Ontology-driven Event Type Classification in Images

Eric Müller-Budack[1], Matthias Springstein[1], Sherzod Hakimov[1], Kevin Mrutzek[2], Ralph Ewerth[1,2]
[1]TIB - Leibniz Information Centre for Science and Technology, Hannover, Germany
[2]Leibniz University Hannover, L3S Research Center, Hannover, Germany
{eric.mueller, matthias.springstein, sherzod.hakimov, ralph.ewerth}@tib.eu

Abstract

Event classification can add valuable information for semantic search and the increasingly important topic of fact validation in news. So far, only few approaches address image classification for newsworthy event types such as natural disasters, sports events, or elections. Previous work distinguishes only between a limited number of event types and relies on rather small datasets for training. In this paper, we present a novel ontology-driven approach for the classification of event types in images. We leverage a large number of real-world news events to pursue two objectives: First, we create an ontology based on Wikidata comprising the majority of event types. Second, we introduce a novel large-scale dataset that was acquired through Web crawling. Several baselines are proposed including an ontology-driven learning approach that aims to exploit structured information of a knowledge graph to learn relevant event relations using deep neural networks. Experimental results on existing as well as novel benchmark datasets demonstrate the superiority of the proposed ontology-driven approach.

1. Introduction

Digital media and social media platforms such as *Twitter* have become a popular resource to provide news and information. To handle the sheer amount of daily published articles in the Web, automated solutions to understand the multimedia content are required. The computer vision community has focused on many visual classification tasks such as object recognition [18, 19, 20, 23, 42], place (scene) classification [41], or geolocation estimation [26, 30, 34, 37] to enable semantic search or retrieval in archives and news collections. But news typically focus on events with a high significance for a target audience. Thus, event classification in images is an important task for various applications. Multimedia approaches [22, 27, 29] have exploited visual descriptors to quantify image-text relations that can help to understand the overall multimodal message and sentiment or might even indicate misinformation, i.e., *Fake News*.

Despite its clear potential, so far only few approaches [6, 11, 21, 24, 39] were proposed for the classification of real-world event types. Datasets for event classification mostly cover only specific event categories, e.g., social [2, 6, 28], sports [24], or cultural events [13]. To the best of our knowledge, the *Web Image Dataset for Event Recognition (WIDER)* [39] is the largest corpus with 50,574 images that considers a variety of event types (61). Nonetheless, many types that are important for news, like *epidemics* or *natural disasters*, are missing. Due to the absence of large-scale datasets, related work has focused on ensemble approaches [4, 5, 36] typically based on pre-trained models for object and place (scene) classification and the integration of descriptors from local image regions [3, 14, 17, 39] to learn rich features for event classification. We believe that one of the main challenges is to define a complete lexicon of important event categories. For this purpose, Ahsan *et al*. [6] suggest to mine *Wikipedia* and gathered 150 generic social events. However, the experiments were only conducted on *WIDER* as well as on two datasets, which cover eight social event types and a selection of 21 real-world events. Progress in the field of Semantic Web has shown that it is possible to define a knowledge graph for newsworthy events [15, 16] but has not been leveraged by computer vision approaches yet. Particularly the relations between events extracted from a knowledge base such as *Wikidata* [35] provide valuable information that can be utilized to train powerful models for event classification.

In this paper, we introduce a novel ontology along with a dataset that enable us to develop a novel ontology-driven deep learning approach for event classification. Our **primary contributions** can be summarized as follows: (1) Based on a set of real-world events from *EventKG* [15, 16], we propose a *Visual Event Ontology (VisE-O)* containing 409 nodes describing 148 unique event types such as different kinds of sports, disasters, and social events with high news potential that can be created with little supervision. It covers the largest number of event types for image classification to date. (2) In order to train deep learning models, we have gathered a large-scale dataset, called

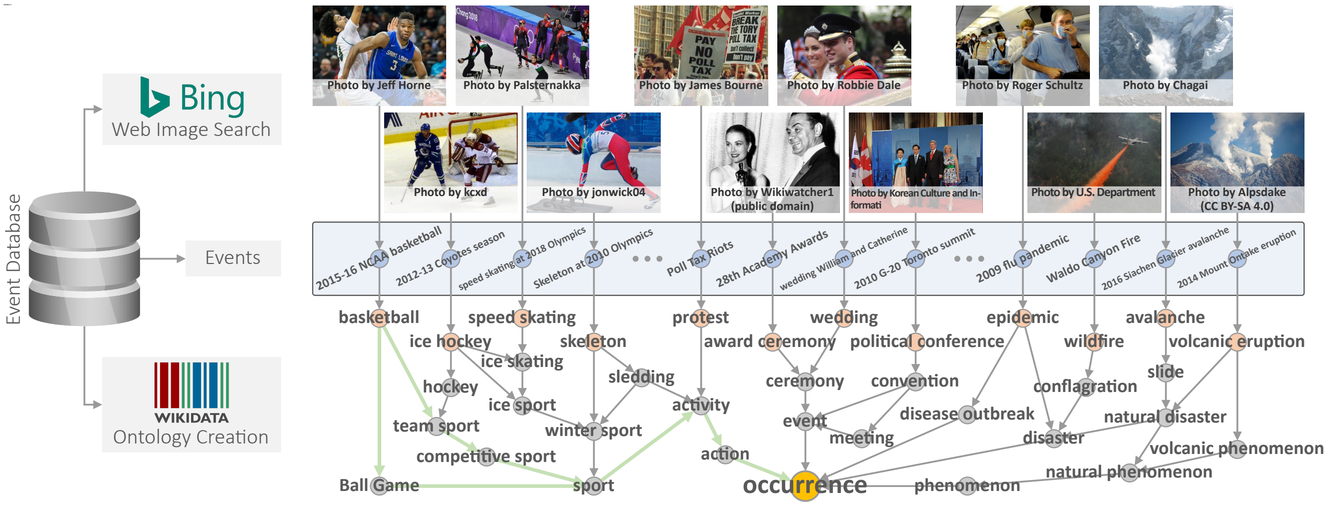

Figure 1. Exemplary subset of the *Ontology* (complete version is provided on our *GitHub* page[1]) and images of the proposed *Visual Event Classification Dataset (VECD)*. *Leaf Event Nodes* (orange) and *Branch Event Nodes* (gray) are extracted based on relations (e.g., *"subclass of"*) to a set of *Events* (blue) using the *Wikidata* knowledge base. The nodes connected by the green path define the *Subgraph* of *basketball* to the *Root Node* (yellow). The combination (union) of all *Subgraphs* defines the *Ontology*. Definitions are according to Section 3.1.

