Grounding Visual Concepts for Zero-Shot Event Detection and Event Captioning

Zhihui Li+
Shandong Normal University
University of New South Wales
zhihuilics@gmail.com

Xiaojun Chang+
Monash University
Melbourne, Australia
cxj273@gmail.com

Lina Yao
University of New South Wales
Sydney, NSW
lina.yao@unsw.edu.au

Shirui Pan
Monash University
Melbourne, Australia
shirui.pan@monash.edu

Ge Zongyuan
Monash University
Airdoc Research, Australia
zongyuan.ge@monash.edu

Huaxiang Zhang*
Shandong Normal University
Jinan, Shandong
huaxzhang@163.com

ABSTRACT

The flourishing of social media platforms requires techniques for understanding the content of media on a large scale. However, state-of-the art video event understanding approaches remain very limited in terms of their ability to deal with data sparsity, semantically unrepresentative event names, and lack of coherence between visual and textual concepts. Accordingly, in this paper, we propose a method of grounding visual concepts for large-scale Multimedia Event Detection (MED) and Multimedia Event Captioning (MEC) in zero-shot setting. More specifically, our framework composes the following: (1) deriving the novel semantic representations of events from their textual descriptions, rather than event names; (2) aggregating the ranks of grounded concepts for MED tasks. A statistical mean-shift outlier rejection model is proposed to remove the outlying concepts which are incorrectly grounded; and (3) defining MEC tasks and augmenting the MEC training set by the videos detected in MED in a zero-shot setting. To the best of our knowledge, this work is the first time to define and solve the MEC task, which is a further step towards understanding video events. We conduct extensive experiments and achieve state-of-the-art performance on the TRECVID MEDTest dataset, as well as our newly proposed TRECVID-MEC dataset.

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

KEYWORDS

Multimedia Event Detection, Multimedia Event Captioning, Grounding Visual Concepts, Zero-shot Learning

+ indicates Zhihui Li and Xiaojun Chang contribute equally to this paper. Corresponding author: Huaxiang Zhang (huaxzhang@163.com).

KDD '20, August 23–27, 2020, Virtual Event, USA

ACM ISBN 978-1-4503-7998-4/20/08…$15.00
https://doi.org/10.1145/3394486.3403072

ACM Reference Format:
Zhihui Li+, Xiaojun Chang+, Lina Yao, Shirui Pan, Ge Zongyuan, and Huaxiang Zhang*. 2020. Grounding Visual Concepts for Zero-Shot Event Detection and Event Captioning. In *26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '20), August 23–27, 2020, Virtual Event, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394486.3403072

1 INTRODUCTION

The abundance of consumer recording devices, such as smart phones and tablets, has lead to the unprecedented sharing of personal videos on social media platforms such as YouTube and Flickr. This growing plethora of visual content necessitates the development of robust and scalable techniques designed to tackle video understanding, including those for video indexing, search, retrieval, and summarization. While these topics have attracted substantial research attention in the multimedia community over the years, the core challenges related to dealing with large-scale video data, the inter- and intra-class variability of multimedia events, and the ability to learn from limited labeled data.

Previous research [3, 6, 7, 9, 12, 16, 27, 28, 30, 45–47, 58, 62] has mostly focused on recognizing video actions/activities (*e.g.*, hammering, pull-ups, and push-ups), which has limited the scope of such research to human-centric events and video content. Multimedia Event Detection (MED) addressed this shortcoming by introducing a generic video categorization task, where the goal is to detect more generic user-defined events in the videos (such as 'parade', 'grooming an animal', and 'dog show'). However, although MED focuses on the general video categorization task, the label set is within a predetermined set of events, potentially limiting specificity and expressiveness. Accordingly, in this paper, we define for the first time the more generic video understanding task of Multimedia Event Captioning (MEC). The MEC task is defined as *generating event-centric sentence descriptions of the video content.* The MEC task is inspired by, and yet different from, the task of video caption generation, as the later is aimed at generating sentences that "describe the image/video content" [36, 37]. However, generated captions typically contain trivial/common knowledge that is often less important for video understanding. By contrast, in the MEC task, we aim to generate event-centric sentence descriptions.

MED and MEC tasks are, however, extremely difficult and require at least three challenges to be addressed: (1) **Data sparsity.** While video data is abundant, labeled video data is not; therefore,

developing the ability to learn from few (or, in zero-shot cases, no) labeled instances is both important and difficult. This problem is further exacerbated by the high intra-class variance of visual content. (2) **Semantically uninformative event names.** Many previous zero-shot learning methods [5, 8, 18, 29, 53, 60] have semantically embedded the event names themselves (using *e.g.* word2vec) as the underlying representation to transfer the knowledge from known to unknown classes. However, certain event names may be too short to describe the main concepts depicted in the event video; such representations are thus less semantically relevant and informative for knowledge transfer. (3) **Unmatched visual and textual concepts.** Existing semantic approaches in the image domain (*e.g.* DeViSE [17]) seek to construct a joint visual-semantic embedding by matching the visual content of an image against the corresponding textual descriptions in the embedding space. However, such approaches tend to be more challenging in video domains since both event descriptions and event videos may be more diverse and exhibit greater variety. As such, only certain aspects of the event description may be exhibited (or supported) by a particular video.

To address these challenges, we propose ranking grounded visual concepts for zero-shot MED tasks. More specifically, rather than using event names directly, we extract the semantic representation for each event from its textual description. Large-scale auxiliary concept detectors are employed to ground the visual concepts of each testing video and predict the likelihood (score) of the concept being present in videos. Here, we define grounding [63] as "providing the video visual frames and concepts/events correspondences". We further aggregate the scores to ground events while removing outliers via the mean-shift outlier rejection model. As a more generic video understanding task, we define the MEC task such that the training set of MEC is augmented by the videos detected in MED in the zero-shot setting. Our experimental results validate the effectiveness of our framework for both MEC and MED tasks.

Contributions. We make the following contributions in this paper:

(1) To the best of our knowledge, Our work is the first time to define and solve the MEC task, which represents substantial progress in the understanding of video events. Specifically, we propose a framework for aggregating the ranks of grounded concepts for MED and MEC tasks; this framework can better utilize the grounded concepts to detect the zero-shot video event and generate multimedia event captioning.
(2) We present novel semantic representations of events by fusing both the frequency feature – Term Frequency-Inverse Document Frequency (TF-IDF) with the semantic word vector representations of key words from the events' textual descriptions.
(3) In order to robustly aggregate grounded visual concepts, the Statistical Mean-shift outlier model is formulated to prune the outlier concepts.

