TransFusion: Multi-Modal Fusion for Video Tag Inference via Translation-based Knowledge Embedding

Di Jin
University of Michigan
dijin@umich.edu

Zhongang Qi
Applied Research Center (ARC), PCG, Tencent
zhongangqi@tencent.com

Yingmin Luo
Applied Research Center (ARC), PCG, Tencent
yingminluo@tencent.com

Ying Shan
Applied Research Center (ARC), PCG, Tencent
yingsshan@tencent.com

ABSTRACT

Tag inference is an important task in the business of video platforms with wide applications such as recommendation, interpretation, and more. Existing works are mainly based on extracting video information from multiple modalities such as frames or music, and then infer tags through classification or object detection. This, however, does not apply to inferring generic tags or taxonomy that are less relevant to video contents, such as video originality or its broader category, which are important in practice. In this paper, we claim that these generic tags can be modeled through the semantic relations between videos and tags, and can be utilized simultaneously with the multi-modal features to achieve better video tagging. We propose TransFusion, an end-to-end supervised learning framework that fuses multi-modal embeddings (*e.g.*, vision, audio, texts, etc.) with the knowledge embedding to derive the video representation. To infer the diverse tags following heterogeneous relations, TransFusion adopts a dual attentive approach to learn both the modality importance in fusion and relation importance in inference. Besides, it is general enough and can be used with the existing translation-based knowledge embedding approaches. Extensive experiments show that TransFusion outperforms the baseline methods with lowered mean rank and at least 9.59% improvement in HITS@10 on the real-world video knowledge graph.

CCS CONCEPTS

• **Information systems** → **Multimedia and multimodal retrieval**; *Data mining*; • **Computing methodologies** → **Knowledge representation and reasoning**; **Machine learning**.

KEYWORDS

Multi-modal fusion, Knowledge graph embedding, Video tagging

ACM Reference Format:
Di Jin, Zhongang Qi, Yingmin Luo, and Ying Shan. 2021. TransFusion: Multi-Modal Fusion for Video Tag Inference via Translation-based Knowledge Embedding. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3474085.3481535

MM '21, October 20–24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3481535

1 INTRODUCTION

As a rapidly developing form of media, video platforms such as Youtube and Tiktok have become one of the most important ways of information acquiring, entertaining, and socializing for people in the world. Long videos and short videos have thus drawn much attention in both industry and academia, and one most fundamental task in this field is tag inference. In tag inference, brief descriptions about the videos are summarized as a set of tags, and the goal is to assign the most relevant ones to a specific video. Tag inference has wide application such as searching, personalization, and more.

Existing approaches for video tagging in CV (Computer Vision) are mainly based on multi-label classification, object detection, or label propagation [7, 13, 26]. These approaches generally focus on the content of the video to decide *what* to tag, such as specific objects, scenes, actions, or locations. For example in Figure 1, the tag "Cats" and "Pets" are assigned to video 1 that describes the casual life of house cats. However, as a form of complex media, real-world videos are often associated with tags related to the high-level knowledge such as the meta-info or categorization. These tags and their inter-relationship are critical for downstream tasks such as recommendation or personalization. In the previous example, video 1 is tagged "Catlover" to characterize its author, which cannot be inferred using the CV techniques as it does not appear in the video. Besides, the relationship between "Pets" and "Nature" are also important to infer the similarity between the well-tagged (video 1) and weakly-tagged (video 3) videos, but such relation cannot be established through the above mentioned approaches. The key issue is that simply inferring tags through the single relation "content appearing in the video" cannot handle tags that are beyond video contents and assigned with different semantic meanings.

To model the large-scale multi-relational data, knowledge graphs (KG) or knowledge bases such as DBpedia [16] and YAGO [22] are successfully applied in the industry. Knowledge embedding has been widely studied due to its superiority representing the plausibility between entities under heterogeneous relations, and has achieved promising results in knowledge-drive tasks such as link prediction [12, 39] and knowledge alignment [4]. The main idea is to encode entities as low-dimensional vectors and the relation is encoded in the form of vector algebra by following the knowledge graph connectivity, *e.g.*, the translation vector from a header to the tail [2]. However, existing approaches mainly focus on the single

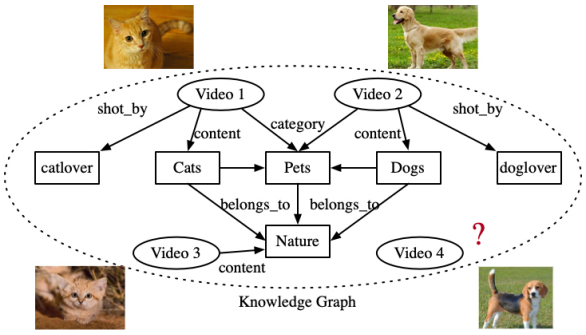


Figure 1: Knowledge embeddings based on graph connectivity cannot infer tags for videos with limited connections or new videos (*e.g.*, video 3 and 4). On the contrary, this task can be better addressed by incorporating information from multi-modality (*e.g.*, visual images). In this example, video 3 and 4 can be reasonably tagged based on the similarity of key frames.

textual modality and embed entities based on the graph connectivity. As videos contain rich information across multiple modalities such as vision, sound, and texts, embeddings based on graph connectivity only are insufficient for tag inference. For example in the knowledge graph depicted in Figure 1, while both video 1 and 2 are well tagged, it is hard to infer reasonable tags for video 3 as no videos are connected to "Nature". It is even harder for new videos such as video 4 since it has not connected to entities in the established KG. On the other hand, this task could be better addressed if the associated key frames (visual modality) are leveraged. In the above example, both video 3 and 4 can be reasonably tagged by looking into their visually similar videos, *i.e.*, video 1 and 2, respectively.

In this paper, we claim that the multi-modality features for videos, and the semantic relation between videos and tags can be utilized simultaneously to achieve a better video tagging. We propose TransFusion, a general embedding fusion approach for knowledge graphs that integrates the pretrained video embeddings from multiple modalities. Specifically, we propose a partially-trainable model to fuse both the learnable entity KG embeddings and the pretrained video embeddings from multiple modalities. By following the predefined scoring function, the fused video embeddings are then used to derive embeddings for semantic relations, which are further used to infer tags as the regular link prediction task. TransFusion leverages a dual attention mechanism to look into both the importance of multiple modalities in representing the videos, as well as the importance of multiple relations in tag inference. "TransFusion" gets the name from the use of attentive fusion for video modalities. It also works seamlessly with a broad class of translation-based embedding models for knowledge graph, pioneered and represented by the TransE algorithm. We hope this work could benefit researchers and practitioners addressing similar tasks in the video media industry, and motivates its usage due to its generality to be applied to a broad class of existing works in the field, which is desiring property in practice. Our contributions are:

- **New problem formulation** We formulate video tag inference through heterogeneous relations as a knowledge embedding problem, which is the first study in the field with such formulation.
- **Novel framework** We propose a novel dual attentive approach that learns both the modality importance in fusion and relation importance in tag inference. Besides, the framework is generally applicable to translation-based graph embedding approaches.
- **Extensive experiments** We conduct extensive experiments with TransFusion and show that it achieves state-of-the-art performance on the tag inference task with promising generality and scalability.