Visual Event Classification Dataset (VisE-D), of 570,540 images crawled automatically from the Web. It contains 531,080 training and 28,543 validation images as well as two test sets with 2,779 manual annotated and 8,138 *Wikimedia* images. Figure 1 depicts some example images. (3) We provide several baselines including an ontology-driven deep learning approach that integrates the relations of event types extracted from structured information in the ontology to understand the fundamental differences of event types in different domains such as *sports*, *crimes*, or *natural disasters*. Experimental results on several benchmark datasets demonstrate the feasibility of the proposed approach. Dataset and source code are publicly available.[1]

The remainder of this paper is organized as follows. In Section 2 we review related work. The ontology and dataset for newsworthy event types is presented in Section 3. In Section 4 we propose an ontology-driven deep learning approach for event classification. Experimental results for several benchmarks are presented in Section 5. Section 6 summarizes the paper and outlines areas of future work.

2. Related Work

Since there are different definitions of an event, approaches for event classification are diverse and range from specific actions in videos [33, 40] over the classification of more personal events in photo collections [10, 11, 38] to the classification of social, cultural, and sport events in photos [17, 24, 36, 39]. In the sequel, we mainly focus on works and datasets for the recognition of events and event types in images with potential news character.

Early approaches for event classification have used handcrafted features such as *SIFT* (Scale-Invariant Feature Transform) to classify events in particular domains like sports [21, 24]. As one of the first deep learning approaches Xiong *et al*. [39] trained a multi-layer framework that leverages two convolutional neural networks to incorporate the visual appearance of the whole image as well as interactions among humans and objects. Similarly, several approaches integrated local information from image patches or regions extracted by object detection frameworks [3, 14, 17] to learn rich features for event classification. In this respect, Guo *et al*. [17] proposed a graph convolutional neural network to leverage relations between objects. Another kind of approaches applies ensemble models and feature combination [4, 5, 36] to exploit the capabilities of deep learning models trained for different computer vision tasks, most typically for object recognition and scene classification. In the absence of a large-scale dataset for many event types, Ahsan *et al*. [6] suggest to train classifiers based on images crawled for a set of 150 social event concepts mined from *Wikipedia*, while Wang *et al*. [36] apply transfer learning to object and scene representations to learn compact representations for event recognition with few training images. For a more detailed review of deep learning techniques for event classification, we refer to Ahmad and Conci's survey [1].

There are many datasets and also challenges such as the *MediaEval Social Event Detection Task* [28] and *ChaLearn Looking at People* [13] for event classification. But they mostly cover specific domains such as social events [2, 28],

[1] Our project is available on the *EventKG* website and on *GitHub*:
EventKG: `http://eventkg.l3s.uni-hannover.de/VisE`
GitHub: `https://github.com/TIBHannover/VisE`

cultural events [13], or sports [24]. In addition, the datasets are either too small [24] to train deep learning models or contain very few event classes [2]. Other proposals have introduced datasets and approaches to detect concrete real-world news events [6, 13, 14], but only distinguish between a small predefined selection. To the best of our knowledge, *WIDER (Web Image Dataset for Event Recognition)* [39] is the most complete dataset in terms of the number of event categories that can be leveraged by deep learning approaches. It contains $50{,}574$ images for 61 event types. But many important event types for news such as *epidemics* or *natural disasters* are missing.

3. Ontology and Dataset

In contrast to prior work, this section presents an ontology and dataset for event classification that covers a larger number of event types with news character across all domains such as *sports*, *crimes*, and *natural disasters*. Based on definitions for terms and notations (Section 3.1), we suggest an approach that leverages events identified by *EventKG* [15, 16] to automatically retrieve an ontology that can be refined with little supervision (Section 3.2). Images for event types in the resulting *Visual Event Ontology (VisE-O)* are crawled from the Web to create the *Visual Event Classification Dataset (VisE-D)* according to Section 3.3.

3.1. Definitions and Notations

In this section, we introduce definitions and notations that are used in the remainder of the paper. Figure 1 contains supplementary visualizations to clarify the definitions.

Event: As in the *EventKG* [15], we define a set E of contemporary and historical events of global importance (e.g., *2011 NBA Finals* in Figure 1) in this paper.

Ontology, Root Node, Event Node, and Relation: The *Ontology* is a directed graph composed by a set of *Event Nodes* N and their corresponding *Relations* R as edges. *Relations* R are knowledge base specific properties such as *"subclass of"* in *Wikidata* that describe the interrelations of *Event Nodes* N. All parent nodes $n \in N$ that connect a specific *Event* $e \in E$ to the *Root Node* are denoted as *Event Nodes*. The *Root Node* $n_R \in N$ (e.g., *occurrence* in Figure 1) matches the overall definition of an *Event* and represents a parent node that is shared by all *Events*.

Leaf and Branch Event Node: The *Leaf Event Nodes* $N_L \subset N$ such as *basketball* are the most detailed *Event Nodes* without children in the *Ontology*. They group *Events* of the same type, e.g., *2011 NBA Finals* → *basketball* (Figure 1). *Event Nodes*, e.g., *ball game* with at least one child node are referred to as *Branch Event Nodes* $N_B \subset N$.

Subgraph: A *Subgraph* S_L is a set of all *Event Nodes* $S_L = \{n_L, \ldots, n_R\} \subset N$ that relate to a specified *Leaf Event Node* $n_L \in N_L$ while traversing to the *Root Node* n_R.

3.2. VisE-O: Visual Event Ontology

3.2.1 Knowledge Base and Root Node Selection

Several knowledge bases such as *DBpedia* [9], *YAGO* [31], or *Wikidata* [35] are available. We investigated them in terms of event granularity and correctness. At this time, the whole *DBpedia* ontology contains less than $1{,}000$ classes. Thus, the granularity of potential event types is very coarse and for instance some types of natural disasters are either assigned to wrong (*Tsunami* → *television show*) [8] or generic classes (*Earthquake* → *thing*) [7]. As mentioned by Gottschalk and Demidova [15], *YAGO* also contains noisy event categories. On the contrary, *Wikidata* offers fine-granular event types and relations, as shown in Figure 2, and is therefore used as knowledge base in this work. We have selected *occurrence (Q1190554)* as the *Root Node* of the *Ontology* since it matches our definition of an *Event*.

3.2.2 Initial Event Ontology

In this paper, a *bottom-up approach* is applied to automatically create an event ontology. Based on a large set of $|E| = 550{,}994$ real-world events from *EventKG* [15, 16], we recursively obtain all parent *Event Nodes* from *Wikidata*. For *Event Nodes* only relations of the type *"subclass of" (P279)* are considered since they already describe specific categories. For *Events* we additionally allow the properties *"instance of" (P31)* and *"part of" (P361)* as possible relations to increase the coverage, because some events like *2018 FIFA World Cup Group A* are not a *"subclass of"* an *Event Node* but *"part of"* a superordinate event, in this case *2018 FIFA World Cup*. Finally, we remove all *Event Nodes* that are not connected to the *Root Node*. As illustrated in Figure 1, the resulting *Subgraphs* define the *Ontology*.