2 RELATED WORKS

Zero-shot Video Classification. The problem of recognizing novel classes without the aid of any example has been widely explored in the image and video domains [11, 15, 18, 20, 22, 26, 31, 35, 54, 55, 59] through the learning paradigm of transferring the knowledge of existing auxiliary classes to recognize the unknown novel classes. This paradigm, however, is very problematic when it comes to obtaining the "Encyclopedia" auxiliary classes. Thus, we use a more realistic zero-shot setting for event analysis, *i.e.* only the novel testing videos are available, rather than accompanying these with the large-scale auxiliary video event classes.

Concepts in Video Event Detection. Consequently, an event is a higher-level semantic entity that is typically composed of multiple concepts. For example, a "birthday party" event consists of concepts such as "blowing candle" and "birthday cake". Concept detection has been studied for decades in the multimedia community, and there are a number of huge well-labeled video datasets [23, 24, 50, 57] available for use in concept detection tasks as a result. As for event detection, previous research has investigated the use of concept detectors [4, 19, 32, 41]. Chang *et al.* evaluated the semantic correlation of each pretrained video concept to predict the video event of interest. Such semantic correlation, however, may potentially suffer from the problems of *Irrepresentable event names* and *Unmatched visual* and *textual concepts*. By contrast, we compute the semantic relatedness between the textual descriptions of events and pre-defined auxiliary concepts, which yields in semantically richer concept weights (Eq. 2).

Video Captioning. Early works on video captioning considered tagging videos with metadata [1] and clustering captions and videos [21] for retrieval tasks. Recently, Recurrent Neural Networks (RNNs) and Long Short Term Memory (LSTM) have been introduced for video captioning [36, 52, 56]. These approaches use memory cells to store, modify, and access internal state, allowing for the discovery of long-range temporal relationships. Stacked RNN is proposed in [34, 46] and used as a visual encoder and language decoder for video captioning in [51]. Pan *et al.* further improved stacked RNN by introducing a pyramid-shaped hierarchical RNN model [36]. Compared with previous captioning tasks, our MEC task aims at generating sentences more focused on interpreting the events in the video. Specifically, due to the high interclass variance of video content, previous video captioning works may only output sentences pertaining to trivial details or the video background, which may not necessarily explain the depicted video event. By contrast, our MEC task *requires the captions to reflect the event properties in the videos and provide event-level sentence descriptions*.

Rank aggregation. This is a classical problem most closely associated with social choice theory that dates back to Condorcet's famous treatise [2, 14, 61]. Two of the best-known rank aggregation methods are Kemeny [25] and Borda count approaches. The Kemeny approach minimized the sum of the Kendall tau distances to all the voters' lists. Although this method theoretically approximates the best rank aggregation very well, it is NP-hard to compute. Borda count ranks items by the number of times they beat other times, and will converge to an optimal ranking under certain moderate conditions [40]. Due to the computational feasibility of this method and the theoretical guarantee it provides of approximating the optimal ranking, we thus develop an extension ranking aggregation algorithm based on Borda count.

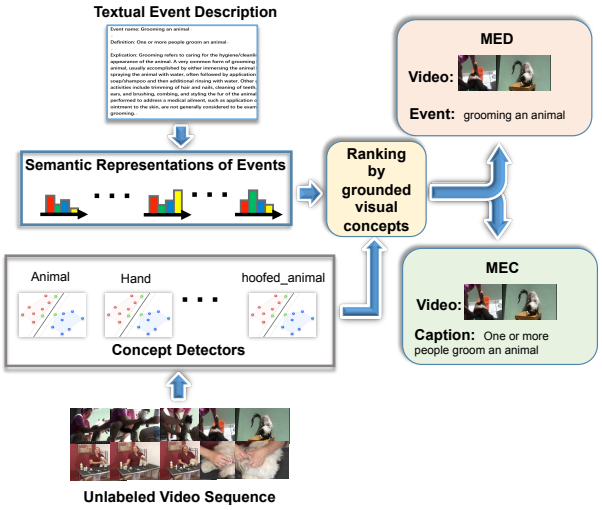


Figure 1: The overview of our framework.

3 THE PROPOSED APPROACH

Our work aims to rank grounded visual concepts for multimedia event detection and multimedia event captioning tasks. The overview of the framework is presented in Figure 1. Each component will be discussed below.

3.1 Problem Context and Definition

We focus on zero-shot video event detection for cases where neither labeled training instances, nor large-scale auxiliary video event data are available for transfer. Specifically, we have only an unlabeled testing video dataset,

$$\mathcal{D} = \{X_i\}_{i=1,\ldots,N} \tag{1}$$

where the i-th video is assumed to be represented by the feature vector $X_i = [x_1, \ldots, x_{n_i}]$; here, x_j is the feature representation of the j-th frame and video i contains a total of d_i frames, while N is the number of test videos in the dataset.

Given X_i, our goal is to estimate the likelihood of video i belonging to one of the pre-defined events, $e \in \mathcal{E}$, which are described (only) by a textual description. The textual description of each event is assumed to be derived from one or more of a multitude of sources, *e.g.* Wikipedia or event ontology definition. Since we do not have access to labeled video event data, we employ large-scale concept detectors to help accomplish the event detection task. The concept detectors are trained on five different concept datasets: YFCC 100M [50], UCF101 [47], Google Sports [24], TRECVID SIN [23] and DIY [57]. The semantic concepts are used here as intermediate latent representation for use in detecting video events (MED) and describing videos (MEC), and can be more formally defined as follows:

$$C = \{(f_m(X), \Phi_m)\}_{m=1,\ \ldots,\ M}, \tag{2}$$

where M is the total number of concepts, Φ_m is the textual description/name of the concept m, and $f_m(X)$ is the corresponding concept detector that gives us concept-level grounding.

We aggregate the scores of various grounded visual concepts to detect video events. The key here is to use concepts as intermediate representation; moreover, it is also necessary to establish which concepts are useful for representing which events, and to what extent, based on the textual description of both the events and concepts. For example, a "birthday party" event is highly associated with several concepts, such as "blowing candle", and "birthday cake", which, if all of these concepts are found to be grounded in one video, this constitutes good evidence to support the identification of the video as a "birthday party" event.

3.2 Methodology

To establish the correspondence and importance of specific concepts to an event description, we semantically match the "atomic" concept descriptions (which typically consist of 1-3 words) to the longer textual descriptions of events. This yields a weighted contribution of each concept to an event. We then compute the likelihood of grounding the concept, given the test video, and aggregate across concepts using computed weighting to produce the likelihood of the event. The whole framework has four steps, as follows:

Extracting Semantic Representations of Events. As described above, we extract the weighted semantic concepts in terms of the textual descriptions of each event. This involves the following two sub-steps:

(1) Weighted keyword extraction. We first extract keywords from the textual event descriptions. The key word set of event $e \in \mathcal{E}$ is denoted as follows:

$$A_e = (a_{1,e},\ \ldots,\ a_{i,e},\ \ldots\ a_{m,e}), \tag{3}$$

where $a_{i,e} \in \mathcal{V}$ is the i-th unique keyword obtained from the description of event e and $\mathcal{V}$ is the keyword vocabulary. These keywords can be interpreted as either semantic concepts or attributes [4, 18]. For each keyword, we compute the Term Frequency-Inverse Document Frequency (TF-IDF) weight, which reflects the extent to which a particular keyword is important in the event descriptions for a particular event[1]. The TF-IDF weight is denoted by $\mathbf{tfidf}(a_{i,e})$ for keyword $a_{i,e}$ in event e.