2 RELATED WORK

Knowledge Graph (KG) Embeddings Most works in the literature represent entities and relations into the low-dimensional vector space $\mathbb{R}^d$ [2, 19, 39]. Some works use other space such as tensor space [30].The existing works can be categorized into transnational distance models or semantic matching models. Distance-based models measure the plausibility between $\mathbf{h}$ and $\mathbf{t}$ through translation based on distance of the entity representation. For example, the forerunner work TransE [2] proposes the additive translation as $f_r(h,t) = \| \mathbf{h} + \mathbf{r} - \mathbf{t} \|_{L_1/L_2}$, and achieves promising results on 1-to-1 relations. Based on TransE, a series of extensions such as TransH [39], TransD [12], etc. were proposed to better handle multi-way relations. Semantic-matching-based methods propose to match entities and relations through linear or bilinear transformation. For example, SME [1] proposes both linear and bilinear matching blocks. DistMult [41] proposes a simplified bilinear transformation as $f_r(h,t) = \mathbf{h}^T \text{diag}(\mathbf{M}_r)\mathbf{t}$. Unlike the above KG embedding methods, TransFusion incorporates video embeddings from 3 different modalities (vision, audio and text) to represent the entities in a KG, and could be generally applied to existing distance-based models.

Multi-modality Embeddings & Fusion Early works in video visual representation learning are mainly based on manually-crafted features [15, 36]. Recently, spatial-temporal features have been widely used with convolutional neurual networks (CNNs) to achieve advancement, Some work propose to learn the spatial and temporal features separately, for example, [7] adopts the two-stream networks, [9, 14, 37] learn spatial features with 2D-CNNs and learn temporal features with specially designed modules. There are also works learning them jointly through 3D-CNNs [3, 34]. As for extracting auditory representation, [23] proposes to use MFCC features, and [11] learn audio semantic representation with spectrogram features feeding to a VGG network. To derive sentence-level semantic features, early work adopts manually-crafted kernels or NLP tools [28]. As deep learning develops, recent works proposes to use end-to-end models with raw sentences and pre-trained word representations learned by Skip-gram and Continuous Bag-of-Words as inputs [24], and learn through CNN [44] or RNN [45]. Most recently, [6, 42] propose fine-tuning base representation models with transformers and achieve the state of the art performance for several NLP tasks. In terms of fusion, DeepCrossing [29] proposes a deep neural network model that automatically combines different forms of textual features such as tri-letter grams to produce superior results in user click prediction. In the fusion process, the attention mechanisms [21, 35] are generally adopted by the deep models where the goal is to improve the performance by highlighting valuable latent features. Relevant techniques such as gated

Table 1: Data statistics and properties. For "Modalities", "V": vision, "A": audio, and "T": title.

Data	# Entities	# Edges	# Relation_types	Modalities
Company-200K	32 527	120 001	20	{V, A, T}
Company-300K	53 150	330 001	20	{V, A, T}
Company	1 000 014	9 846 740	86	{V, A, T}
FB15K	14 541	292 582	237	{V}

fusion networks [17, 27] or graph-based fusion [43] are proposed to handle CV or NLP tasks. For KG, there are works with specific targets, such as using the auxiliary information from the texts [38] for fact reasoning, or from images [40] for classification. However, these works either align the single external modality, or focus on image classification. To the best of our knowledge, our work is the first that fuses embeddings from visual, auditory and textual modalities to conduct video tag inference in multi-relational data.

3 DATA

In this section, we introduce the datasets, discuss how we processed and cleaned them, and give descriptive analyses that motivate our methodology in Section 4. We provide data statistics in Table 1.

3.1 Video KG data

Our new video corpus is collected from the content library of a major consumer video platform, which consists of over 1 million videos during 3 weeks in June, 2020 with over 200 attributes. In the JSON format raw data, each record describes a video with attributes that are either AI generated or manually labeled, such as "`tag`", "`has_person`", "`originality`", etc. We filtered out irrelevant attributes such as those describe the visual quality (*e.g.*, resolution, picture scales) and formed the corpus `Company`s.

3.1.1 Videos. Given a video, we consider 3 modalities: $\mathcal{V}$: visual frames (images), $\mathcal{A}$: background music (BGM) and $\mathcal{T}$: textual title. To get embeddings from 3 modalities, we leverage a dataset in larger scale that consists of over 7 million videos in 380 classes from the same business. The 380 tags include visual objects such as "Cats" and "Vehicles", as well as general concepts or taxonomy such as "Science & technology". This labeled dataset is collected in 2019 and does not overlap with `Company`. We train a series of classification models for the 3 modalities, and extract features inputting to the last fully-connected output layer as the embedding of that modality. We detail the discussion in Section 7.1 of the supplementary material.

3.1.2 Relation & Tags. To facilitate the inference task, we construct the knowledge graph by transferring the attribute values from the input data into semantic relations and tags. We defined a mapper to parse the attributes of each video and generate their semantic tags. For example, given a video record with the attribute-value pair "`has_characters: male`", we use "`has_characters`" as the relation and gender "`Male`" as the tag. The same relation can also be used to describe the other videos with different tags such as "`Female`" or "`Neutral`". Another attribute-value pair "`src_flag: original`" indicates the originality of the video source, and can be transferred into relation"`src_flag`" and tags "`Originality/Derivation`". Based on the characteristics of possible tags related to the videos and general notions, we have the following categorization for the tags.

- *Modality-related* This type of tags are based on information from a specific modality such as an object in the video frames, the music style of the BGM, or the keyword in the video title. An example is the above "`{Male/Female/Neutral}`" which can be inferred through the relation "`has_characters`". Another example is the names of actors/celebrities appearing in the video following "`has_actor`".
- *Meta-info* This type contains flexible tags generated by manual annotation to describe the video, a general notion or the video taxonomy. Examples in this category include the above "`{Originality / Derivation}`" following the "`src_flag`", and "`{Frontpage / Regularpage}`" indicating the popularity of the video following "`if_frontpage`".

Tags are created carefully to avoid ambiguity in the constructed knowledge graph, and to ensure each tag is unique and specific to one relation. Also, we manually connect these tags based on their semantic relations according to the public knowledge base DBpedia. For example, actors participating in different videos are connected with the relation "`befriends`" to indicate their personal relationship. Therefore, the resultant knowledge graph `Company` contains entities that indicate either a video or a tag, as well as rich "video-tag" and "tag-tag" relations as shown in Figure 1.

3.2 Public KG data

Due to the lack of public video-tag data, we leverage the benchmark knowledge graph `Freebase` (`FB15K237`) in the KG literature and the associated public images to show the effectiveness of TransFusion in embedding fusion. For the entities in `Freebase`, we adopt the associated visual information from ImageGraph [25], which is also leverged in MMKG (Multimodal Knowledge Graphs) [20] to perform entity alignment between pairs of KGs. Each entity in `Freebase` has up to 25 relevant images scaled into the same size. To derive the visual embeddings, we pretrained the classification model following ResNet-101 [10] on the public ImageNet dataset [5] that consists of 1.28 million training images in 1000 classes. For each entity, we randomly sampled one representative image as the input to the model, and extracted the dense feature immediately prior to the last fully-connected output layer as the visual embedding.