However, we identified several problems in the initial *Ontology* as illustrated in Figure 2. (1) There are differences in the granularity and some of the fine-grained *Leaf Event Nodes*, e.g., *ATP tennis tournament* or *Nepalese local election*, might be hard to recognize; (2) In particular, sports-centric *Leaf Event Nodes* such as *association football match* and *association football team season* are ambiguous; (3) Some *Event Nodes*, e.g., *software license* do not represent an *Event* according to the definition in Section 3.1.

3.2.3 Event Class Disambiguation

As pointed out in the previous section, most *Leaf Event Nodes* related to sports are visually ambiguous since they represent the same type of sport. The *Wikidata* knowledge base distinguishes between *sports seasons*, *sports competitions*, etc. Although this structure might make sense for some applications, we aim to combine *Event Nodes* that relate to the same sports type. Unfortunately, this is not possible with the initial *Ontology* that relies on *Relations* of the

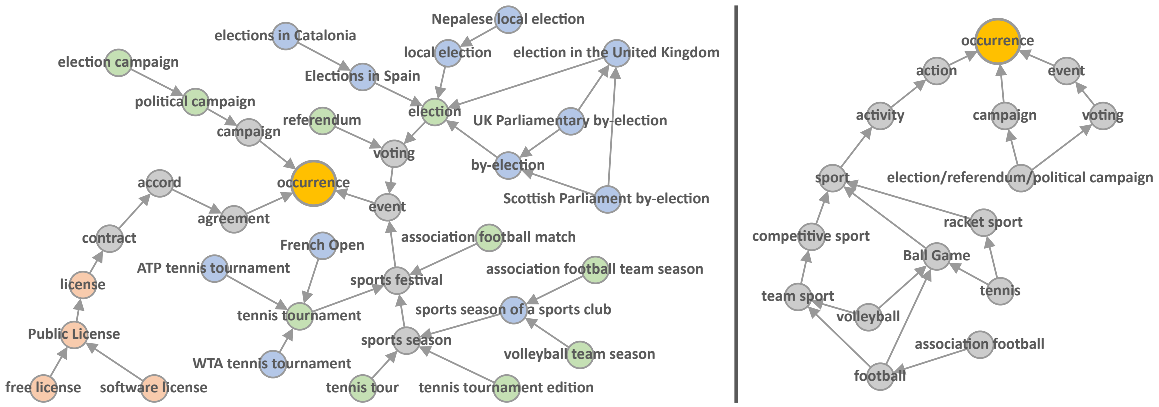

Figure 2. Exemplary subset of the initial *Ontology* after the extraction of all relations from *Wikidata* (left) and respective final *Ontology* after applying the proposed approaches for event class disambiguation and refinement (right). Blue *Event Nodes* might be too fine-granular. Green nodes are semantically and visually similar to other *Event Nodes* in the *Ontology*. Orange nodes do not represent an *Event* according to the definition in Section 3.1. Best viewed in color. Different versions of the ontologies can be explored on our *GitHub* page[1].

type *"subclass of"*. As illustrated in Figure 2 (green nodes), *Event Nodes* of different sports domains (e.g., *volleyball team season* and *association football team season*) relate to a particular type of competition (in this case *team season*) before they relate to another *Event Node* of the same sports type (*association football match*). In order to solve this issue, value(s) for the *Wikidata* property *"sport" (P641)* (if available) for each *Event* and *Event Node* were extracted and used as *Relation*. As a result, sports events were combined according to their sports category rather than the type of the competition as shown in Figure 2 (right). In addition, we delete all *Event Nodes* that are a parent of less than a minimum number of $|E|_{min} = 10$ *Events* to reduce the granularity of the resulting *Leaf Event Nodes*.

These strategies lead to an *Ontology* that is more appropriate for computer vision tasks. However, it can still contain irrelevant *Event Nodes*. Furthermore, scheduled events such as *elections* or *sports festivals* occur more frequently than unexpected or rare events such as *epidemics* or *natural disasters*. Therefore, *Leaf Event Nodes* that represent scheduled event types more likely fulfill the filtering criteria $|E|_{min}$ and are consequently very fine-grained (e.g., elections in different countries) making them hard to distinguish. Thus, we decided to manually refine the *Ontology*.

3.2.4 Event Ontology Refinement

Two co-authors were asked to manually refine the *Ontology* to create a challenging yet useful and fair *Ontology* for image classification. To pursue this goal, the *Ontology* was refined according to two criteria: (1) reject *Event Nodes* that do not match the *Event* definition in Section 3.1 and (2) select the most suitable *Leaf Event Nodes* to prevent ambiguities. For example, *election* was chosen as a representative *Leaf Event Node* since its children contain different types of elections (e.g., *by-election*) and elections in different countries (e.g., *elections in Spain*) that might be too hard to distinguish. As we can use the hierarchical information to automatically assign the children to the selected *Leaf Event Nodes* and simultaneously remove all resulting *Branch Event Nodes* as candidates, only around 500 annotations were necessary to label all (2,288) *Event Nodes*. Finally, we manually merged 30 *Leaf Event Nodes* such as *award* and *award ceremony* that are semantically similar but could not be fused using the *Ontology*.

Ontology	Ontology Statistics					Dataset Statistics			
	$\|E\|$	$\|\hat{E}\|$	$\|N\|$	$\|N_L\|$	$\|R\|$	$\|I_T\|$	$\|I_V\|$	$\|I_B\|$	$\|I_W\|$
Initial	527k	236k	6,114	3,578	7,545	—	—	—	—
Disamb.	530k	164k	2,288	1,081	3,144	—	—	—	—
Refined	530k	447k	409	148	635	531k	29k	2,779	8,138

Table 1. Number of *Event Nodes* $|N|$, *Leaf Event Nodes* $|N_L|$, *Relations* $|R|$, and images $|I|$ for training (T), validation (V) and test (B - *VisE-Bing*, W - *VisE-Wiki*). $|E|$ is the number of *Events* that relate to any *Event Node* in the *Ontology*, and $|\hat{E}|$ the number of *Events* that can be linked unambiguously to a *Leaf Event Node*.

The statistics for all variants of the *Ontology* are shown in Table 1 and reveal that the refined *Ontology* is able to link the most *Events* to *Leaf Event Nodes*. In the preliminary *Ontologies*, many *Events* are children of *Branch Event Nodes* and it is not possible to use them to query example images for *Leaf Event Node* as explained in the next section.

3.3. VisE-D: Visual Event Classification Dataset

Data Collection: To create a large-scale dataset for the proposed *Ontology* we defined different queries to crawl representative images from *Bing*. A maximum of 1,000 images (500 without restrictions and another 500 uploaded within the last year) using the names of the *Leaf Event*

Nodes were crawled. In addition, the names of popular *Events* related to a *Leaf Event Node* that happened after 1900 were used as queries to increase the number of images and reduce ambiguities (e.g., *Skeleton at the 2018 Winter Olympics* for *Skeleton* in Figure 1). In this regard, a sampling strategy (details are provided on our *GitHub* page[1]) was applied to set the number of images downloaded for an *Event* based on its popularity (number of *Wikipedia* page views) and date to prevent spam in the search results.