(2) Word vector representation of keywords. While extracted keywords can be treated as concepts, conducting exact matching between a keyword and the pre-defined concept set C may be difficult. For a more semantic treatment of the problem, we convert keywords to a distributed word vector representation [33]. More specifically, we learn a K-dimensional embedding ϕ_a for each $a \in \mathcal{V}$. In this work, we use $K = 100$, while the skip-gram model [33] is used to train ϕ_a on a large-scale text corpus that includes around 7 billion words: these are derived from the UMBC WebBase (3 billion words), the latest Wikipedia articles (3 billion words) and some other documents (1 billion words).

The semantic representation of event e is thus,

$$W_e = \{\mathbf{w}_{a_{i,e}}\}_{i=1}^{|A_e|}, \tag{4}$$

where $\mathbf{w}_{a_{i,e}} = \mathbf{tfidf}(a_{i,e})\dot{\phi}_{a_{i,e}}$. The TF-IDF weight term suppresses keywords that appear too frequently in all event descriptions, while the word vector representation of keywords enables semantic similarity matching with pre-computed concept detectors.

[1]For events that have multiple descriptions, we compute TF-IDF over all descriptions.

Grounded visual concept description. Concept detectors trained on large-scale concept datasets are used here for event detection. To this end, we train concept classifiers by (1) extracting improved dense trajectory features (including HOG, HOF and MBH) for each concept video dataset and encoding them with Fisher vector representations and (2) training a SVM model for each concept classifier. Platt scaling [39] is used to calibrate the output of SVM classifiers into probability distributions, which yields the probability of a given concept being present in a video. Formally, we define the SVM concept mapping as follows:

$$f_j : X \rightarrow s_j, \tag{5}$$

where the score of the i-th test video containing the j-th video concept is $f_j(X_i) = s_{j,i}$.

To correlate the events with the concepts, moreover, we also define the semantic similarity of event e and the j-th concept as follows:

$$\mathrm{w}(A_e, \Phi_j) = (1-\lambda) < W_e, \phi_{\Phi_j} > + \lambda \cos(\phi_e, \phi_{\Phi_j}), \tag{6}$$

where

$$< W_e, \phi_{\Phi_j} >= \sum_{i=1}^{|A_e|} \mathbf{tfidf}(a_{i,e}) \dot{\cos}(\phi_e, \phi_{\Phi_j}) \tag{7}$$

measures the semantic relatedness between the event keywords and the concept name Φ_j; $\dot{\cos}()$ denotes the cosine similarity. The second smoothing term is the average of the semantic relatedness between event names and concept names. ϕ_e and ϕ_{Φ_j} indicate the word vector that averages all textual words from the names of event e and concept Φ_j, respectively. Note that cosine similarity is used here, since high-dimensional word embedding vectors are naturally directional and cosine distance is more robust to noise than the other metrics such as Euclidean distance [33].

Multimedia Event Detection. Recall that $s_{j,i} \in \mathbb{R}$ is the score probability of concept j occuring in test video i. By taking these scores, the weighted similarity between concepts and the video event class e defined in the last two steps, we aim to recover the likelihood vector (that can also be taken as ranking) across all the test instances in $\mathcal{D}$ of them belonging to event e. To this end, we first define the score vector for concept j across the test dataset as follows:

$$\mathbf{s}_j = \{s_{j,i}\}_{i=1}^{N} \in \mathbb{R}^N. \tag{8}$$

We can then obtain the score vectors for $\mathbf{s}^{(e)} \in \mathbb{R}^N$; here, each element is the likelihood of the corresponding test video belonging to event e. The natural formulation for this goal is to solve the following optimization problem:

$$\mathbf{s}^{(e)} = \arg\min_{\mathbf{s}} \sum_{j=1}^{M} \mathrm{w}(A_e, \Phi_j) \cdot \delta(\mathbf{s}, \mathbf{s}_j), \tag{9}$$

where $\delta(\mathbf{s}, \mathbf{s}_j)$ measures the distance between two score vectors/ranked lists. In terms of statistics, there are two potential choices for the distance metric: Spearman footrule distance and Kendall's tau distance. As explained in the related work and an extension of the Borda count, Spearman footrule distance is more suitable for $\delta(\cdot)$ as a compromise between computational cost and providing a good approximation to the optimal ranking [40]. The Spearman footrule distance is defined as follows:

$$\delta(\cdot) = \sum_{t \in s^\dagger \cup s^\ddagger} |r^{s^\dagger} - r^{s^\ddagger}|, \tag{10}$$

where $r^{s^\dagger}(t)$ is the rank of item t in the list $s^\dagger$; $r^{s^\ddagger}(t)$ is defined likewise.

Our model further employs the mean-shift outlier rejection model, which is an extension of the model proposed in [44], to account for potential outlying errors in the concept grounding $\mathbf{s}_j$, as follows,

$$\mathbf{s}^{(e)} = \arg\min_{\mathbf{s}, \triangle^{(j)}} \sum_{j=1}^{M} \mathrm{w}(A_e, \Phi_j) \cdot \delta(\mathbf{s}, \mathbf{s}_j + \delta_j) + \gamma |\triangle_j|, \tag{11}$$

where $\triangle_j$ is the error vector that accounts for outliers within the grounding of the j-th concept across all videos. Moreover, Equation (11) can be efficiently solved by means of the generalized conditional gradient [4].

Multimedia Event Captioning. The MEC is defined for the first time in this paper, and is a more generic video understanding task. Since there is no MEC dataset available, we here construct the first event captioning dataset, TRECVID-MEC, which will be released upon acceptance. To construct the TRECVID-MEC dataset, we use videos from the 30 events in the MEDTest-13 and MEDTest-14 datasets along with all corresponding videos. More specifically, the training instances of the TRECVID MED-EX100 tasks (around 100 training instances per event) are used here; we split videos into training (1000 videos), validation (996 videos) and test (1000 videos) sets. We further construct a caption pool to describe each of the events. In the caption pool, each event is described by at least nine different sentences that describe the important concepts and contents of the event. The sentences are constructed by expert annotators refer to the training, validation and testing videos contained in our dataset.