4 METHOD

While the semantic/taxonomic tags of a video can be inferred through KG embeddings, many tags are assigned based on the specific modality (*e.g.*, frames or BGM), and such information cannot be captured by connectivity-based KG embedding approaches only. Thus, in the constructed KG that reflects multiple relationships between videos and tags, the fused embeddings of videos should consist of **(1)** the knowledge segment that facilitates tag inference through semantic relations, and **(2)** the supplementary segment containting pretrained embeddings from the considered modalities such as vision or audio. In addition, as a tag can be relevant to more than one modalities through multiple semantic relations (*e.g.*, the tag "`music_video`" can be assigned based on the audio or vision-related relation), highlighting the impact of each modality and relation in deriving the fused KG representation is central in designing TransFusion. Specifically, we consider the following components: **(C1)** Modeling the importance of each modality in the

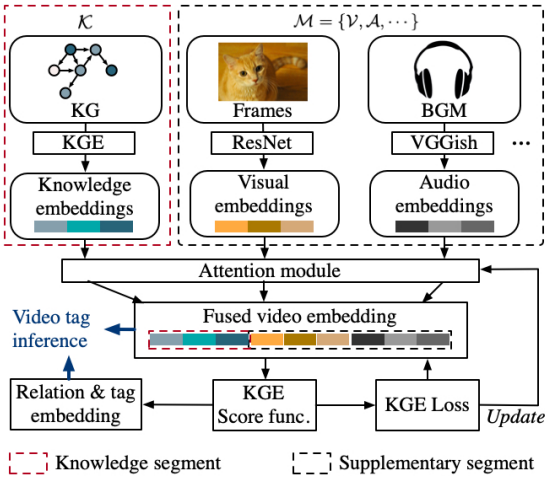


Figure 2: Workflow. TransFusion concatenates the knowledge (red dashed box) and supplementary multi-modal embeddings (black dashed box) as the fused video represntation and controls their impacts with the shared attention module. Then, TransFusion uses the predefined KG scoring function to jointly update the attention values and the embedding of tags, relations and the knowledge segment of videos (green rectangles) based on KG connectivity.

fused video embedding (Section 4.2), **(C2)** Highlighting important relations (Section 4.3), and **(C3)** Tag inference as the link prediction task via a given translation-based KG scoring function and loss (Section 4.4). Next, we describe each component in detail. The workflow is depicted in Figure 2.

4.1 Preliminaries

First, we briefly overview the important notations used in this paper, (symbols are listed in Table 2). As mentioned in Section 3, our constructed KG contains the "video-tag" and "tag-tag" semantic relations. Thus, in the tuple format representation (h, r, t), h indicates either a video (h_{vid}) or tag (h_{tag}) and t indicates the tag only. An important design target is to generally facilitate the broad class of translation-based knowledge embedding approaches represented by TransE for modality fusion in tag inference. Therefore, we assume a predefined scoring function f_r is given, such as $f_r(\mathbf{h}, \mathbf{t}) = ||\mathbf{h} + \mathbf{r} - \mathbf{t}||$ that describes the plausibility between entities. In addition, f_r is also the rule to infer tags. Next we specify the notations used to represent the fused video embedding $\mathbf{h}_{\text{vid}}$. At a high level, TransFusion fuses video embeddings through concatenation (shown in Figure 2) as it is a practical and lossless technique to preserve video characteristics across different modalities [29]. We denote the general form of the fused video embedding as $\mathbf{h}_{\text{vid}} = [\mathbf{z}^{(\mathcal{K})}, \{\mathbf{z}^{(\mathcal{M}_i)}\}]$ where $\mathbf{z}^{(\mathcal{K})}$ indicates the video knowledge embedding, and $\mathbf{z}^{(\mathcal{M}_i)}$ indicates the video embedding from the supplementary modality ($\mathcal{M}_i \in \{\mathcal{V}, \mathcal{A}, \mathcal{T}\}$). In this work, we take account of up to 3 modalities and do not consider duplication in the fused entity embedding, *i.e.*, $\mathcal{M}_{i-1} \neq \mathcal{M}_i$, $i \leq 3$.

4.2 Multi-modality Fusion for Videos

As a practical way to preserve video info from the supplementary modalities $\mathcal{M}$, concatenating the pretrained embeddings implicitly assumes that each modality is equally important, which does not hold for real-world video data. To capture the importance between the knowledge and the supplementary modalities, TransFusion adopts a partially-trainable embedding fusion model that consists of both the trainable knowledge segment and the fixed pretrained supplementary segment (shown in Figure 3). For the trainable segment, TransFusion uses the predefined f_r to update the knowledge embedding that captures video connectivity in the knowledge graph. For the pretrained segment, TransFusion concatenates the video embeddings projected to the knowledge space from the original modal space $\mathcal{M}$. Altogether, TransFusion leverages the attention mechanism to measure the importance of each modality **(C1)** in the fused video embedding. Next we introduce the projection and attention component of our partially-trainable fusion model.

Table 2: Major symbols and their definition

Symbol	Definition
$G = \{(h, r, t)\}$	a knowledge graph (KG) consisting a set of tuples
E, R	the set of entities and relations in the KG, $h, t \in E, r \in R$
$(\mathbf{h}, \mathbf{r}, \mathbf{t})$	embeddings of the header, relation and tail
$\mathbf{h}_{\text{vid}}, \mathbf{h}_{\text{tag}}$	fused video embedding and tag embedding as the header
$\mathcal{K}$	the knowledge modality
$\mathcal{M} = \{\mathcal{V}, \mathcal{A}, \mathcal{T}\}$	set of supplementary modalities (vision, audio and text)
$\mathbf{z}^{(*)}$	the video embedding from an arbitrary modality $(*)$
f_r	the scoring function of the knowledge embedding
$\{\mathbf{W}, \mathbf{b}\}, \{\mathbf{V}, \mathbf{a}\}$	learnable parameters in the projection and attention layer

4.2.1 Cross-Modality Transformation. In TransFusion, the first step of embedding fusion is to project the pretrained video embeddings from the original modal space into the knowledge space through a non-linear affine transformation (projection layer):

$$\mathbf{z}^{(\mathcal{M}_i)} = \sigma(\mathbf{e}^{(\mathcal{M}_i)} \cdot \mathbf{W}^{(\mathcal{M}_i)} + \mathbf{b}) \tag{1}$$

where $\mathbf{e}^{(\mathcal{M}_i)} \in \mathbb{R}^{d^{(\mathcal{M}_i)}}$ denotes the pretrained video embedding from modality $\mathcal{M}_i$, $i \in \{1, 2, 3\}$ indexes the supplementary modality and $\mathcal{M}_i \in \{\mathcal{V}, \mathcal{A}, \mathcal{T}\}$. $\mathbf{W} \in \mathbb{R}^{d^{(\mathcal{M}_i)} \times d^{(\mathcal{K})}}$ and $\mathbf{b} \in \mathbb{R}^{d^{(\mathcal{K})}}$ are the learnable transform matrix and bias vector, respectively. σ indicates the non-linear operator. Thus, the pretrained embeddings in the original space are transformed into the knowledge space $\mathcal{K}$, and the vanilla form of the fused video embedding is as follows.

$$\mathbf{h}_{\text{vid}} = [\underbrace{\mathbf{z}^{(\mathcal{K})}}_{trainable}, \underbrace{\mathbf{z}^{(\mathcal{M}_1)}, \cdots, \mathbf{z}^{(\mathcal{M}_i)}}_{pretrained.}] \tag{2}$$

we use $\mathcal{M}'$ to denote all modalities including $\mathcal{K}$, and thus the dimension of $\mathbf{h}_{\text{vid}} = |\mathcal{M}'|d^{(\mathcal{K})}$. In this way, TransFusion projects the pretrained visual or textual embeddings from their original modal space into the same knowledge space with the same dimension. More importantly, each modality in the concatenated

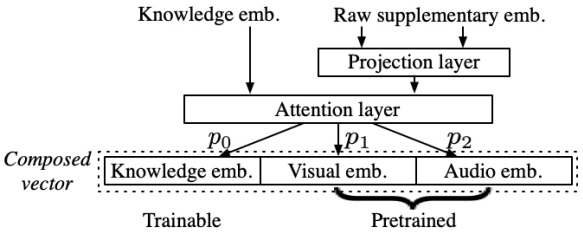


Figure 3: The partially trainable embedding fusion model. The composed vector is used as the fused video embedding $\mathbf{h}_{\text{vid}}$ in KG. Each modality contributes independently to derive embeddings of semantic relations as translation through vector operation.

