to deal with the out-of-distribution (OOD) inputs. We adopt a three level training strategy for our model. For the first stage, we pre-train the translation module (DVC) with video captioning and video retrieval task, following the strategy in (Luo et al. 2020) and (Xu et al. 2016). JEDDi-Net is trained on both the ActivityNet Captions (Krishna et al. 2017) dataset and MSR-VTT (Xu et al. 2016) dataset. MSR-VTT has open domain video clips, and each clip has 20 captioning sentences labeled by human. For both cases, we used the SPN module from (Xu et al. 2016) trained with the temporal annotation of ground truth segments in the ActivityNet Captions dataset with Sports-1M pretrained C3D weight initialized in (Tran et al. 2015). For the text encoder, we choose between the pre-trained BERT (Devlin et al. 2018) Base uncased model and the pre-trained skip-thought model (Kiros et al. 2015). Next, we extract the weak features from the dense captions and document text using topic and event extraction models such as (Angelov 2020; Wadden et al. 2019).

Finally, we train our weakly supervised model using the joint objective loss function.

During inference, the translation module is able to generate captions for OOD inputs, which includes input from new modalities as in *(covar-2)*, i.e., image or LIDAR. Both image and LIDAR modality can be handled as a variant of the video translation module. Both BERT and skip-thought can generate text embedding for any textual input. This takes care of the novel weak features *(prior-1)*. Finally, the linear layers in WeS-JEm would be able to map the OOD inputs into the pre-trained joint embedding space. The final similarity score between data samples in the embedding space would produce the relevance between new samples.

With the encounter of a *(prior-2)* novelty, if the system is allowed to $learn$, in proposed joint embedding model, the set of selected learning objectives becomes, $O_s = \{x^D x^V, x^D x^D, x^V x^V\}$ where only the $xx - embedding$ objective function is considered, as each data sample would include an initial representation from the textual descriptions.

Finally, when encountered with any of the last of the three types of novelties, an information retrieval system needs to re-learn. In case of a novel modality introduction, the long-time response is to $learn$ or $gather$ a new translation method. In our model, we include this as a **Relevance Feedback** module. The new relevance label provided by a human annotator holds more importance than previous relevance labels. To make a distinction between this newly provided label and old label between (x_p, x_q), during re-training, we encode this by assigning more priority to the new similarity label, $A^{new}_{sim} > A^{old}_{sim}$.

Experiments

The first set of experiment compares the embedding approach for single modality information retrieval (text $\rightarrow$ text) when there is only topics are available as feature. We call this model **D**ata with **T**opics to **D**ata **Vec**tors **(DT2DVec)**. We experimented with the following representations for caption document of videos, documents, and topics -

- random initialization of the document,
- average of pre-trained GloVe word embedding (300d) of filtered tokens from the document,
- Skip-Thought (Kiros et al. 2015) for capturing the global context of the document,
- BERT-Base uncased model for generation of text representation T from the token sequence t of the document.

These initial representations are mapped into a hidden layer to map them into the joint embedding space as part of the retrieval model, as shown in equation 4.

Negative Sampling. As in (Mikolov et al. 2013), we used negative sampling to train the model. Our goal is to minimize the similarity of the target object, o_i and samples drawn from the noise distribution, $\mathcal{P}_n(o_i)$ with k negative samples for each data sample. DT2DVec investigated with a number of choices for $\mathcal{P}_n(o_i)$.

Table 1: Performance Comparison Results of DT2DVec

	LSA	LDA	DT2DVec
Inter-similarity	0.76	0.66	0.61
Intra-similarity	0.45	0.28	0.047

1. Following (Mikolov et al. 2013), we pick $\mathcal{P}_n(o_i)$ from the uniform distribution of the objects in the dataset. Objects in the dataset consist of the documents from video captions, text, and topics of the texts and videos. The uniform distribution of the objects in the dataset d is $U(d)$ raised to the 3/4rd power with $U(d)$ being the frequency of objects in the respective dataset. Documents are different from words, as words can appear multiple times in a document where often documents do not appear multiple times in a dataset.
2. Given, we have the annotated topics for documents, we consider the noise distribution of each topic t, $\mathcal{P}_n(t)$ from the document samples of other topics. Any data sample that is not annotated with topic t belongs to $\mathcal{P}_n(t)$.