Ground-truth Labels: We provide two ground-truth vectors for each image based on the search query. (1) The ***Leaf Node Vector*** $\mathbf{y}_L \in \{0,1\}^{|N_L|}$ indicates which of the $|N_L| = 148$ *Leaf Event Nodes* are related to the image, and serves for classification tasks without using *Ontology* information. Note that $\mathbf{y}_L$ is multi-hot encoded as a queried *Event* (e.g, *SpaceX Lunar Tourism Mission* → *spaceflight* and *expedition*) can relate to multiple *Leaf Event Nodes*. (2) The multi-hot encoded ***Subgraph Vector*** $\mathbf{y}_S \in \{0,1\}^{|N|}$ denotes which of the $|N| = 409$ *Event Nodes* (*Leaf* and *Branch*) are in the *Subgraphs* of all related *Leaf Event Nodes* and allows to learn from *Ontology* information.

Splits: We were able to download about 588,000 images, which are divided into three splits for training (90%), validation (5%), and test (5%). For the test set we only use images from *Events* that relate to exactly one *Leaf Event Node*. Test images that are a duplicate (using the image hash) of a training or validation image are removed.

VisE-Bing Test Set: Two co-authors verified whether or not a test image depicts the respective *Leaf Event Node*. Each co-author annotated a maximum of ten valid images for each *Leaf Event Node* to prevent bias. They received different sets to increase the number of images. We obtained 2,779 verified test images, with 20 images for most (109) of the 148 *Leaf Event Nodes*. The dataset statistics are reported in Table 1 and in the supplemental material on *GitHub*[1].

VisE-Wiki Test Set: To create another larger test set, we downloaded all *Wikimedia* images for each *Leaf Event Node* and its child *Events* using the *Commons category (P373)* linked in *Wikidata*. Despite *Wikimedia* is a trusted source, we noticed some less relevant images for news, e.g., historic drawings or scans. We applied a k-nearest-neighbor classifier based on the embeddings of a *ResNet-50* [18] trained on *ImageNet* [12]. For each test image in *VisE-Bing*, we selected the $k = 100/|I_a^n|$ nearest images, where $|I_a^n|$ is the number of annotated images of the *Leaf Event Node* $n \in N_L$ in *VisE-Bing*. The test set comprises 8,138 images for 146 of 148 classes (statistics available on *GitHub*[1]).

4. Event Classification

In this section, we propose a baseline classification approach (Section 4.1) and more advanced strategies as well as weighting schemes to integrate event type relations from the *Ontology* in the network training (Section 4.2).

4.1. Classification Approach

As shown in Table 1 the refined *Ontology* contains $|N_L| = 148$ *Leaf Event Nodes*. As a baseline classifier, we train a convolutional neural network that predicts *Leaf Event Nodes* without using ontology information. The *Leaf Node Vector* $\mathbf{y}_L = (y_L^1, \ldots, y_L^{|N_L|})$ from Section 3.3 is used as target for optimization. During training the cross-entropy loss L_c based on the sigmoid activations $\hat{\mathbf{y}}_L$ of the last fully-connected layer is optimized:

$$L_c = -\sum\nolimits_{i=1}^{|N_L|} y_L^i \cdot \log \hat{y}_L^i \tag{1}$$

4.2. Integration of Ontology Information

In order to integrate information from the proposed *Ontology* in Section 3.2, we use the multi-hot encoded *Subgraph Vector* $\mathbf{y}_S = (y_S^1, \ldots, y_S^{|N|})$ introduced in Section 3.3 that includes the relations to all $|N| = 409$ *Event Nodes* as a target. We consider two different loss functions. As for the classification approach, we apply the cross-entropy loss on the sigmoid activations $\hat{\mathbf{y}}_S$ of last fully-connected layer to define an ontology-driven loss function:

$$L_o^{cel} = -\sum\nolimits_{i=1}^{|N|} y_S^i \cdot \log \hat{y}_S^i \tag{2}$$

As an alternative, we minimize the cosine distance of the predicted $\hat{\mathbf{y}}_S$ and the ground truth $\mathbf{y}_S$ *Subgraph Vectors*:

$$L_o^{cos} = 1 - \frac{\mathbf{y}_S \cdot \hat{\mathbf{y}}_S}{\|\mathbf{y}_S\|_2 \, \|\hat{\mathbf{y}}_S\|_2} \tag{3}$$

The granularity and the number of *Event Nodes* within the *Subgraphs* of *Leaf Event Nodes* varies for different domains, e.g., *sports*, *elections*, or *natural disasters*. As a consequence, the loss might be difficult to optimize. In addition, *Branch Event Nodes* such as *action* or *process* represent general concepts that are shared by many *Leaf Event Nodes*. Some *Branch Event Nodes* are also redundant since they do not include more *Leaf Event Nodes* as their children.

4.2.1 Redundancy Removal

We delete every *Branch Event Node* that relates to the same set of *Leaf Event Nodes* compared to its child nodes in the *Ontology*. These nodes are redundant since they do not include any new relationship information. As a result, we are able to reduce the size of the *Subgraph Vector* $\mathbf{y}_S \in \{0,1\}^{|N|}$ from $|N| = 409$ to $|N_{RR}| = 245$.

4.2.2 Node Weighting

To encourage the neural network to focus on *Leaf Event Nodes* and more informative *Branch Event Nodes* in the *Ontology*, we investigated two weighting schemes. Based on *one* of the schemes, each entry in the ground-truth $\mathbf{y}_S$ and predicted $\hat{\mathbf{y}}_S$ *Subgraph Vectors* is multiplied with its corresponding weight before the loss according to Equation (2) or (3) is calculated.

We propose a **Distance Weight** γ^n based on the distance of an *Event Node* $n \in N$ to all connected *Leaf Event Nodes* in the *Ontology*. First, the length l^n of the shortest path including self loops (a node is always in its own path $l^n > 0$) to each connected *Leaf Event Node* is determined. The average length $\overline{l^n}$ of these paths is used to calculate the weight:

$$\gamma^n = \frac{1}{2^{\left(\overline{l^n}-1\right)}} \; . \tag{4}$$

This weighting scheme encourages the network to learn from *Event Nodes* that are close to the *Leaf Event Nodes*. They describe more detailed event types which are harder to distinguish. Please note, that the average length $\overline{l^n}$ changes if the redundancy removal (Section 4.2.1) is applied.