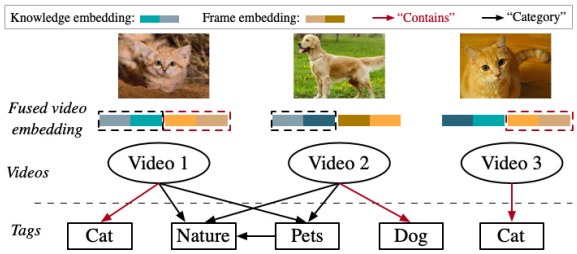


Figure 4: Given a tuple (h, r, t) that indicates the video, semantic relation and the associated tag, TransFusion learns the modality importance in inferring tag t through the relation r. In this example, the knowledge segment is more important to infer tags following the "belongs_to" relation, while the frame segment is more important to infer tags following the "contains" relation.

video representation could contribute independently to inferring the plausibility between a video and the associated tags through vector operation in general translation models. We detail the discussion in Section 4.4. Next we discuss how to model the importance of each modality through the self-attention mechanism **(C1)**.

4.2.2 Modality Attention. Given the fused video representation containing both the knowledge the pretrained supplementary embeddings, TransFusion learns the importance vector $\mathbf{p}$ that represents the impact of each modality in referring tag t following the semantic relation r as a tuple (h, r, t). As the example shown in Figure 4, visual embeddings are more important to infer content-related tags such as "`Cat`" or "`Pets`", and thus should be assigned higher weights when r indicates semantics such as "`contains`". On the other hand, the knowledge embeddings are more important to infer semantics-related tags such as the meta-info or the general description, as well as the tag-tag relations that are irrelevant to video contents. TransFusion adopts the self-attention mechanism to learn the modality importance as the weights for both the trainable and pretrained embeddings across all modalities. Putting them together, the fused video embedding is denoted as:

$$\mathbf{h}_{\text{vid}} = \mathbf{p} \odot \mathbf{z} = [p_0 \cdot \mathbf{z}^{(\mathcal{K})}, p_1 \cdot \mathbf{z}^{(\mathcal{M}_1)}, \cdots, p_i \cdot \mathbf{z}^{(\mathcal{M}_i)}] \tag{3}$$

where $\odot$ denotes the element-wise multiplication, and p_i measures the importance of each modality. To compute p_i, we first define the energy coefficient of modality $\mathcal{M}_i$ as $e_i = a((\mathbf{z}^{(\mathcal{M}_i)}\mathbf{V})^\top)$, where $\mathbf{V} \in \mathbb{R}^{d^{(\mathcal{K})} \times d^H}$ represents the shared linear transformation to the hidden space. a represents the attention mechanism $\mathbb{R}^{d^H} \rightarrow \mathbb{R}$ as the single-layer neural network, which is parameterized with $\mathbf{a}$. Then, TransFusion normalizes p_i as follows:

$$p_i = \text{softmax}(\tanh(e_i)) = \frac{\exp\left(\mathbf{a}^\top \tanh((\mathbf{z}^{(\mathcal{M}'_i)}\mathbf{V})^\top)\right)}{\sum_{j=1}^{|\mathcal{M}'|} \exp\left(\mathbf{a}^\top \tanh((\mathbf{z}^{(\mathcal{M}'_j)}\mathbf{V})^\top)\right)} \tag{4}$$

where $\mathcal{M}'$ denotes all modalities including $\mathcal{K}$, and $\mathcal{M}' = \{\mathcal{K}, \mathcal{V}, \mathcal{A}, \mathcal{T}\}$ if considering all 3 supplementary modalities (vision, audio and text). As the output, $\mathbf{p} = [p_i]$ denotes the normalized importance of both the trainable and pretrained embeddings in $\mathbf{h}_{\text{vid}}$, and $\sum_i p_i = 1$.

In practice, TransFusion adopts the basic MLP architecture for generality and computational efficiency. It also supports other fusion strategies such as graph-based fusion, which we leave it in the future work.

4.3 Relation Attention

In addition to measuring the modality importance, TransFusion also adopts a dual attention mechanism to highlight important relations **(C2)**. For example, multiple relations can be used to infer the tag "`music_video`" such as "`has_person`" and "`BGM_style`", and these relations contribute differently in the inference. In this example, it can be seen that the importance of relations is relevant to specific video modalities as discussed in Section 4.2.2, therefore, we propose a dual attention mechanism to learn the relation weights that is similar to the modality attention as follows:

$$q_i = \frac{\exp\left(\mathbf{a}_r^\top \tanh((\mathbf{r}_i\mathbf{V}_r)^\top)\right)}{\sum_{j=1}^{|R|} \exp\left(\mathbf{a}_r^\top \tanh((\mathbf{r}_j\mathbf{V}_r)^\top)\right)} + 1 \tag{5}$$

where $\mathbf{V}_r \in \mathbb{R}^{d^{(|\mathcal{M}'|\mathcal{K})} \times d^H}$ represents the shared linear transformation to the hidden space. q_i indicates the importance of relation r_i, which impacts all tuples with the same relation $\{(h, r_i, t)\}$ in the training process. The constant offset 1 in Equation (5) scales the attention value q_i and ensures that the "unimportant" tuples can be used to derive KG embeddings in the same way without attention.

4.4 Model Training & Inference

TransFusion casts the tag inference as the link prediction task **(C3)**. Given the predefined translation-based KG embedding approach, TransFusion uses the fused video embedding following Section 4.2 as the basis to compute the representation of the relations and tags. We use TransE as the base translation model (loss function given in Equation (6)) and show this process in Algorithm 1.

$$L = \sum \nabla(\gamma + d(\mathbf{h} + \mathbf{r}, \mathbf{t}) - d(\mathbf{h}' + \mathbf{r}, \mathbf{t}'))_+ \tag{6}$$

where d indicates the distance metrics, such as L_1 or L_2. In Algorithm 1, TransFusion first initializes the embedding of headers, relations and tails (Line 1 - 6). As a header in G indicates either a video or a tag, we follow Equation (3) to derive the fused video embeddings, and initialize the tag embeddings following the base KG embedding approach (TransE). In Line 7 - 13, TransFusion follows the given scoring function f_r and loss L to update the learnable knowledge embedding segment $\mathbf{z}^{(\mathcal{K})}$ in $\mathbf{h}_{\text{vid}}$ as well as $\mathbf{r}$ and $\mathbf{t}$. In the inference stage (Line 16 - 19), give a video and relations of interest, TransFusion first uses f_r to compute the tail embedding $\hat{\mathbf{t}}$, and then rank all tags based on the distance $d(\mathbf{t}, \hat{\mathbf{t}})$ to select the closest one as the inferred tag.