For the MEC captioning, we use the hierarchical structure and attention model in LSTM layers, which is a variant of the state-of-the-art video captioning model, Hierarchical Recurrent Neural Encoder (HRNE) [36]. Recurrent Neural Networks (RNNs) are widely used in video captioning tasks due to their superior performance in sequence to sequence learning. Unlike former video captioning works [52] where stacked RNN [48] is used. Hierarchical Recurrent Neural Encoder (HRNE) [36] proposes a pyramid-shaped RNN structure. The basic component of HRNE is a two-layer RNN. The input sequence is divided into several chunks; the first RNN layer consists of several RNN chains, which are used as temporal filters to calculate the features of these chunks. The second RNN layer then takes these features as input and extracts the feature of the whole video.

Compared with stacked RNNs, HRNE is able to efficiently explore temporal information in a longer range. The computation operations are also significantly lessened due to its pyramid-shaped structure. Furthermore, it is able to uncover temporal transitions between frame chunks with different granularities.

The model is trained from scratch on our TRECVID-MEC dataset. During training time, the captions of training videos are dynamically selected from the corresponding caption pool. In more detail,

each sentence of the caption pool can be represented as the semantic vector as in Eq (1); and visual concepts of each video can be grounded using concept detectors. By using Equation (6), we can compute the similarity between each sentence and each concept grounded in one video. So the best training caption is selected as the one with the highest similarity of all concepts to the sentence. The selected sentence together with the corresponding video are used as the new training data to train our model.

We also augment the training data for MEC task in the zero-shot setting, to better train our model. Specifically, the training set is expanded by using the original testing video in the MEDTest-13 and MEDTest-14 datasets, accompanied with the corresponding event labels estimated by MED. In other words, we train our model in a more realistic setting, *i.e.* given the unlabeled testing videos, we first employ MED to estimate their event labels, then use these videos to train the model for event captioning.

4 EXPERIMENTS

We conduct three different experiments to validate our framework. In each subsection, we first introduce the experimental setting, then discuss the experimental findings. For all experiments, we generate unigrams, bigrams and trigrams and compute the TF-IDF feature vectors for all of these tokens. A pre-defined word list is used to filter out stop words (such as "and", "is"); while keywords are defined as those with high values in the TF-IDF vector.

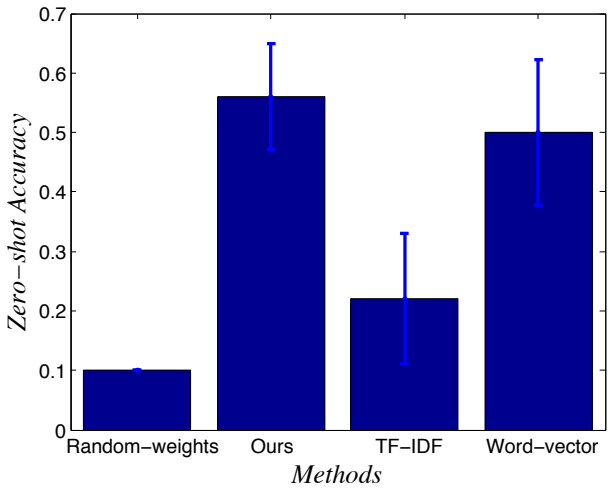


Figure 2: Results of Guessing Novel Event Names.

4.1 Guessing Novel Event Names

This experiment validates whether the weighted keywords extracted from the event description provide good semantic representations that can be generalized for novel events. This experiment is motivated by the fact that humans are able to guess event names when provided only with key descriptions. For example, when asked which event is characterized by "blowing candle", "birthday cakes" and "clamping hands", a person is likely to respond "birthday party".

(1) Experimental Setup. Following the standard zero-shot learning settings [26], we use the 30 events from both MEDTest-13 and MEDTest-14 and randomly divide them into 20 auxiliary ($\mathcal{E}_{au}$ and 10 testing ($\mathcal{E}_{te}$) event classes respectively. Experiments are then repeated five times in order to reduce variance.

Here, we use only the textual descriptions of each event. The semantic representation of each event can then be computed using Eq (1). Furthermore, we use the same 100-dimensional word vectors to directly represent event names (as opposed to event descriptions), denoted as ϕ_e. We compute word vector representations for both auxiliary and test event names. The support vector regressors are learned from semantic representation to each dimension of vectors of all auxiliary event names; given the semantic representation of testing events, we can predict the 100-dimensional word vectors of testing event names. The event names are assigned by matching the predicted vectors against ϕ_e, $e \in \mathcal{E}_{te}$ using cosine nearest neighbour distance.

Three different baselines are compared against the keyword representation $\{W_e\}$: (1) word-vector: we calculate the weighted average of a keyword's word vector embedding with the TF-IDF weight in Equation (4), *i.e.* $\{\phi_{a_{i,e}}$; (2) TF-IDF: we directly use the TF-IDF feature vector; (3) random-weights, which use the representation generated randomly.

(2) Experimental Results. The results are presented in Figure 2 with the chance-level 10% performance (also corresponding to Random-weights). Our method achieves the best accuracy overall, and also outperforms the methods of word-vector and TFIDF methods by 10 and 30 absolute percentage points respectively. Notably, the improved results achieved by our method are mainly due to the application of the TF-IDF feature vector to measure more reliable visual concepts. In short, the improvements illustrated here validate the effectiveness of our weighted keyword semantic representation.

4.2 The MED Experiments

MEDTest-13 and MEDTest-14 were introduced by NIST for all participants in the TRECVID competition in 2013 and 2014 respectively. They consequently serve as the standard benchmark for zero-shot video event detection algorithms. The main experiment is conducted on MEDTest-14; we also include the additional validation on MEDTest-13 in supplementary material.

(1) Experimental setup. We use the standard setups of zero-shot event detection on MEDTest-14 dataset. Specifically, the official test split released by the NIST is used here and we strictly adopt standard procedure. The results are measured by the classification accuracy by using mean Average Precision (mAP). MEDTest-14 has 20 events with descriptions. MEDTest-14 includes 24,000 videos and the event IDs are shown in the X-axis of Fig 3.

We extract from videos the improved dense trajectory features (including HOG, HOF and MBH) and we further employ fisher vectors [43] to encode these features. Particularly, the dimension of each descriptor is first reduced by a factor of 2 and for each type of features, we use Gaussian Mixture Models with 256 Gaussians to compute Fisher vectors. The λ is set to 0.01 in Equation (6).

We implement and compare against several different methods developed for MED task; all these methods use features identical to ours. Particularly, we compare three different variances of [19]: (1) *Prim*, (2) *OR* and (3) *Fu*. Furthermore, we compare against (4) *Sel* [32], (5) *Bi* [41], (6) *Bor*: the Borda rank aggregation with equal

Table 1: Event IDs and their corresponding names.