Similarly, we calculate a **Degree of Centrality Weight** ω_n for each *Event Node* $n \in N$ based on the number c^n of *Leaf Event Nodes* connected to an *Event Node* n and the total number of *Leaf Event Nodes* $|N_L| = 148$:

$$\omega^n = 1 - \frac{c^n - 1}{|N_L|} \; . \tag{5}$$

According to Equation (5) the weights of all *Leaf Event Nodes* are set to $\omega^n = 1, \forall n \in N_L$ (denoted as ω_L), while, for instance, the *Root Node* n_R is weighted with $\omega^{n_R} \approx 0$ because it is connected to all *Leaf Event Nodes*. Thus, the network should focus on learning unique event types such as *tsunami* or *carnival* rather than coarse superclasses that relate to many *Leaf Event Nodes*. While the maximum weight of *Branch Event Nodes* using the *Distance Weights* is 0.5 and defined by the nodes closest to the *Leaf Event Nodes* ($\overline{l^n} = 2$), their corresponding *Degree of Centrality Weight* can be close to ω_L. To put more emphasis on *Leaf Event Nodes*, we set their weights to $\omega_L > 1$. We set these weights to $\omega_L = 6$ as discussed in detail in Section 5.3.1.

4.2.3 Inference

The classification approach predicts a *Leaf Node Vector* $\hat{\mathbf{y}}_L$ that contains the probabilities of the $|N_L| = 148$ *Leaf Event Nodes* that can be directly used for event classification. On the other hand, the ontology-driven network outputs a *Subgraph Vector* $\hat{\mathbf{y}}_S$ with probabilities for all $|N| = 409$ or $|N_{RR}| = 245$ (with redundancy removal) *Event Nodes* in the *Ontology*. There are several options to retrieve a *Leaf Node Vector* $\hat{\mathbf{y}}_L$ for classification using $\hat{\mathbf{y}}_S$.

(1) We retrieve the probabilities $\hat{\mathbf{y}}_L^o$ that are part of the predicted *Subgraph Vector* $\hat{\mathbf{y}}_S$. (2) Similar to Equation (3), the cosine similarity of the predicted *Subgraph Vector* $\hat{\mathbf{y}}_S$ to the multi-hot encoded *Subgraph Vector* $\mathbf{y}_S^n$ of each *Leaf Event Node* $n \in N_L$ is measured to leverage the probabilities of *Branch Event Nodes*. Note that the ground truth and predicted *Subgraph Vectors* are first multiplied with the used weights during network training. As a result, we obtain $|N_L| = 148$ similarities that are stored as $\hat{\mathbf{y}}_L^{cos} \in \mathbb{R}^{|N_L|}$.

The elementwise product $\hat{\mathbf{y}}_L = \hat{\mathbf{y}}_L^o \odot \hat{\mathbf{y}}_L^{cos}$ is used as prediction for the ontology approach, we found that this combination worked best in most cases. Results using the individual probabilities can be found on our *GitHub* page[1].

5. Experimental Setup and Results

In this section, the utilized network architecture and parameters (Section 5.1), evaluation metrics (Section 5.2) as well as experimental results (Section 5.3) are presented.

5.1. Network Parameters

We used a *ResNet-50* [18] as the basic architecture for the proposed approaches. They were optimized using *Stochastic Gradient Descent (SGD)* with *Nesterov momentum* term [32], weight decay of 1×10^{-5}, and a batch size of 128 images. To speed-up the training, the initial learning rate of 0.01 is increased to 0.1 using a linear ramp up in the first $10{,}000$ iterations. Then, a cosine learning rate annealing [25] is applied to lower the learning rate to zero after a total of $100{,}000$ iterations. The model that achieves the lowest loss on the validation set is used for the experiments.

5.2. Evaluation Metrics

We report the top-1, top-3, and top-5 accuracy using the top-k predictions in the *Leaf Node Vector* $\hat{\mathbf{y}}_L$ (Section 4.2.3). But the accuracy does not reflect the similarity of the predicted to the ground-truth *Leaf Event Node* with respect to the *Ontology* information. For this reason, we create a multi-hot encoded *Subgraph Vector* $\mathbf{y}_{\hat{S}} \in \{0,1\}^{|N|}$ representing the whole *Subgraph* $\hat{S}$ of the predicted (top-1) *Leaf Event Node* $\hat{n}$. Note, that the full *Subgraph Vector* with dimension $|N| = 409$ is created to generate comparable results for models trained with and without redundancy removal. We propose to measure the cosine similarity (*CS*; similar to Equation (3)) and *Jaccard Similarity Coefficient* (*JSC*; Equation (6)) between $\mathbf{y}_{\hat{S}}$ and the ground-truth *Subgraph Vector* $\mathbf{y}_S$ of the test image to quantify the similarity based on all $|N| = 409$ *Event Nodes*:

$$JSC = \frac{\left\|\mathbf{y}_S \odot \mathbf{y}_{\hat{S}}\right\|_1}{\left\|\mathbf{y}_S\right\|_1 \cdot \left\|\mathbf{y}_{\hat{S}}\right\|_1 \cdot \left\|\mathbf{y}_S \odot \mathbf{y}_{\hat{S}}\right\|_1} \tag{6}$$

5.3. Experimental Results

In this section, the results of our proposed approaches are presented. It includes a comparison of the ontology-driven approaches to the classification baseline (Section 5.3.1), an analysis of results for specific event types (Section 5.3.2), and an evaluation on other benchmarks (Section 5.3.3).

	Loss	WS	RR	Accuracy			*JSC*	*CS*
				Top1	***Top3***	***Top5***		
C	L_c			77.4	89.8	93.6	84.7	87.7
O^{cel}	L_o			67.5	83.3	88.5	81.1	85.4
O^{cel}_{ω}	L^{cel}_o	$\omega, \omega_L = 1$		68.1	83.7	88.9	81.1	85.3
$O^{cel}_{6\omega}$	L^{cel}_o	$\omega, \omega_L = 6$		79.8	91.0	94.0	86.6	89.2
$O^{cel}_{6\omega}$+RR	L^{cel}_o	$\omega, \omega_L = 6$	✓	81.7	91.5	**94.5**	**87.9**	90.3
O^{cel}_{γ}	L^{cel}_o	γ		66.6	83.5	89.1	78.3	82.8
O^{cel}_{γ}+RR	L^{cel}_o	γ	✓	73.2	86.8	91.3	82.6	86.2
O^{cos}	L^{cos}_o			67.6	77.8	81.8	82.6	86.7
O^{cos}_{ω}	L^{cos}_o	$\omega, \omega_L = 1$		72.7	84.1	87.2	84.5	87.9
$O^{cos}_{6\omega}$	L^{cos}_o	$\omega, \omega_L = 6$		80.2	90.6	93.4	86.3	88.9
$O^{cos}_{6\omega}$+RR	L^{cos}_o	$\omega, \omega_L = 6$	✓	80.8	90.1	93.1	86.9	89.4
O^{cos}_{γ}	L^{cos}_o	γ		81.1	90.2	93.1	87.1	89.7
O^{cos}_{γ}+RR	L^{cos}_o	γ	✓	80.7	90.3	93.1	86.9	89.5
$CO^{cel}_{6\omega}$+RR	$L_c + L^{cel}_o$	$\omega, \omega_L = 6$	✓	81.5	**91.8**	94.3	87.5	90.0
CO^{cos}_{γ}	$L_c + L^{cos}_o$	γ		**81.9**	90.8	93.2	**87.9**	**90.4**

Table 2. Results on *VisE-Bing* using different loss functions, weighting schemes (WS), and ontology redundancy removal (RR).