Generality It is worth noting that for the fused video embedding $\mathbf{h}_{\text{vid}}$, TransFusion only updates the knowledge embedding segment ($\mathbf{z}^{(\mathcal{K})}$ in Equation (3)) and the corresponding importance q. Thus, the embeddings of tags $\mathbf{t}$ and relations $\mathbf{r}$ are learned with the constraint to incorporate the fixed pretrained video embeddings from supplementary modalities to minimize the loss. Because the embedding from each modality is losslessly concatenated to represent a video $\mathbf{h}_{\text{vid}}$ and contributes independently to measure the distance between $\mathbf{h}$ and $\mathbf{t}$ using $\mathbf{r}$ as the translation, we claim that TransFusion can be applied to other translation-based approaches In practice, TransFusion is general to be integrated into another translation-based knowledge embedding approach by changing the training process (Line 7 - 13 of Algorithm 1) with corresponding f_r and L. We empirically show experimental results using translation-based models other than TransE in Section 5.3.

Algorithm 1 TransFusion-TransE

Input: Training set $\mathcal{S} = \{(h, r, t)\}$ in G, video embeddings $\mathbf{e}^{(\mathcal{M})}$ from supplementary modalities $\mathcal{M}$, $\mathcal{S}_{\text{query}}$, f_r and loss L.
Output: $\mathbf{h}, \mathbf{r}, \mathbf{t}$ in G and inferred set of tags $\mathcal{S}_{\text{tag}}$.

Initialize $\mathbf{z}^{(\mathcal{K})}$, $\mathbf{h}_{\text{tag}}$, $\mathbf{r}$, $\mathbf{t}$ and trainable parameters, $\mathbf{a}, \mathbf{b}, \mathbf{V}, \mathbf{W}$.
$p_i = \frac{1}{|\mathcal{M}|+1}$ for $i \in [0, M]$, $q_i = \frac{1}{|R|}$ for $i \in [0, |R| - 1]$
for $\mathcal{M}_i \in \mathcal{M}$ **do** ▷ Project pretrained emb.
  $\mathbf{z}^{(\mathcal{M}_i)} = \delta(\mathbf{e}^{(\mathcal{M}_i)} \cdot \mathbf{W}^{(\mathcal{M}_i)} + \mathbf{b})$ ▷ Equation (1)
$\mathbf{h}_{\text{vid}} = [p_0 \cdot \mathbf{z}^{(\mathcal{K})}, p_1 \cdot \mathbf{z}^{(\mathcal{M}_1)}, \cdots, p_i \cdot \mathbf{z}^{(\mathcal{M}_i)}]$ ▷ Equation (3)
loop training epochs ▷ Base method, TransE
  $\mathbf{h} \leftarrow \mathbf{h}/||\mathbf{h}||$, $\mathbf{t} \leftarrow \mathbf{t}/||\mathbf{t}||$
  $\mathcal{S}_{\text{batch}} \leftarrow \text{SAMPLE}(\mathcal{S}, b)$, $\mathbf{T}_{\text{batch}} \leftarrow \emptyset$
  for $(h, r, t) \in \mathcal{S}_{\text{batch}}$ **do** ▷ Sample corrupted triples
    $(h', r, t') \leftarrow \text{SAMPLE}(\mathcal{S}'_{(h,r,t)})$
    $\mathbf{T}_{\text{batch}} \leftarrow \mathbf{T}_{\text{batch}} \cup \{((h, r, t), (h', r, t'))\}$
  Update $\mathbf{z}^{(\mathcal{K})}$, $\mathbf{r}, \mathbf{t}$ w.r.t loss in base method ▷ Equation (6)
  Update $\mathbf{W}, \mathbf{b}$ and $\mathbf{V}, \mathbf{a}$ ▷ Equation (4) and (5)
end loop
for $(h, r) \in \mathcal{S}_{\text{query}}$ **do** ▷ Tag inference
  $\hat{\mathbf{t}} \leftarrow \mathbf{h} + \mathbf{r}$
  $\mathcal{S}_{\text{tag}} \leftarrow \mathcal{S}_{\text{tag}} \cup \text{argmin}_t\, d(\mathbf{t}, \hat{\mathbf{t}})$
return $\mathcal{S}_{\text{tag}}$.

5 EXPERIMENTS

We conduct experiments to evaluate TransFusion from the following perspectives: **Q1.** How well does TransFusion incorporate multi-modal embedding information in inferring video tags? **Q2.** Is TransFusion general enough to be applied to other translation-based knowledge embedding approaches? **Q3.** How much extra workload does TransFusion bring to the basic knowledge embedding methods, and how does it scale with the graph sizes? **Q4.** Can TransFusion infer meaningful tags for given videos? We ran all experiments on a machine with a 14-core 2.40GHz Intel Xeon CPU with 256GB memory and Tesla P40 GPU. For reproducibility, we provide the source code: https://github.com/TencentARC/TransFusion.

5.1 Setup

TransFusion-variants. To testify the capability of TransFusion in integrating the pretrained video embeddings from $\mathcal{V}$, $\mathcal{A}$ and $\mathcal{T}$ modality, we conduct experiments using TransFusion to integrate the single, dual and triple modalities. Namely, they are TransFusion-V/A/T, TransFusion-VA/VT/AT and TransFusion-VAT. To testify the generality, we adopt TransH, TransR, and TransD as the base method and conduct implementation by following OpenKE [8].
Baselines. Due to the lack of baselines with the same goal (*i.e.*, assigning the tag/label through the semantic relation), we adopt three empirical embedding fusion approaches in KG settings. Namely, they are: **B1** application-concatenation, **B2** application-summation, and **B3** multi-modal transformation. Specifically, **B1** concatenates video embeddings from the knowledge and supplementary modalities to form $\mathbf{h}_{\text{vid}}$, and use it to learn $\mathbf{r}$ and $\mathbf{t}$ following the predefined f_r. This approach can be seen as a special case of TransFusion without the modality- and relation-attention module. **B2** is similar to **B1** except that in the learning process, it sums the learned knowledge embedding and multi-modal embeddings to get $\mathbf{h}_{\text{vid}}$. **B3** transforms the pretrained video embeddings from $\mathcal{M}$ only with the modality attention to form $\mathbf{h}_{\text{vid}}$. It can be seen as another special case of TransFusion without having $\mathbf{z}^{(\mathcal{K})}$ in $\mathbf{h}_{\text{vid}}$.
Training. For TransFusion, we set the dimension of the knowledge embedding segment $d^{(\mathcal{K})} = 64$, which is the same for projected multi-modal embeddings from their original space. The dual attention module in TransFusion adopts the multilayer perceptron (MLP) with 1 hidden layer ($d_{\text{hidden}} = 256$) and outputs the scalar attention value. The activation σ is set to be ELU (Exponential Linear Unit). We use the default setup to achieve optimal performance for base methods. To train the models, we use Adam optimizer for 100 epoches with learning rate = 10^{-4} and batche size= 100. As described in Section 4.4, we cast tag inference as the link prediction task, so we evaluate the model performance using two widely-use metrics in knowledge graph embedding: Mean Rank (MR) and HITS@k, *i.e.*, the proportion of correct tags ranked in the top k.

5.2 Tag Inference Performance

We first experiment the effectiveness of TransFusion to integrate video embeddings from 3 modalities $\mathcal{M} = \{\mathcal{V}, \mathcal{A}, \mathcal{T}\}$ in tag inference (**Q1**). Intuitively, $\mathcal{V}$ is more informative than $\mathcal{A}$ and $\mathcal{T}$ as most tags are relevant to the objects. Therefore, our first step is to verify this conjecture by integrating different combinations of $\mathcal{M}$ and determine the optimal. Then, we run TransFusion that integrates the optimal modality combos on both Company datasets. We divide the Company datasets into chunks of size $50K$ to train the models and test on a separate chunk of $20K$ to illustrate the capability of TransFusion in handling data at different scales. For consistency, we use TransE as the base model.