Event ID	Event Names	Event ID	Event Names
E006	Birthday party	E026	Renovating
E007	Changing a vehicle tire	E027	Rock climbing
E008	Flash mob gathering	E028	Town hall meeting
E009	Getting a vehicle unstuck	E029	Winning a race without a vehicle
E010	Grooming an animal	E030	Working on a metal crafts project
E011	Making a sandwich	E031	Beekeeping
E012	Parade	E032	Wedding shower
E013	Parkour	E033	Non-motorized vehicle repair
E014	Repairing an appliance	E034	Fixing musical instrument
E015	Working on a sewing project	E035	Horse riding competition
E021	Attempting a bike trick	E036	Felling a tree
E022	Cleaning an appliance	E037	Parking a vehicle
E023	Dog show	E038	Playing fetch
E024	Giving directions to a location	E039	Tailgating
E025	Marriage proposal	E040	Tuning a musical instrument

Table 2: Experiment results for Zero-Shot event detection on MEDTest 2014. Mean average precision (mAP), in percentages, is used as the evaluation metric. Larger mAP indicates better performance.

MEDTest 2014									
ID	Prim	Sel	Bi	OR	Fu	Bor	PCF	DCC	Ours
E021	2.1	3.0	2.6	3.9	4.0	3.1	4.6	6.4	**7.9**
E022	0.8	1.0	0.8	1.4	1.5	1.2	1.5	2.9	**3.6**
E023	33.9	36.9	35.2	39.2	40.9	38.7	41.8	44.3	**46.8**
E024	2.6	3.8	3.0	4.7	4.9	4.1	4.9	6.1	**7.7**
E025	0.5	0.8	0.6	1.0	1.4	0.8	1.0	1.3	**1.9**
E026	1.0	1.6	1.3	2.4	3.0	2.0	2.7	4.2	**5.7**
E027	11.2	13.6	12.5	15.9	16.3	15.1	16.5	19.6	**21.5**
E028	0.8	0.7	1.1	1.6	2.0	1.7	2.3	4.0	**5.3**
E029	8.4	10.7	12.2	14.0	14.9	13.2	14.8	**17.7**	16.8
E030	0.4	0.6	0.5	0.9	1.0	0.4	0.5	0.5	**0.7**
E031	32.8	53.2	45.9	69.5	69.7	67.5	72.6	77.5	**79.1**
E032	3.1	5.9	4.4	8.1	8.5	7.5	8.7	**11.4**	10.8
E033	15.3	20.2	18.5	22.1	22.2	21.5	23.3	26.6	**28.3**
E034	0.3	0.5	0.4	0.7	0.8	0.5	0.8	0.9	**1.1**
E035	9.3	13.3	11.1	16.5	16.7	15.8	18.7	21.8	**23.1**
E036	1.9	2.6	2.1	3.2	3.4	2.9	3.8	5.5	**7.2**
E037	2.2	4.5	3.8	6.8	6.9	5.4	6.8	8.5	**9.6**
E038	0.7	0.7	0.6	1.0	1.2	0.9	2.3	**2.9**	2.6
E039	0.4	0.6	0.4	0.7	0.8	0.6	0.9	**2.3**	1.9
E040	0.7	1.0	0.7	1.6	1.6	1.2	1.8	3.1	**3.7**
mean	6.4	9.6	7.9	10.8	11.1	10.2	11.4	13.4	**14.7**
MEDTest 2013									
ID	Prim	Sel	Bi	OR	Fu	Bor	PCF	DCC	Ours
mean	7.1	7.9	6.9	9.5	9.9	8.4	10.0	12.6	**15.3**

weights on the discovered semantic concepts, (7)*wBor* [5], and (8) *PCF* [4].

Experimental Results. The results on MEDTest 2014 and MEDTest 2013 are listed in Table 2. From the results, we can see that our method outperforms all other approaches by a large margin: our approach achieves a mean mAP performance of 14.7 on MEDTest-14, which is an improvement of 9.7% over the next closest competitor. We also highlight several observations based on the mean performance: (1) The results of the Bor approach are better than those of Sel, Bi and Prim. This further validates that our weighted keyword event description is better and more semantic than other ways of utilizing concepts. (2) Our approach represents a further improvement over Bor and wBor. The performance improvement validates that our approach to the rank aggregation of concepts can efficiently remove the outlying errors of the concept ranking scores; moreover our rank aggregation method is intrinsically better than directly using the Borda count on our zero-shot video event detection problem. (3) Compared with PCF which also detects outliers in the concept ranking scores, our method is still vastly superior. This

demonstrates that efficient grounding of visual concepts, which our method achieves using weighted keyword representation, is important. We present some qualitative results in Figure 3.

4.3 The MEC Experiments

To the best of our knowledge, this is the first work on MEC; hence, the experiments are conducted on our TRECVID-MED dataset.

Experimental setup. We extract the GoogLeNet features in each frame of our experiment [49]. We utilize the following evaluation protocol: (1) Existing Metrics. We directly measure the quality of the generated video caption. The standard metrics for evaluating caption generation are used in MEC, such as BLEU [38], and METEOR [13] scores to evaluate the MEC results. These metrics are widely used in machine translation literature. (2) Object Metric. We argue that event caption represents a further generalization of video understanding and can be directly used for video classification. To this end, an objective measure – classification accuracy - is introduced here. More specifically, with the caption sentence generated for each testing instance, we compute the semantic similarity between the sentence and each event description. In particular, for a given testing video instance, we assume that the generated sentence is represented by $Y = (\cdots, y_j, \cdots,)$ where y_j is the j-th keyword of the sentence; moreover, semantic similarity is computed as follows:

$$\mathrm{w}(E_e, Y) = \frac{1}{Z}\sum_{j=1}^{|Y|} \mathrm{tfidf}(y_j)\cdot \mathrm{tfidf}(a_{i_j})\cdot \cos(\phi_{a_{i_j}}, \phi_{y_j}) \qquad (12)$$

where $i_j = \arg\max_i \{\cos(\phi_{a_{i,e}}, \phi_{y_j})\}_{a_{i,e}\in A_e}$, the normalization term is $Z = \sum_i^{|Y|} \mathrm{tfidf}(a_{i_j,e}\cdot \mathrm{tfidf}(y_i)$, and $\mathrm{tfidf}(y_j)$ indicates the TF-IDF weight of y_j of the sentences in all generated sentences. The testing instance is thus classified as the event with the highest semantic similarity.

We compare the following different methods:

(1) HENE-MSVD: we directly use the HENE video captioning model [36] trained on an existing video caption dataset – Microsoft Research Video Description Corpus (MSVD) [10] and predict the sentence captioning on our task.
(2) SL: Our approach trained from scratch on the training split of the TRECVID-MEC dataset.
(3) Aug-SL-10, Aug-SL-50, and Aug-SL-100: Our approach is trained with the augmented data, which include the training split of TRECVID-MEC and top-10, top-50, and top-100 ranking videos from auxiliary videos in TRECVID-MEC dataset respectively (i.e., top-10, top-50, top-100 ranks of $\mathbf{s}$ in Equation (11).