5.3.1 Ablation Study

The results of the proposed approaches on *VisE-Bing* are presented in Table 2. The performances of the ontology-driven approaches are significantly worse without applying any weighting scheme, because the correct prediction of the majority of *Event Nodes* in a *Subgraph* is already sufficient to achieve low loss signals. However, the ontology-driven approaches greatly benefit from the weighting schemes and clearly outperform the classification baseline. As discussed in Section 4.2.2, a higher weight ω_L for *Leaf Event Nodes* needs to be assigned using the *Degree of Centrality Weights* to balance the impact of *Branch* and *Leaf Event Nodes* on the overall loss. Thus, we increased the weight to $\omega_L = 6$ as it approximately corresponds to the average number of *Branch Event Nodes* in all $|N_L| = 148$ *Subgraphs*.

Both loss functions L^{cel}_o and L^{cos}_o achieve similar results in their best setups. Models trained with L^{cos}_o work well with both weighting schemes, while models optimized with L^{cel}_o are better with the *Degree of Centrality Weight*. We believe they are more tailored towards single-label classification tasks and benefit from the higher weights $\omega_L = 6$ of *Leaf Event Nodes*. We were able to achieve slightly better results combining both loss functions, since it puts more emphasis on the prediction of *Leaf Event Nodes* while still considering ontology information.

The best results with respect to *top-1 accuracy*, *JSC*, and *CS* were achieved by combining the classification and ontology-driven cosine loss term with *Distance Weights*. The cosine loss is in general more stable when training with and without redundancy, which could indicate that it is more robust to changes in depth and size of the *Ontology*. Furthermore it works well with the *Distance Weight* which does not require an extra weight for *Leaf Event Nodes*.

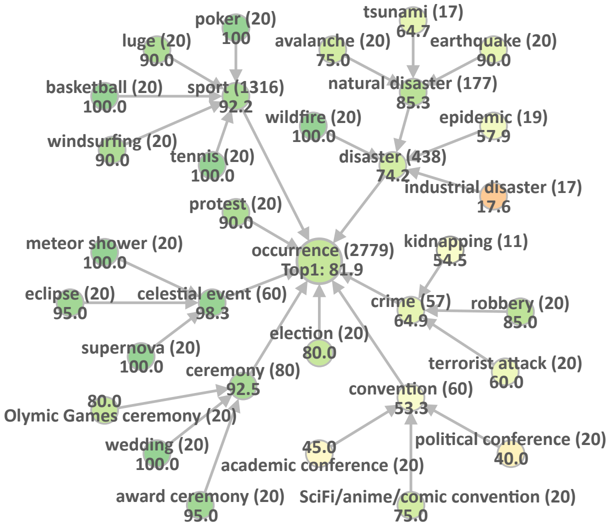

Figure 3. *Top-1 accuracy* and number of images (in brackets) for a selection of *Event Nodes* on *VisE-Bing* using the CO^{cos}_{γ} approach. The results correspond to the mean *top-1 accuracy* of all (also those that are not shown) related *Leaf Event Nodes*. The *Ontology* is simplified for better comprehensibility.

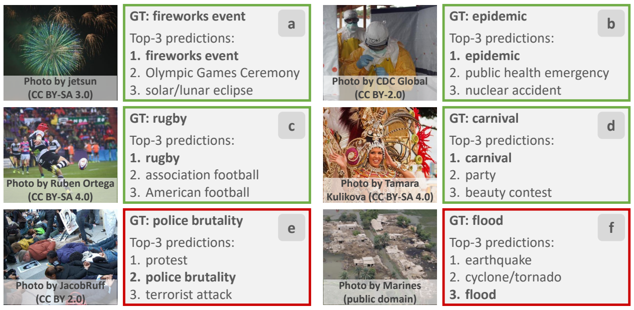

Figure 4. Correctly (green) and incorrectly (red) classified examples of the CO^{cos}_{γ} network model from *VisE-Wiki*.

5.3.2 Performance for Individual Event Types

The *top-1 accuracy* for a selection of *Event Nodes* and qualitative results of the CO^{cos}_{γ} model are provided in Figure 3 and 4. The proposed approach achieves good results for a majority of event types. Misclassification can be typically explained by the visual similarity of the respective events. For example, images for *tornado*, *tsunami*, and *earthquake* are often captured after the actual event and the consequences of these natural disasters can be visually similar as illustrated in Figure 1 and 4f. It also turned out, that classes such as *protest*, *earthquake*, and *explosion* are predicted very frequently, because they depict visual concepts that are also part of other events. For instance, images of the event types *police brutality*, *vehicle fire*, and *economic crisis* are frequently classified as *protest* since they depict typical scenes of riots or demonstrations (Figure 4e). The best results were achieved for sports-centric event types, which is not surprising as they are usually unambiguous. In general, the performance for scheduled event types such as *election* and *sport* is better compared to unexpected or rare events. We assume the main reason is that journalists usually broadcast live coverage of scheduled events, while photos of crimes (e.g., *robbery*, *terrorist attack*) and *natural disasters* are rare and mostly captured by amateurs.

5.3.3 Comparisons on other Benchmarks

We considered several benchmarks including the novel *VisE-Wiki* (Section 3.3) test dataset as well as *WIDER* [39], *SocEID* [6], and *RED* [6]. These benchmarks have different characteristics, which allows us to evaluate the ontology-driven approach in various setups. *WIDER* comprises $50{,}574$ Web images for 61 event types. The *Social Event Image Dataset (SocEID)* consists of circa 37,000 images but contains only eight social event classes, while *Rare Events Dataset (RED)* is comparatively small and contains around $7{,}000$ images from 21 real-world events. We used the splits provided by the authors for *WIDER* [39] and *SocEID* [6]. For *RED* we randomly used 70% of the dataset for training and 30% for testing, as suggested by Ahsan *et al*. [6]. The splits are provided[1] to allow fair comparisons.