For batch training, DT2DVec adopted the following approaches: **(2a)** Let us assume there are p number of positive pairs in a batch, and the set of topics of these pairs is POS_T. If the mode of POS_T is topic t, then we pick negative examples from $\mathcal{P}_n(t)$ for this batch. The intuition behind the approach is the closer graph objects in the embedding space should have similar distribution; **(2b)** We select variable number of negative examples for each batch. Negative examples are selected from $\mathcal{P}_n(t)$ of each topic in the positive pairs, weighted by the number of positive pairs from each topic.

Dataset and Experimental Setup. For the performance evaluation of DT2DVec, we used the 20 Newsgroups dataset (Lang 1995) with twenty annotated topics. We compared DT2DVec with two baseline topic modeling approaches - LSA (Deerwester et al. 1990) and LDA (Blei, Ng, and Jordan 2003). For evaluation, we split each document into two parts, and test if **(1)** the topics of the first half are similar to topics of the second half (inter-similarity); **(2)** halves of different documents are mostly dissimilar (intra-similarity). We use cosine similarity to measure the difference between the two vectors of half document topics. For inter-similarity, higher similarity score is better. For intra-similarity, the lower the similarity, the better the vectors are. We present the result for the random initialization of initial representation of text in Table 1. We used Pytorch (Paszke et al. 2017) with binary cross entropy to train the unsupervised retrieval model. Mini-batch gradient descend was used for optimization with SGD (Duda 2019).

Results. **(DT2DVec - Rand)** performed significantly better in recognizing the dissimilar documents than the baseline models, LSA and LDA. There was around 43% improvement in similarity score for dissimilar documents whereas LSA outperformed our method by 10% for simiar documents. Figure 3 shows the embedding space of the documents using the DT2DVec model.

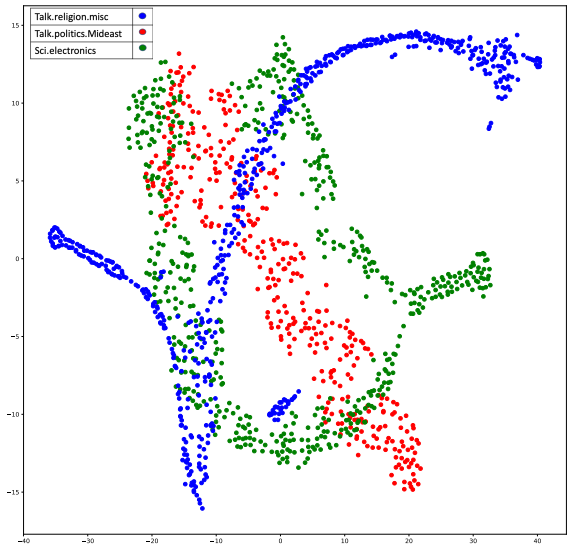

Figure 3: t-SNE Embedding of Documents using DT2DVec with Random Initial Representation of Text

Multi-modal Retrieval. For the second set of experiments, we are using the cross modal datasets described in FemmIR (Solaiman and Bhargava 2021), Youcook2 (Zhou, Xu, and Corso 2018), and MSR-VTT (Xu et al. 2016). We evaluate only text-based video retrieval task on Youcook2 and MSR-VTT. For the text-based video retrieval task, we use the captions as the input text queries to find the corresponding video clips. We compare the performance of our model with the following models - UniVL (Luo et al. 2020), SDML (Hu et al. 2019), and CMPM+CMPC (Zhang and Lu 2018). We will include the results in future work.

Conclusion

This paper proposed a *weakly supervised open world learning* framework for multi-modal information retrieval. Our methods involve no human annotation, show promising performance compared to unsupervised approaches, and formalize novelties encountered during testing. In the future, we would test our novelty characterization, detection and adaptation framework with different datasets. We would also include different modalities in our framework, and would test the capability of the framework for domain generalization.