Experimental results. The results are presented in Table 3. Compared with all of the other competitors trained on the TRECVID-MEC dataset, the HENE-MSVD results are the worst on all metrics. Since HRNE-MSVD is trained on MSVD, this actually validates the uniqueness of our event captioning tasks, and captions of the previous video captioning dataset are very different from ours. Note also that the worst results of HENE-MSVD are not primarily caused by the domain shift of the different video datasets, since (1) both video datasets are large enough and are not targeting one specific type of visual domain; and (2) the GooLeNet features used here can serve as a type of generic features that is less sensible to the domain shift problem [42].

The results of SL are reasonably good if compared with those of HENE-MSVD. However, our Aug-SL-10 is much better than SL due to the addition of high-quality video instances added as training data. This validates the idea that our MED algorithm is able to effectively ground the visual concepts that are most important for detecting the events. Thus, with these augmented video instances, our Aug-SL-10 can almost double the classification accuracy and improve the results from 26.7% of SL to 58.8%. Nevertheless, when more instances are included, some 'noise' video instances are also used for the training set. These noise instances are less likely to be confidence to ground the video event. Consequently, the results of Aug-SL-50 and Aug-SL-100 improve the results of Accuracy by 10 absolute percentage points but achieve slightly worse performance on the METOR and BLUE metrics. Finally, an alternative means of augmenting the training set is to directly train supervised classifiers on the MEC training data. However, the results of using this method are much lower than those achieved using our method Aug-SL-100.

Table 3: Experimental results on our dataset. B@n indicates BLEU score that uses up to n-grams.

Model	MET.	B@1	B@2	B@3	B@4	Acc.
Aug-SL-10	11.1	**393.**	**26.1**	**19.8**	**15.3**	58.8
Aug-SL-50	**11.3**	36.3	24.0	18.3	14.5	68.6
Aug-SL-100	10.8	36.1	23.2	17.4	13.6	**68.7**
SL	10.0	34.5	21.4	14.7	10.3	26.7
HENE-MSVD	5.6	24.2	4.3	1.6	0.8	9.4

Table 4: Experimental results on our dataset. B@n indicates a BLEU score that uses up to n-grams.

Model	MET.	B@1	B@2	B@3	B@4	Acc.
SVM-RBF-10	10.1	35.8	23.4	17.1	12.7	54.3
Aug-SL-10	**11.1**	**39.3**	**26.1**	**19.8**	**15.3**	**58.8**
SL	10.0	34.5	21.4	14.7	10.3	26.7
HENE-VAD	5.6	24.2	4.3	1.6	0.8	9.4

We construct the supervised classifiers by using the training set of the TRECVID-MEC dataset. More specifically, we use SVM with an RBF kernel and train each event classifier with all training instances. Among the unlabeled instances, we select the top- 10 ranked videos to augment the training set in order to train the MEC tasks (i.e. SVM-RBF-10). The results are listed in Table 4.

As can be seen from the tables, the results of SVM-RBF-10 are much lower than those achieved by our method Aug-SL-10. This is due to the data sparsity and high intra-class variance of visual content. Thus the question of how best to utilize the training set of MEC to augment the unlabeled instances is a potential avenue for future work.

Some qualitative examples are illustrated in Figure 4. Comparably, our captions provide more event-related details than the HRNE-MSVD captions from Pan *et al.* [36].

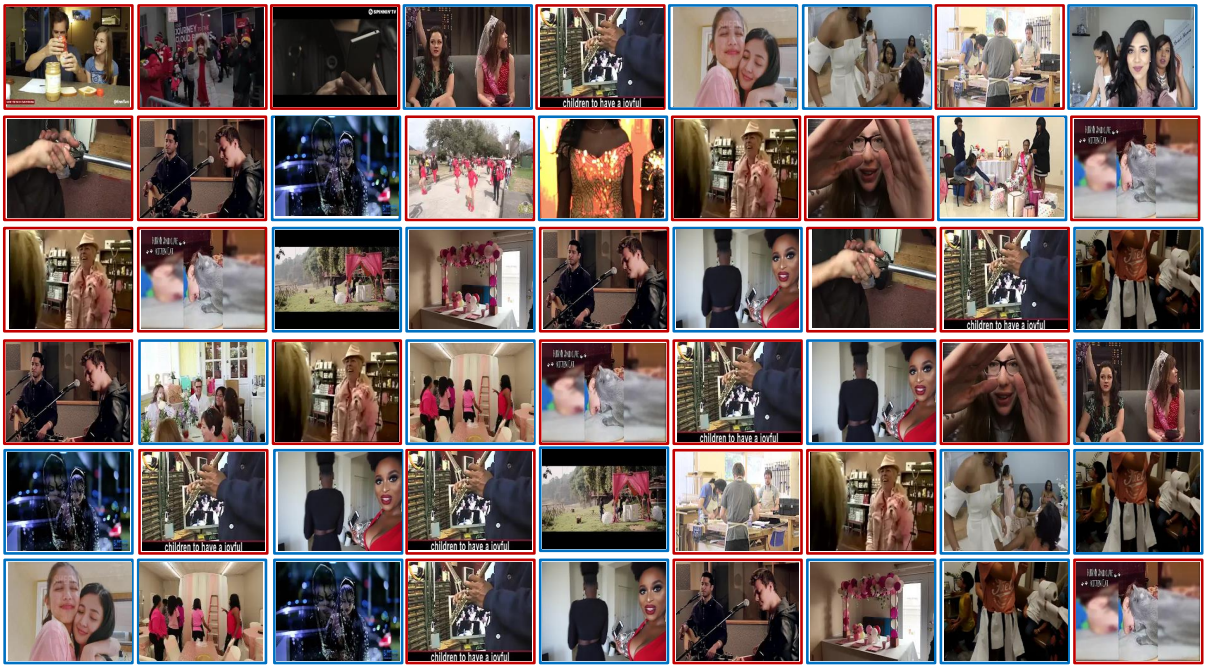


Figure 3: Top ranked videos for the event *Wedding shower*. From top to below: OR, Fu, PCF, DCC and the proposed approach.

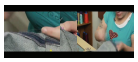
Event	Video	*Aug-SL-20*	*HRNE* (Pan et al. 2016)
E036		one or more people cut down a tree by hand or with a motorized machine	a man is walking through the woods
		one or more people cut a tree	a man is walking
		several people work together to fell a tree	a man is performing a stunt with a crowd
E011		construct sandwich from ingredients	a man is cutting a paper
		construct an edible food item from ingredients	a man is eating
		construct an edible food from one or more of bread plus fillings	a man is cutting a piece of paper
E015		several people work to construct a garment	a woman is cleaning a sheet of paper
		several people work to construct clothes by hand or or machine	a man is removing the skin from a piece of paper

Figure 4: Qualitative examples of our MEC results.