As *WIDER*, *SocEID*, and *RED* do not provide an *Ontology*, we have manually linked the classes (e.g., *soccer* to *association football (Q2736)* in *WIDER*) to *Wikidata* to define the set of *Leaf Event Nodes*. Then, we created the *Ontologies* (provided on our *GitHub* repository[1]) according to Section 3.2.2. The models are mostly trained with the parameters from Section 5.1. Due to the smaller dataset sizes the number of training iterations was reduced to $2{,}500$ for *RED* and $10{,}000$ for *SocEID* and *WIDER*. Cosine learning rate annealing [25] was applied from the beginning to lower the learning rate from 0.01 to zero after the specified amount of iterations. The results for our approach and other comparable solutions that use a single network and the whole image as input are presented in Table 5.3.3.

The ontology-driven approaches (CO) clearly outperform the classification baseline (C) on *VisE-Wiki*, *WIDER*, and *RED*. As expected, the results on *SocEID* just slightly improved, because less *Ontology* information are provided due to the lower number of eight classes and thus *Event Nodes*. Compared to the results for *VisE-Bing* (Table 2), the performances are worse on *VisE-Wiki*, because the test set is not manually annotated and contains noisy or ambiguous imagery, particularly for rare event types such as *city fire*. The same applies for *WIDER*. Superior performances are achieved in comparison to similar solutions. It is worth noting that the proposed ontology-driven approach can also be easily integrated in frameworks that utilize ensemble models [4, 5, 36] or additional image regions [17, 39].

Approach	VisE-Wiki 148 classes		WIDER [39] 61 classes		SocEID [6] 8 classes		RED [6] 21 classes	
	Top1	***JSC***	***Top1***	***JSC***	***Top1***	***JSC***	***Top1***	***JSC***
AlexNet [39]	—	—	38.5	—	—	—	—	—
AlexNet-fc7 [6]	—	—	77.9	—	86.4	—	77.9	—
WEBLY-fc7 [6]	—	—	77.9	—	83.7	—	79.4	—
Event conc. [6]	—	—	78.6	—	85.4	—	77.6	—
AlexNet [5]	—	—	41.9	—	—	—	—	—
ResNet152 [5]	—	—	48.0	—	—	—	—	—
C	61.7	72.7	45.6	56.9	91.2	92.7	76.1	82.1
$CO_{6\omega}^{cel}$+RR	63.4	73.9	**51.0**	**61.6**	91.4	**92.9**	79.1	84.3
CO_{γ}^{cos}	**63.5**	**74.1**	49.7	60.3	**91.5**	**92.9**	**80.9**	**85.4**

Table 3. Results on different benchmarks. While our results are superior on *SocEID* and *RED*, Ahsan *et al*. [6] achieved better results (77.9%) on *WIDER* using random splits (gray, not provided on request) also compared to other baselines by training a SVM on *AlexNet* embeddings, which is a similar approach for which Ahmad *et al*. [5] reported 41.9%. Their results for *WIDER* and *RED* are nearly identical, although *WIDER* contains more classes and is in general more challenging. We believe these results are not explainable and need to be verified in a reproducibility experiment.

6. Conclusions and Future Work

In this paper, we have presented a novel ontology, dataset, and ontology-driven deep learning approach for the classification of newsworthy event types in images. A large number of events in conjunction with a knowledge base were leveraged to retrieve an ontology that covers many possible real-world event types. The corresponding large-scale dataset with $570{,}540$ images allowed us to train powerful deep learning models and is, to the best of our knowledge, the most complete and diverse public dataset for event classification to date. We have proposed several baselines including an ontology-driven deep learning approach that exploits event relations to integrate structured information from a knowledge graph. The results on several benchmarks have shown that the integration of structured information from an ontology can improve event classification.

In the future, we plan to further explore strategies to leverage ontology information such as graph convolutional networks. Other interesting research directions are the combination of several knowledge bases and the investigation of semi-supervised approaches to learn from noisy Web data.

Acknowledgement

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no 812997 (project name: "CLEOPATRA"), and the German Research Foundation (DFG: Deutsche Forschungsgemeinschaft) under "VIVA - Visual Information Search in Video Archives" (project number: 388420599).

References

[1] Kashif Ahmad and Nicola Conci. How deep features have improved event recognition in multimedia: A survey. *TOMM*, 15(2):39:1–39:27, 2019.

[2] Kashif Ahmad, Nicola Conci, Giulia Boato, and Francesco G. B. De Natale. USED: a large-scale social event detection dataset. In Christian Timmerer, editor, *Proceedings of the 7th International Conference on Multimedia Systems, MMSys 2016, Klagenfurt, Austria, May 10-13, 2016*, pages 50:1–50:6. ACM, 2016.

[3] Kashif Ahmad, Nicola Conci, and Francesco G. B. De Natale. A saliency-based approach to event recognition. *Signal Process. Image Commun.*, 60:42–51, 2018.

[4] Kashif Ahmad, Mohamed Lamine Mekhalfi, Nicola Conci, Giulia Boato, Farid Melgani, and Francesco G. B. De Natale. A pool of deep models for event recognition. In *2017 IEEE International Conference on Image Processing, ICIP 2017, Beijing, China, September 17-20, 2017*, pages 2886–2890. IEEE, 2017.

[5] Kashif Ahmad, Mohamed Lamine Mekhalfi, Nicola Conci, Farid Melgani, and Francesco G. B. De Natale. Ensemble of deep models for event recognition. *TOMM*, 14(2):51:1–51:20, 2018.

[6] Unaiza Ahsan, Chen Sun, James Hays, and Irfan A. Essa. Complex event recognition from images with few training examples. In *2017 IEEE Winter Conference on Applications of Computer Vision, WACV 2017, Santa Rosa, CA, USA, March 24-31, 2017*, pages 669–678. IEEE Computer Society, 2017.

[7] Internet Archive. Internet Archive snapshot for ”Earthquake” from 18th February 2020, 2020. `https://web.archive.org/web/20200218100604/http:/dbpedia.org/page/Earthquake`, last accessed on 2020-04-20.

[8] Internet Archive. Internet Archive snapshot for ”Tsunami” from 14th February 2020, 2020. `https://web.archive.org/web/20200214202750/http:/dbpedia.org/page/Tsunami`, last accessed on 2020-04-20.

[9] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. Dbpedia: A nucleus for a web of open data. In Karl Aberer, Key-Sun Choi, Natasha Fridman Noy, Dean Allemang, Kyung-Il Lee, Lyndon J. B. Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber, and Philippe Cudré-Mauroux, editors, *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007*, volume 4825 of *Lecture Notes in Computer Science*, pages 722–735. Springer, 2007.

[10] Siham Bacha, Mohand Saïd Allili, and Nadjia Benblidia. Event recognition in photo albums using probabilistic graphical models and feature relevance. *J. Vis. Commun. Image Represent.*, 40:546–558, 2016.

[11] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Event recognition in photo collections with a stopwatch HMM. In *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*, pages 1193–1200. IEEE Computer Society, 2013.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society, 2009.