Acknowledgments

This research is supported, in part, by the Defense Advanced Research Projects Agency (DARPA) and the Air Force Research Laboratory (AFRL) under the contract number W911NF2020003. 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 DARPA, AFRL, or the U.S. Government. We thank our team members on this project for all the discussions to develop this paper. Some of the ideas in this paper are based on our learning from the SAIL-ON meetings.

References

Andrew, G.; Arora, R.; Bilmes, J.; and Livescu, K. 2013. Deep canonical correlation analysis. In *International conference on machine learning*, 1247–1255. PMLR.

Angelov, D. 2020. Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*.

Bellet, A.; Habrard, A.; and Sebban, M. 2013. A survey on metric learning for feature vectors and structured data. *arXiv preprint arXiv:1306.6709*.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, 3(null): 993–1022.

Boult, T.; Grabowicz, P.; Prijatelj, D.; Stern, R.; Holder, L.; Alspector, J.; Jafarzadeh, M.; Ahmad, T.; Dhamija, A.; Li, C.; et al. 2021. Towards a Unifying Framework for Formal Theories of Novelty. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 15047–15052.

Deerwester, S.; Dumais, S. T.; Furnas, G. W.; Landauer, T. K.; and Harshman, R. 1990. Indexing by latent semantic analysis. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE*, 41(6): 391–407.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Duda, J. 2019. SGD momentum optimizer with step estimation by online parabola model. arXiv:1907.07063.

Faghri, F.; Fleet, D. J.; Kiros, J. R.; and Fidler, S. 2017. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*.

Feng, F.; Wang, X.; and Li, R. 2014. Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the 22nd ACM international conference on Multimedia*, 7–16.

Frome, A.; Corrado, G.; Shlens, J.; et al. 2013. DeViSE: A Deep Visual-Semantic Embedding Model. In *Advances in Neural Information Processing Systems*, volume 26.

Hu, P.; Zhen, L.; Peng, D.; and Liu, P. 2019. Scalable Deep Multimodal Learning for Cross-Modal Retrieval. In *Proceedings of the 42Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, 635–644. New York, NY, USA: ACM. ISBN 978-1-4503-6172-9.

Ji, S.; Xu, W.; Yang, M.; and Yu, K. 2012. 3D convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1): 221–231.

Kan, M.; Shan, S.; and Chen, X. 2016. Multi-view deep network for cross-view classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4847–4855.

Kiros, R.; Zhu, Y.; Salakhutdinov, R. R.; Zemel, R.; Urtasun, R.; Torralba, A.; and Fidler, S. 2015. Skip-thought vectors. In *Advances in neural information processing systems*, 3294–3302.

Krishna, R.; Hata, K.; Ren, F.; Fei-Fei, L.; and Carlos Niebles, J. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, 706–715.

Lang, K. 1995. Newsweeder: Learning to filter netnews. In *Machine Learning Proceedings 1995*, 331–339. Elsevier.

Langley, P. 2020. Open-world learning for radically autonomous agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 13539–13543.

Liong, V. E.; Lu, J.; Tan, Y.-P.; and Zhou, J. 2016. Deep coupled metric learning for cross-modal matching. *IEEE Transactions on Multimedia*, 19(6): 1234–1244.

Liu, B.; Robertson, E.; Grigsby, S.; and Mazumder, S. 2021. Self-Initiated Open World Learning for Autonomous AI Agents. arXiv:2110.11385.

Luo, H.; Ji, L.; Shi, B.; Huang, H.; Duan, N.; Li, T.; Li, J.; Bharti, T.; and Zhou, M. 2020. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; and Dean, J. 2013. Distributed Representations of Words and Phrases and their Compositionality. arXiv:1310.4546.