5 CONCLUSION

We propose grounding visual concepts for MED and MEC in a zero-shot setting. We organize the entire framework into a rank aggregation problem while a mean-shift outlier rejection model is used to solve the optimization and remove outliers in the process. For the first time, we introduce a multimedia event captioning task and investigate the use of our framework to argument the training instances of MEC tasks.

ACKNOWLEDGEMENTS

This work is partially supported by the National Natural Science Foundation of China (Nos. 61772322 ,U1836216), the major fundamental research project of Shandong, China (No. ZR2019ZD03), and the Taishan Scholar Project of Shandong, China(No. ts20190924), partially supported by the Air Force Research Laboratory, and partially supported by NSFC grant 61973250, 61906109 and 61702415.

REFERENCES

[1] Hrishikesh B. Aradhye, George Toderici, and Jay Yagnik. 2009. Video2Text: Learning to Annotate Video Content. In *ICDM Workshops 2009*.
[2] Ioannis Caragiannis, Xenophon Chatzigeorgiou, George A. Krimpas, and Alexandros A. Voudouris. 2019. Optimizing positional scoring rules for rank aggregation. *Artif. Intell.* 267 (2019), 58–77.
[3] Xiaojun Chang, Zhigang Ma, Yi Yang, Zhiqiang Zeng, and Alexander G. Hauptmann. 2017. Bi-Level Semantic Representation Analysis for Multimedia Event Detection. *IEEE Trans. Cybernetics* 47, 5 (2017), 1180–1197.
[4] Xiaojun Chang, Yi Yang, Alexander G. Hauptmann, Eric P. Xing, and Yaoliang Yu. 2015. Semantic Concept Discovery for Large-Scale Zero-Shot Event Detection. In *IJCAI*.
[5] Xiaojun Chang, Yi Yang, Guodong Long, Chengqi Zhang, and Alexander G. Hauptmann. 2016. Dynamic Concept Composition for Zero-Example Event Detection. In *AAAI*, Dale Schuurmans and Michael P. Wellman (Eds.).
[6] Xiaojun Chang, Yi Yang, Eric P. Xing, and Yaoliang Yu. 2015. Complex Event Detection using Semantic Saliency and Nearly-Isotonic SVM. In *ICML*, Francis R. Bach and David M. Blei (Eds.).
[7] Xiaojun Chang, Yaoliang Yu, Yi Yang, and Alexander G. Hauptmann. 2015. Searching Persuasively: Joint Event Detection and Evidence Recounting with Limited Supervision. In *ACM MM*, Xiaofang Zhou, Alan F. Smeaton, Qi Tian, Dick C. A. Bulterman, Heng Tao Shen, Ketan Mayer-Patel, and Shuicheng Yan (Eds.).
[8] Xiaojun Chang, Yaoliang Yu, Yi Yang, and Eric P. Xing. 2016. They are Not Equally Reliable: Semantic Event Search Using Differentiated Concept Classifiers. In *CVPR*.
[9] Xiaojun Chang, Yaoliang Yu, Yi Yang, and Eric P. Xing. 2017. Semantic Pooling for Complex Event Analysis in Untrimmed Videos. *IEEE Trans. Pattern Anal.*