5.2.1 Impact of modalities. We run TransFusion to integrate all possible single, dual and triple supplementary modality combinations from $\{\mathcal{V}, \mathcal{A}, \mathcal{T}\}$ over the first chunk from Company-200K. Table 3 gives the results. It can be seen that when considering a single modality, TransFusion with $\mathcal{V}$ performs the best in all metrics, $\mathcal{A}$ performs second to the best, and $\mathcal{T}$ performs the worst. This is as expected as most tags & relations are related to images or video contents. The video title given by users is mixing: most of the titles tend to describe the video content, but there are also less-relevant words & phrases such as click-baits. Nevertheless, the title is still more relevant to the 380 video classes comparing to the auditory BGM. When integrating two modalities, we observe that the variants TransFusion-VA and -AT achieve better performance than using a single modality. Particularly, TransFusion-VA performs the best with significant improvement in HITS. TransFusion-VT

Table 3: Comparison of model performance in terms of Mean Rank (MR) and HITS@k. Models combining the single, dual and triple modality combos with the optimal performance are marked in bold.

Modalities	Method	MR	HITS@1	HITS@10
Single	TransFusion-V	**1157.8392**	**0.2032**	**0.4688**
	TransFusion-A	1250.9954	0.1785	0.4536
	TransFusion-T	1278.0450	0.1855	0.4547
Dual	TransFusion-VA	**1068.5327**	**0.2162**	**0.5435**
	TransFusion-VT	1212.9903	0.1672	0.4654
	TransFusion-AT	1068.8405	0.1895	0.5094
Triple	TransFusion-VAT	**1038.3549**	**0.2089**	**0.5390**

Table 4: TransFusion (base model: TransE) performance in MeanRank (MR) and HITS@10. Model performing the best and second to the best is marked in bold and $*$, resp.. Overall, all TransFusion variants outperform the base model, and outperform the baselines in most cases, especially when data scale increases. Among the variants, TransFusion-VAT performs the best with at least 9.59% improvement in HITS@10.

Data	Method	50K		100K		150K		200K		250K		300K	
		MR	HITS@10	MR	HITS@10	MR	HITS@10	MR	HITS@10	MR	HITS@10	MR	HITS@10
Comp. 200K	TransE	2655.9757	0.1893	2179.3452	0.2647	1822.1693	0.3109	1668.7373	0.3387				
	Baseline1-VAT	1696.3206	0.1302	1728.1139	0.2475	1661.7216	0.3167	1549.2969	0.3612				
	Baseline2-VAT	1158.4259	0.2859	1213.9927	0.3957	1207.7767	0.4515	1212.2670	0.4839				
	Baseline3-VAT	4636.4943	0.0003	6240.3623	0.0005	7458.4704	0.0006	8430.2834	0.0001	–	–	–	–
	TransFusion-0	2153.6474	0.3123	1622.3542	0.3444	1548.9553	0.3808	1443.0488	0.4191				
	TransFusion-V	1157.8392	0.4688	1036.3844	0.4507	**980.0499**	**0.5927**	993.0402*	0.5990*				
	TransFusion-VA	1068.5327*	**0.5435**	1031.2422*	0.4507*	1023.0957	0.5664	1078.1203	0.5872				
	TransFusion-VAT	**1038.3549**	0.5390*	**999.2196**	**0.4916**	988.6463*	0.5907*	**952.4052**	**0.6038**				
Comp. 300K	TransE	3750.2084	0.1443	3095.4598	0.2148	2698.5626	0.2372	2433.0914	0.2745	2166.9205	0.2842	2010.6706	0.3208
	Baseline1-VAT	2311.6552	0.0650	2154.8289	0.1701	2076.5137	0.2434	2008.5320	0.2953	2745.0868	0.2625	1824.2388	0.3429
	Baseline2-VAT	**1431.2027**	0.2154	1378.2324	0.3235	1414.2717	0.3868	1419.3751	0.4142	2410.1631	0.3123	1443.1374	0.4782
	Baseline3-VAT	4505.9492	0.0004	5878.1708	0.0001	7948.9965	0.0004	9304.8260	0.0006	10837.2260	0.1343	11082.0582	0.0001
	TransFusion-0	3171.0726	0.2516	2514.6353	0.2996	2102.2660	0.3375	1961.0081	0.3390	1796.0561	0.3691	1738.3757	0.3636
	TransFusion-V	2055.5578	0.2808*	1341.5043*	0.4547*	1249.8854*	0.5423*	1213.8453*	0.5653*	1412.4491	0.4851	1270.0986	0.5796
	TransFusion-VA	1813.3729*	0.2682	**1312.0373**	0.4469	1463.1938	0.5079	1389.3730	0.5132	1239.1945*	0.5022*	1102.1073*	0.5977*
	TransFusion-VAT	1942.8047	**0.3829**	1571.6603	**0.4594**	**1218.2960**	**0.5534**	**1064.1529**	**0.5971**	**1115.2981**	**0.5150**	**1085.4599**	**0.6004**

performs worse that TransFusion-V, and it is likely due to the miss-matching between the visual contents and their titles (such as click-baits) in the data. For TransFusion-VAT that integrates all modalities, the performance is slightly worse than TransFusion-VA in terms of HITS, this is likely due to the noise incurred by the inconsistency between modalities, but still, we observe the improved overall Mean Rank. Therefore, we use TransFusion-V, -VA and -VAT as representatives in the following experiments.

5.2.2 Tag Inference. In this experiment, we tune TransFusion to show the effectiveness of each component as the ablation study. We use TransFusion without any modalities (we name this variant as TransFusion-0) to testify the edge attention module only. We also use TransFusion to integrate the optimal single, dual and triple modality combinations shown in Section 5.2.1 to testify the modality attention module. For all methods including the baselines, we incrementally aggregate chunks of $50K$ as the training set.

As shown in Table 4, we first observe that all TransFusion variants outperform the base model TransE, including TransFusion-0 that only contains the edge attention module without any pretrained video embeddings. This indicates the effectiveness of the edge attention module in tag inference. Furthermore, we observe that TransFusion that integrates any extra combination (single, dual or triple) of modalities performs significantly better than TransFusion-0 and all the baselines with reduced MR and at least 9.59% improvement in HITS@10, which shows the effectiveness of the modality attention module used in the fused video embeddings. In terms of the baselines, the application-concatenation **B1** and summation **B2** approach outperform the base model in some cases, which shows the usefulness of these empirical approaches in integrating the supplmentary modalities. The multi-modality transformation, **B3** performs badly, which is as expected as it cannot learn the knowledge embedding segment that is necessary for tag inference. Among the variants, TransFusion-VAT performs the best or second to the best in almost all cases, especially when the scale of training data increases on both Company datasets. This demonstrates the ability of the dual attention modules in fusing the important multi-modal embedding and highlighting the important relations, which is critical in dealing with massive multi-relational data that contains noise or inconsistency between modalities.

5.3 Generality

In this section we evaluate the important design target of TransFusion, generality in facilitating multi-modality fusion for existing translation-based knowledge embedding approaches to perform tag inference (**Q2**). We apply TransFusion to integrate the pretrained embeddings from $\mathcal{M} = \{\mathcal{V}, \mathcal{A}, \mathcal{T}\}$ for 3 translation-based knowledge embedding methods (*i.e.*, TransH, TransR and TransD) that fall into the same categorization as the extensions of TransE, and compare the performance with their vanilla form on the Company-200K dataset. As the reference, we use the result given by TransE from Table 4. Since TransFusion integrating all three modalities tend to perform best (Section 5.2.2), we only report the result given by TransFusion-VAT for brevity. In addition, we employ the public benchmark dataset FB15K237 and evaluate TransFusion on the general task of link prediction. To keep consistent with tag inference for videos, we only follow the "$(h \rightarrow t)$" prediction manner.