Mitchell, T.; Cohen, W.; Hruschka, E.; Talukdar, P.; Betteridge, J.; Carlson, A.; Dalvi, B.; Gardner, M.; Kisiel, B.; Krishnamurthy, J.; Lao, N.; Mazaitis, K.; Mohamed, T.; Nakashole, N.; Platanios, E.; Ritter, A.; Samadi, M.; Settles, B.; Wang, R.; Wijaya, D.; Gupta, A.; Chen, X.; Saparov, A.; Greaves, M.; and Welling, J. 2015. Never-Ending Learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15)*.

Moreno-Torres, J. G.; Raeder, T.; Alaiz-Rodríguez, R.; Chawla, N. V.; and Herrera, F. 2012. A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1): 521–530.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in PyTorch.

Peng, Y.; Qi, J.; Huang, X.; and Yuan, Y. 2017. CCL: Cross-modal correlation learning with multigrained fusion by hierarchical network. *IEEE Transactions on Multimedia*, 20(2): 405–420.

Rasiwasia, N.; Costa Pereira, J.; Coviello, E.; Doyle, G.; Lanckriet, G. R.; Levy, R.; and Vasconcelos, N. 2010. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM international conference on Multimedia*, 251–260.

Sah, S.; Gopalakrishnan, S.; and Ptucha, R. 2020. Aligned attention for common multimodal embeddings. *Journal of Electronic Imaging*, 29: 023013 – 023013.

Solaiman, K.; and Bhargava, B. 2021. Feature Centric Multi-modal Information Retrieval in Open World Environment (FeMMIR). Unpublished.

Song, Y.; and Soleymani, M. 2019. Polysemous visual-semantic embedding for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1979–1988.

Tanon, T. P.; Weikum, G.; and Suchanek, F. 2020. Yago 4: A reason-able knowledge base. In *European Semantic Web Conference*, 583–596. Springer.

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 4489–4497.

Wadden, D.; Wennberg, U.; Luan, Y.; and Hajishirzi, H. 2019. Entity, relation, and event extraction with contextualized span representations. *arXiv preprint arXiv:1909.03546*.

Wang, B.; Yang, Y.; Xu, X.; Hanjalic, A.; and Shen, H. T. 2017. Adversarial cross-modal retrieval. In *Proceedings of the 25th ACM international conference on Multimedia*, 154–162.

Wang, W.; Arora, R.; Livescu, K.; and Bilmes, J. 2015. On deep multi-view representation learning. In *International conference on machine learning*, 1083–1092. PMLR.

Wang, X.; Hu, P.; Zhen, L.; and Peng, D. 2021. DRSL: Deep Relational Similarity Learning for Cross-modal Retrieval. *Inf. Sci.*, 546: 298–311.

Wei, J.; Xu, X.; Yang, Y.; Ji, Y.; Wang, Z.; and Shen, H. T. 2020. Universal weighting metric learning for cross-modal matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13005–13014.

Xu, H.; Li, B.; Ramanishka, V.; Sigal, L.; and Saenko, K. 2018. Joint Event Detection and Description in Continuous Video Streams. *CoRR*, abs/1802.10250.

Xu, J.; Mei, T.; Yao, T.; and Rui, Y. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5288–5296.

Xu, X.; He, L.; Lu, H.; Gao, L.; and Ji, Y. 2019. Deep adversarial metric learning for cross-modal retrieval. *World Wide Web*, 22(2): 657–672.

Zhai, X.; Peng, Y.; and Xiao, J. 2013. Learning cross-media joint representation with sparse and semisupervised regularization. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(6): 965–978.

Zhang, L.; Ma, B.; Li, G.; Huang, Q.; and Tian, Q. 2017. Generalized semi-supervised and structured subspace learning for cross-modal retrieval. *IEEE Transactions on Multimedia*, 20(1): 128–141.

Zhang, Y.; and Lu, H. 2018. Deep cross-modal projection learning for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 686–701.

Zhou, L.; Xu, C.; and Corso, J. J. 2018. Towards automatic learning of procedures from web instructional videos. In *Thirty-Second AAAI Conference on Artificial Intelligence*.