[13] Sergio Escalera, Junior Fabian, Pablo Pardo, Xavier Baró, Jordi Gonzàlez, Hugo Jair Escalante, Dusan Misevic, Ulrich Steiner, and Isabelle Guyon. Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results. In *2015 IEEE International Conference on Computer Vision Workshop, ICCV Workshops 2015, Santiago, Chile, December 7-13, 2015*, pages 243–251. IEEE Computer Society, 2015.

[14] Michael Goebel, Arjuna Flenner, Lakshmanan Nataraj, and BS Manjunath. Deep learning methods for event verification and image repurposing detection. *Electronic Imaging*, 2019(5):530–1, 2019.

[15] Simon Gottschalk and Elena Demidova. EventKG: A multilingual event-centric temporal knowledge graph. In Aldo Gangemi, Roberto Navigli, Maria-Esther Vidal, Pascal Hitzler, Raphaël Troncy, Laura Hollink, Anna Tordai, and Mehwish Alam, editors, *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings*, volume 10843 of *Lecture Notes in Computer Science*, pages 272–287. Springer, 2018.

[16] Simon Gottschalk and Elena Demidova. EventKG - the hub of event knowledge on the web - and biographical timeline generation. *Semantic Web*, 10(6):1039–1070, 2019.

[17] Xin Guo, Luisa Polania, Bin Zhu, Charles Boncelet, and Kenneth Barner. Graph neural networks for image understanding based on multiple cues: Group emotion recognition and event recognition as use cases. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 2921–2930, 2020.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.

[19] Andrew Howard, Ruoming Pang, Hartwig Adam, Quoc V. Le, Mark Sandler, Bo Chen, Weijun Wang, Liang-Chieh Chen, Mingxing Tan, Grace Chu, Vijay Vasudevan, and Yukun Zhu. Searching for mobilenetv3. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 1314–1324. IEEE, 2019.

[20] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2261–2269. IEEE Computer Society, 2017.

[21] Vidit Jain, Amit Singhal, and Jiebo Luo. Selective hidden random fields: Exploiting domain-specific saliency for event

classification. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA*. IEEE Computer Society, 2008.

[22] Ayush Jaiswal, Ekraam Sabir, Wael Abd-Almageed, and Premkumar Natarajan. Multimedia semantic integrity assessment using joint embedding of images and text. In Qiong Liu, Rainer Lienhart, Haohong Wang, Sheng-Wei "Kuan-Ta" Chen, Susanne Boll, Yi-Ping Phoebe Chen, Gerald Friedland, Jia Li, and Shuicheng Yan, editors, *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*, pages 1465–1471. ACM, 2017.

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pages 1106–1114, 2012.

[24] Li-Jia Li and Fei-Fei Li. What, where and who? classifying events by scene and object recognition. In *IEEE 11th International Conference on Computer Vision, ICCV 2007, Rio de Janeiro, Brazil, October 14-20, 2007*, pages 1–8. IEEE Computer Society, 2007.

[25] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[26] Eric Müller-Budack, Kader Pustu-Iren, and Ralph Ewerth. Geolocation estimation of photos using a hierarchical model and scene classification. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XII*, volume 11216 of *Lecture Notes in Computer Science*, pages 575–592. Springer, 2018.

[27] Eric Müller-Budack, Jonas Theiner, Sebastian Diering, Maximilian Idahl, and Ralph Ewerth. Multimodal analytics for real-world news using measures of cross-modal entity consistency. In Cathal Gurrin, Björn Thór Jónsson, Noriko Kando, Klaus Schöffmann, Yi-Ping Phoebe Chen, and Noel E. O'Connor, editors, *Proceedings of the 2020 on International Conference on Multimedia Retrieval, ICMR 2020, Dublin, Ireland, June 8-11, 2020*, pages 16–25. ACM, 2020.

[28] Timo Reuter, Symeon Papadopoulos, Georgios Petkos, Vasileios Mezaris, Yiannis Kompatsiaris, Philipp Cimiano, Christopher M. De Vries, and Shlomo Geva. Social event detection at mediaeval 2013: Challenges, datasets, and evaluation. In Martha A. Larson, Xavier Anguera, Timo Reuter, Gareth J. F. Jones, Bogdan Ionescu, Markus Schedl, Tomas Piatrik, Claudia Hauff, and Mohammad Soleymani, editors, *Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop, Barcelona, Spain, October 18-19, 2013*, volume 1043 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2013.

[29] Ekraam Sabir, Wael AbdAlmageed, Yue Wu, and Prem Natarajan. Deep multimodal image-repurposing detection. In Susanne Boll, Kyoung Mu Lee, Jiebo Luo, Wenwu Zhu, Hyeran Byun, Chang Wen Chen, Rainer Lienhart, and Tao Mei, editors, *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018*, pages 1337–1345. ACM, 2018.

[30] Paul Hongsuck Seo, Tobias Weyand, Jack Sim, and Bohyung Han. Cplanet: Enhancing image geolocalization by combinatorial partitioning of maps. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part X*, volume 11214 of *Lecture Notes in Computer Science*, pages 544–560. Springer, 2018.

[31] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In Carey L. Williamson, Mary Ellen Zurko, Peter F. Patel-Schneider, and Prashant J. Shenoy, editors, *Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May 8-12, 2007*, pages 697–706. ACM, 2007.

[32] Ilya Sutskever, James Martens, George E. Dahl, and Geoffrey E. Hinton. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 1139–1147. JMLR.org, 2013.

[33] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6450–6459. IEEE Computer Society, 2018.

[34] Nam N. Vo, Nathan Jacobs, and James Hays. Revisiting IM2GPS in the deep learning era. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2640–2649. IEEE Computer Society, 2017.

[35] Denny Vrandecic and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, 2014.

[36] Limin Wang, Zhe Wang, Yu Qiao, and Luc Van Gool. Transferring deep object and scene representations for event recognition in still images. *Int. J. Comput. Vis.*, 126(2-4):390–409, 2018.

[37] Tobias Weyand, Ilya Kostrikov, and James Philbin. Planet - photo geolocation with convolutional neural networks. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII*, volume 9912 of *Lecture Notes in Computer Science*, pages 37–55. Springer, 2016.

[38] Zifeng Wu, Yongzhen Huang, and Liang Wang. Learning representative deep features for image set analysis. *IEEE Trans. Multimedia*, 17(11):1960–1968, 2015.

[39] Yuanjun Xiong, Kai Zhu, Dahua Lin, and Xiaoou Tang. Recognize complex events from static images by fusing deep channels. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1600–1609. IEEE Computer Society, 2015.

[40] Zhongwen Xu, Yi Yang, and Alexander G. Hauptmann. A discriminative CNN video representation for event detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1798–1807. IEEE Computer Society, 2015.

[41] Bolei Zhou, Àgata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(6):1452–1464, 2018.

[42] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. Learning transferable architectures for scalable image recognition. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 8697–8710. IEEE Computer Society, 2018.