Our first observation of Table 5 is that when comparing the vanilla base methods only, TransH and TransD perform comparably well and all 3 methods including TransR perform better than TransE, which follows the finding in the precedent works. More importantly, the base knowledge embedding methods that use TransFusion to integrate three modalities continuously outperform their vanilla forms with reduced MR and $0.91\% - 2.66\%$ improvement in terms of HITS@10. This is as expected because despite the fact that all these base KG embedding approaches adopt different forms of scoring

Table 5: Performance (MR and HITS@k) of TransFusion using different base methods on Company-200K. Base knowledge embedding methods that use TransFusion to integrate pretrained embeddings across modalities continuously outperform their vanilla forms.

Base Method	Variants	MR	HITS@1	HITS@10
TransE	Vanilla	1668.7373	0.0793	0.3387
	TransFusion-VAT	**952.4052**	**0.2286**	**0.6038**
TransH	Vanilla	708.6168	0.3330	0.7860
	TransFusion-VAT	**696.8757**	**0.3434**	**0.7951**
TransR	Vanilla	1061.2629	0.2705	0.6675
	TransFusion-AT	**1019.8175**	**0.2955**	**0.6870**
TransD	Vanilla	737.9201	0.3220	0.7794
	TransFusion-AT	**705.4973**	**0.3283**	**0.8060**

Table 6: Comparison of model performance in MR and HITS@k on FB15K237. Models with optimal performance are marked in bold.

Data	Method	MR	HITS@1	HITS@10
FB15K237	TransE	208.3638	0.2064	0.4919
	Baseline1-V	326.4840	0.1696	0.4452
	Baseline2-V	250.1654	0.1790	0.4448
	Baseline3-V	728.7685	0.1147	0.2682
	TransFusion-0	212.5140	0.2027	0.4908
	TransFusion-V	**195.5638**	**0.2165**	**0.4942**

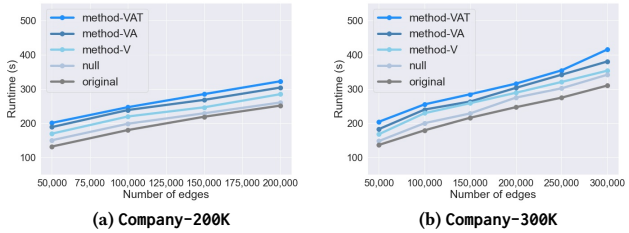
functions or loss, they follow the similar transnational distance manner to derive the embeddings, and thus can adopt TransFusion to integrate multi-modal information. We leave the generality of TransFusion for KG approaches in the other category, semantic matching as one direction of the future work. In Table 6, we observe that TransFusion-V with TransE as the base model achieves improved performance on the benchmark dataset FB15K237. It can be seen that the improvement is limited comparing to the Company datasets, we blame it to the diversity of entities and the relatively low quality of entity images crawled from the public websites. Besides, the images could be biased as we only randomly select one to derive the visual embedding. But still, we observe that by applying TransFusion to the base model TransE, the performance of link prediction is improved, which further demonstrates its generality.

5.4 Scalability Analysis

To evaluate the scalability (**Q3**), we use TransE as the base model and report the runtime of training TransFusion variants to obtain the fused video embeddings on both Company datasets versus their numbers of edges. We also report the runtime of the base model as reference, since it does not have extra computation and should have the lowest runtime. Based on the visualization shown in Figure 5, we observe that the runtime of TransFusion variants is longer than the vanilla TransE, and the runtime increases as the more modalities are integrated. This is as expected as TransFusion requires more computation for the dual attention modules. Nevertheless, TransFusion variants with TransE as the base model still scale linearly with the number of edges in the knowledge graphs, which shows its scalability.

5.5 Tag Inference Case Study

Here we showcase the learned attention values of modalities and important relations to answer **Q4**. Specifically, we report the averaged attention score of each modality on Company-200K and the top-5 relations with highest averaged scores in in Table 7 In Table 8, we list the top-5 ranked tags for 3 randomly selected videos



(a) Company-200K (b) Company-300K

Figure 5: TransFusion variants runtime on two Company datasets. TransFusion scales linearly as the number of edges increase.

Table 7: TransFusion learned modality importance (attention scores) of Compan-200K as well as top-5 important relations.

Modality	Knowledge	Vision	Audio	Text
Score	0.4030	**0.5340**	0.0616	0.0015

Relation	manu_tag	category_3	category_2	category_1	has_person
Score	0.3787	0.2601	0.1283	0.09623	0.0482

outputted by both TransFusion and the base method following the relation with the highest learned attention.

From Table 7, we observe that the knowledge embedding is important with fairly high attention scores on the Company-200K dataset. Besides, $\mathcal{V}$ is the most important, which corresponds to the analysis in Section 5.2.1. Additionally, we observe that the most important relation is manu_tag, which is reasonable as this relation marks the manually-created tags given by the users. The relation category_* denotes the pre-defined tags employed by the business provider, higher values of * indicate finer granularity. An example that corresponds to category_1/2/3 is Animal_Life, Pets and Cats. Table 8 gives the inferred tags following the manu_tag relation. We asked 5 human annotators to watch the 3 videos and mark the inferred tags that they think are relevant to the video content in bold. It can be seen that TransFusion provides high-quality tags that are diverse and close to human interpretation. We also provide the sampled key frames of these 3 videos in Section 7.2 of the appendix to better understand the inferred tags.

Table 8: Inferred tags for 3 videos given by TransFusion and the base model (TransE) on Company-200K.

Training tags	Base Model	TransFusion-VAT
scientific facts, fun experiments, technology, electric shock, meat	animals, show, television drama, comic, magic	**education**, documentary, **scientific news**, **science & technology**, **scientists**
how to make, DIY, recipe, tutorial, pizza	**casual life**, **homemade**, cartoon, temporal work, delicious food	**delicious recipe**, **casual life**, **eating & broadcasting**, street food, **food selfies**
international society, COVID-19, India, disinfection, top news	education, **news**, cartoon, selfies, life skills	**scientific news**, education, **vehicles**, **documentary**, **science & technology**

6 CONCLUSION

In this work, we cast the problem of semantic video tag inference as the semantic link prediction task in knowledge graph and described a general deep learning solution, TransFusion. We propose a partially trainable embedding fusion model to integrate the pretrained video embeddings from multiple modalities and adopts a modality attention module to automatically learn their importance. We also propose an edge attention module to highlight the important semantic relations. Together, the dual attention modules answer the question *what* and *why* in video tag inference. Extensive experiments show the effectiveness of TransFusion on two real-world video datasets in the industry and a publicly-available knowledge base with improvement in both MearRank and HITS. Our experiments also show the linear scalability of TransFusion, and provide the analysis on the learned modality importance in the fused video representation as well as the importance of semantic relations. There are many possibilities for future directions of this work, such as extending it to handle semantic KG embeding approaches, or exploring advanced fusion strategies such as graph-based fusion.

REFERENCES

[1] Antoine Bordes, Xavier Glorot, Jason Weston, and Yoshua Bengio. A semantic matching energy function for learning with multi-relational data. *Machine Learning*, 94(2):233–259, 2014.

[2] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems*, pages 2787–2795, 2013.

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017.