Mach. Intell. 39, 8 (2017), 1617–1632.
[10] David L. Chen and William B. Dolan. [n.d.]. Collecting Highly Parallel Data for Paraphrase Evaluation. In *ACL*.
[11] Wenqing Chu, Hongyang Xue, Chengwei Yao, and Deng Cai. 2019. Sparse Coding Guided Spatiotemporal Feature Learning for Abnormal Event Detection in Large Videos. *IEEE Trans. Multimedia* 21, 1 (2019), 246–255.
[12] Shumin Deng, Ningyu Zhang, Jiaojian Kang, Yichi Zhang, Wei Zhang, and Huajun Chen. 2020. Meta-Learning with Dynamic-Memory-Based Prototypical Network for Few-Shot Event Detection. In *WSDM*.
[13] Michael J. Denkowski and Alon Lavie. 2014. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *WMT@ACL*.
[14] Ícaro Cavalcante Dourado, Daniel Carlos Guimarães Pedronette, and Ricardo da Silva Torres. 2019. Unsupervised graph-based rank aggregation for improved retrieval. *Inf. Process. Manage.* 56, 4 (2019), 1260–1279.
[15] Elise Epaillard and Nizar Bouguila. 2019. Variational Bayesian Learning of Generalized Dirichlet-Based Hidden Markov Models Applied to Unusual Events Detection. *IEEE Trans. Neural Netw. Learning Syst.* 30, 4 (2019), 1034–1047.
[16] Hehe Fan, Xiaojun Chang, De Cheng, Yi Yang, Dong Xu, and Alexander G. Hauptmann. 2017. Complex Event Detection by Identifying Reliable Shots from Untrimmed Videos. In *ICCV*.
[17] Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. DeViSE: A Deep Visual-Semantic Embedding Model. In *NIPS*.
[18] Yanwei Fu, Timothy M. Hospedales, Tao Xiang, and Shaogang Gong. 2012. Attribute Learning for Understanding Unstructured Social Activity. In *ECCV*.
[19] AmirHossein Habibian, Thomas Mensink, and Cees G. M. Snoek. 2014. Composite Concept Discovery for Zero-Shot Video Event Detection. In *ICMR*.
[20] Ryuhei Hamaguchi, Ken Sakurada, and Ryosuke Nakamura. 2019. Rare Event Detection Using Disentangled Representation Learning. In *CVPR*.
[21] Haiqi Huang, Yueming Lu, Fangwei Zhang, and Songlin Sun. 2012. A Multi-modal Clustering Method for Web Videos. In *ISCTCS*.
[22] Radu Tudor Ionescu, Fahad Shahbaz Khan, Mariana-Iuliana Georgescu, and Ling Shao. 2019. Object-Centric Auto-Encoders and Dummy Anomalies for Abnormal Event Detection in Video. In *CVPR*.
[23] Lu Jiang, Deyu Meng, Shoou-I Yu, Zhen-Zhong Lan, Shiguang Shan, and Alexander G. Hauptmann. [n.d.]. Self-Paced Learning with Diversity. In *NIPS*.
[24] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Fei-Fei Li. 2014. Large-Scale Video Classification with Convolutional Neural Networks. In *CVPR*.
[25] John G Kemeny. 1959. Mathematics without numbers. *Daedalus* 88, 4 (1959), 577–591.
[26] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. 2014. Attribute-Based Classification for Zero-Shot Visual Object Categorization. *IEEE Trans. Pattern Anal. Mach. Intell.* 36, 3 (2014), 453–465.
[27] Sangmin Lee, Hak Gu Kim, and Yong Man Ro. 2020. BMAN: Bidirectional Multi-Scale Aggregation Networks for Abnormal Event Detection. *IEEE Trans. Image Processing* 29 (2020), 2395–2408.
[28] Yonggang Li, Rui Ge, Yi Ji, Shengrong Gong, and Chunping Liu. 2019. Trajectory-Pooled Spatial-Temporal Architecture of Deep Convolutional Neural Networks for Video Event Detection. *IEEE Trans. Circuits Syst. Video Techn.* 29, 9 (2019), 2683–2692.
[29] Zhihui Li, Lina Yao, Xiaojun Chang, Kun Zhan, Jiande Sun, and Huaxiang Zhang. 2019. Zero-shot event detection via event-adaptive concept relevance mining. *Pattern Recognit.* 88 (2019), 595–603.
[30] Huan Liu, Qinghua Zheng, Minnan Luo, Dingwen Zhang, Xiaojun Chang, and Cheng Deng. 2017. How Unlabeled Web Videos Help Complex Event Detection?. In *IJCAI*, Carles Sierra (Ed.).
[31] Jian Liu, Yubo Chen, and Kang Liu. 2019. Exploiting the Ground-Truth: An Adversarial Imitation Based Knowledge Distillation Approach for Event Detection. In *AAAI*.
[32] Masoud Mazloom, Efstratios Gavves, Koen E. A. van de Sande, and Cees Snoek. [n.d.]. Searching informative concept banks for video event detection. In *ICMR*.
[33] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *NIPS*.
[34] Joe Yue-Hei Ng, Matthew J. Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. 2015. Beyond short snippets: Deep networks for video classification. In *CVPR*.
[35] Mark Palatucci, Dean Pomerleau, Geoffrey E. Hinton, and Tom M. Mitchell. 2009. Zero-shot Learning with Semantic Output Codes. In *NIPS*.
[36] Pingbo Pan, Zhongwen Xu, Yi Yang, Fei Wu, and Yueting Zhuang. 2016. Hierarchical Recurrent Neural Encoder for Video Representation with Application to Captioning. In *CVPR*.
[37] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. 2016. Jointly Modeling Embedding and Translation to Bridge Video and Language. In *CVPR*.
[38] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *ACL*.
[39] JC Platt. 1999. Probabilities for SV Machines, Advances in Large Margin Classifiers.
[40] Arun Rajkumar and Shivani Agarwal. 2014. A Statistical Convergence Perspective of Algorithms for Rank Aggregation from Pairwise Data. In *ICML*.
[41] Mohammad Rastegari, Ali Diba, Devi Parikh, and Ali Farhadi. 2013. Multi-attribute Queries: To Merge or Not to Merge?. In *CVPR*.
[42] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. In *CVPR Workshops*.
[43] Jorge Sánchez, Florent Perronnin, Thomas Mensink, and Jakob J. Verbeek. 2013. Image Classification with the Fisher Vector: Theory and Practice. *International Journal of Computer Vision* 105, 3 (2013), 222–245.
[44] Yiyuan She and Art B. Owen. 2010. Outlier Detection Using Nonconvex Penalized Regression. *CoRR* abs/1006.2592 (2010).
[45] Lei-Lei Shi, Lu Liu, Yan Wu, Liang Jiang, Muhammad Kazim, Haider Ali, and John Panneerselvam. 2019. Human-Centric Cyber Social Computing Model for Hot-Event Detection and Propagation. *IEEE Trans. Comput. Social Systems* 6, 5 (2019), 1042–1050.
[46] Karen Simonyan and Andrew Zisserman. 2014. Two-Stream Convolutional Networks for Action Recognition in Videos. In *NIPS*.
[47] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *CoRR* abs/1212.0402 (2012).
[48] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. [n.d.]. Sequence to Sequence Learning with Neural Networks. In *NIPS*.
[49] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *CVPR*.
[50] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2015. The New Data and New Challenges in Multimedia Research. *CoRR* abs/1503.01817 (2015).
[51] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond J. Mooney, Trevor Darrell, and Kate Saenko. 2015. Sequence to Sequence - Video to Text. In *ICCV*.
[52] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond J. Mooney, and Kate Saenko. 2015. Translating Videos to Natural Language Using Deep Recurrent Neural Networks. In *NAACL*.
[53] Zuxuan Wu, Yanwei Fu, Yu-Gang Jiang, and Leonid Sigal. 2016. Harnessing Object and Scene Semantics for Large-Scale Video Understanding. In *CVPR*.
[54] Xianjun Xia, Roberto Togneri, Ferdous Sohel, and Defeng Huang. 2019. Auxiliary Classifier Generative Adversarial Network With Soft Labels in Imbalanced Acoustic Event Detection. *IEEE Trans. Multimedia* 21, 6 (2019), 1359–1371.
[55] Haoran Yan, Xiaolong Jin, Xiangbin Meng, Jiafeng Guo, and Xueqi Cheng. 2019. Event Detection with Multi-Order Graph Convolution and Aggregated Attention. In *EMNLP-IJCNLP*.
[56] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher J. Pal, Hugo Larochelle, and Aaron C. Courville. 2015. Describing Videos by Exploiting Temporal Structure. In *ICCV*.
[57] Shoou-I Yu, Lu Jiang, and Alexander G. Hauptmann. 2014. Instructional Videos for Unsupervised Harvesting and Learning of Action Examples. In *ACM MM*.
[58] Dingwen Zhang, Junwei Han, Lu Jiang, Senmao Ye, and Xiaojun Chang. 2017. Revealing Event Saliency in Unconstrained Video Collection. *IEEE Trans. Image Processing* 26, 4 (2017), 1746–1758.
[59] Hao Zhang and Chong-Wah Ngo. 2019. A Fine Granularity Object-Level Representation for Event Detection and Recounting. *IEEE Trans. Multimedia* 21, 6 (2019), 1450–1463.
[60] Lingling Zhang, Jun Liu, Minnan Luo, Xiaojun Chang, and Qinghua Zheng. 2018. Deep Semisupervised Zero-Shot Learning with Maximum Mean Discrepancy. *Neural Computation* 30, 5 (2018).
[61] Yu Zhao, Zhenhui Shi, Jingyang Zhang, Dong Chen, and Lixu Gu. 2019. A novel active learning framework for classification: Using weighted rank aggregation to achieve multiple query criteria. *Pattern Recognition* 93 (2019), 581–602.
[62] Zhicheng Zhao, Xuanchong Li, Xingzhong Du, Qi Chen, Yanyun Zhao, Fei Su, Xiaojun Chang, and Alexander G. Hauptmann. 2018. A unified framework with a benchmark dataset for surveillance event detection. *Neurocomputing* 278 (2018), 62–74.
[63] Yuke Zhu, Oliver Groth, Michael S. Bernstein, and Li Fei-Fei. 2016. Visual7W: Grounded Question Answering in Images. In *CVPR*.