[4] Muhao Chen, Yingtao Tian, Mohan Yang, and Carlo Zaniolo. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. In *IJCAI*, 2017.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[7] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, pages 6202–6211, 2019.

[8] Xu Han, Shulin Cao, Lv Xin, Yankai Lin, Zhiyuan Liu, Maosong Sun, and Juanzi Li. Openke: An open toolkit for knowledge embedding. In *EMNLP*, 2018.

[9] Dongliang He, Zhichao Zhou, Chuang Gan, Fu Li, Xiao Liu, Yandong Li, Limin Wang, and Shilei Wen. Stnet: Local and global spatial-temporal modeling for action recognition. In *Proceedings of the AAAI*, volume 33, pages 8401–8408, 2019.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[11] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 IEEEE ICASSP*, pages 131–135, 2017.

[12] Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. Knowledge graph embedding via dynamic mapping matrix. In *Proceedings of the 53rd ACL and the 7th international joint conference on NLP*, pages 687–696, 2015.

[13] Feng Kang, Rong Jin, and Rahul Sukthankar. Correlated label propagation with application to multi-label learning. In *CVPR*, volume 2, pages 1719–1726, 2006.

[14] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, pages 1725–1732, 2014.

[15] Alexander Klaser, Marcin Marszałek, and Cordelia Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC 2008-19th British Machine Vision Conference*, pages 275–1. British Machine Vision Association, 2008.

[16] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2):167–195, 2015.

[17] Xiangtai Li, Houlong Zhao, Lei Han, Yunhai Tong, Shaohua Tan, and Kuiyuan Yang. Gated fully fusion for semantic segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11418–11425, 2020.

[18] Rongcheng Lin, Jing Xiao, and Jianping Fan. Nextvlad: An efficient neural network to aggregate frame-level features for large-scale video classification. In *Proceedings of the ECCV*, pages 0–0, 2018.

[19] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. Learning entity and relation embeddings for knowledge graph completion. In *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

[20] Ye Liu, Hui Li, Alberto Garcia-Duran, Mathias Niepert, Daniel Onoro-Rubio, and David S Rosenblum. Mmkg: multi-modal knowledge graphs. In *European Semantic Web Conference*, pages 459–474. Springer, 2019.

[21] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

[22] Farzaneh Mahdisoltani, Joanna Biega, and Fabian M Suchanek. Yago3: A knowledge base from multilingual wikipedias. 2013.

[23] Annamaria Mesaros, Toni Heittola, Antti Eronen, and Tuomas Virtanen. Acoustic event detection in real life recordings. In *2010 18th European Signal Processing Conference*, pages 1267–1271. IEEE, 2010.

[24] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[25] Daniel Oñoro-Rubio, Mathias Niepert, Alberto García-Durán, Roberto González-Sánchez, and Roberto J López-Sastre. Answering visual-relational queries in web-extracted knowledge graphs. In *AKBC*, 2018.

[26] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *CoRR*, abs/1804.02767, 2018.

[27] Wenqi Ren, Lin Ma, Jiawei Zhang, Jinshan Pan, Xiaochun Cao, Wei Liu, and Ming-Hsuan Yang. Gated fusion network for single image dehazing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3253–3261, 2018.

[28] Bryan Rink and Sanda Harabagiu. Utd: Classifying semantic relations by combining lexical and semantic resources. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 256–259, 2010.

[29] Ying Shan, T Ryan Hoens, Jian Jiao, Haijing Wang, Dong Yu, and JC Mao. Deep crossing: Web-scale modeling without manually crafted combinatorial features. In *Proceedings of the 22nd ACM SIGKDD*, pages 255–262, 2016.

[30] Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. Reasoning with neural tensor networks for knowledge base completion. In *Advances in neural information processing systems*, pages 926–934, 2013.

[31] Yan Song, Shuming Shi, Jing Li, and Haisong Zhang. Directional skip-gram: Explicitly distinguishing left and right context for word embeddings. In *ACL: Human Language Technologies, Volume 2 (Short Papers)*, pages 175–180, 2018.

[32] J Sun. Jieba chinese word segmentation tool, 2012.

[33] Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019.

[34] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[36] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *ICCV*, pages 3551–3558, 2013.

[37] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, pages 20–36. Springer, 2016.

[38] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. Knowledge graph and text jointly embedding. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1591–1601, 2014.

[39] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. Knowledge graph embedding by translating on hyperplanes. In *AAAI*, volume 14, pages 1112–1119, 2014.

[40] Ruobing Xie, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. Image-embodied knowledge representation learning. *arXiv preprint arXiv:1609.07028*, 2016.

[41] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*, 2014.

[42] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*, 2019.

[43] Yongjing Yin, Fandong Meng, Jinsong Su, Chulun Zhou, Zhengyuan Yang, Jie Zhou, and Jiebo Luo. A novel graph-based multi-modal fusion encoder for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3025–3035, 2020.

[44] Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers*, pages 2335–2344, 2014.

[45] Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. Attention-based bidirectional long short-term memory networks for relation classification. In *The 54th ACL (Volume 2: Short Papers)*, pages 207–212, 2016.

7 SUPPLEMENTARY MATERIALS

7.1 Multi-modal video embedding

In this section we provide the details of obtaining video embeddings from multiple modalities. Specifically, we trained two classification models for visual modality. The first model is a StNet [9] with ResNet-50 [10] as backbone. We train the StNet on our large-scale labeled data with 16 frames uniformly sampled from a video as input. The second model consists of a pretrained EfficientNet-B3 [33] and a NextVlad [18]. 32 frames are uniformly sampled from a video and converted into visual frame-level features with pretrained EfficientNet-B3. We train the Nextvlad aggregating visual frame-level features with our large-scale labeled data. Embeddings extracted from two classification models are concatenated to form the final visual embeddings. One classification model is trained for auditory modality. Our model consists of a pretrained VGGish [11] and a Nextvlad [18]. According to [11], we divide the background audio of a video into non-overlapping 960 ms frames, and only consider the first 32 frames. Log-mel spectrograms of 32 frames are computed and presented to the pretrained VGGish to extract auditory frame-level features. We train the Nextvlad aggregating auditory frame-level features with our large-scale labeled data. One classification model is trained for textual modality. Our model is a LSTM-attention [45]. We tokenize the descriptive title of the video with Jieba [32] and vectorize every word according to the word-book [31]. We train the LSTM-attention with word vectors as input on our large-scale labeled data.

7.2 Detailed Video Frames

Here we provide the titles and detailed key frames sampled from the 3 videos in Section 5.5 to support the tags that are inferred by TransFusion. As shown in Figure 6, the key frames are sampled consecutively over the fixed intervals (12s for video 1, 2s for video 2 and 3).

The titles are as follows:

- video 1: `The electric shock experiment on fresh meat: the consequence and discovery.`
- video 2: `Homemade “pizza” using ingredients at hand.`
- video 3: `A scene of street disinfection in India amid Covid-19 global pandemic.`

Among the tags inferred, we find that TransFusion is capable of inferring more diverse tags, such as “scientists” of video 1 and “street food” of video 2, as well as tags that are relevant to the video content, such as “vehicle” of video 3. Besides, all these tags are closer to human interpretation of the video content.

(a) Sampled frames of video 1

(b) Sampled frames of video 2

(c) Sampled frames of video 3

Figure 6: Sampled frames of videos in the case study of Section 5.5. These key frames are sampled consecutively over the fixed intervals. For video 1 (a), the interval is 12 seconds. For video 2 (b) and 3 (c), the interval is 2 seconds.