Few-Shot Event Classification in Images using Knowledge Graphs for Prompting

Golsa Tahmasebzadeh, Matthias Springstein, Ralph Ewerth, Eric Müller-Budack
{golsa.tahmasebzadeh, matthias.springstein, ralph ewerth, eric.mueller}@tib.eu
TIB – Leibniz Information Centre for Science and Technology, Hannover, Germany
L3S Research Center, Leibniz University Hannover, Hannover, Germany

https://github.com/TIBHannover/PromptImageEvent

Abstract

Event classification in images plays a vital role in multimedia analysis especially with the prevalence of fake news on social media and the Web. The majority of approaches for event classification rely on large sets of labeled training data. However, image labels for fine-grained event instances (e.g., 2016 Summer Olympics) can be sparse, incorrect, ambiguous, etc. A few approaches have addressed the lack of labeled data for event classification but cover only few events. Moreover, vision-language models that allow for zero-shot and few-shot classification with prompting have not yet been extensively exploited. In this paper, we propose four different techniques to create hard prompts including knowledge graph information from Wikidata and Wikipedia as well as an ensemble approach for zero-shot event classification. We also integrate prompt learning for state-of-the-art vision-language models to address few-shot event classification. Experimental results on six benchmarks including a new dataset comprising event instances from various domains, such as politics and natural disasters, show that our proposed approaches require much fewer training images than supervised baselines and the state-of-the-art while achieving better results.

1. Introduction

With the advent of social media, the daily amount of multimodal news available on the Web is enormous. This leads to a rising demand for tools to collect, analyze, organize, and retrieve multimodal news information. Contextualization of images has been widely studied from multiple perspectives useful for news corpora, such as geolocation estimation [19, 29], place classification [34], and argument extraction in events [17]. Nonetheless, few approaches have addressed the identification of real-world events in images [4, 20, 30]; even though it contains crucial information to comprehend multimedia content, e.g., for fact-checking or misinformation detection.

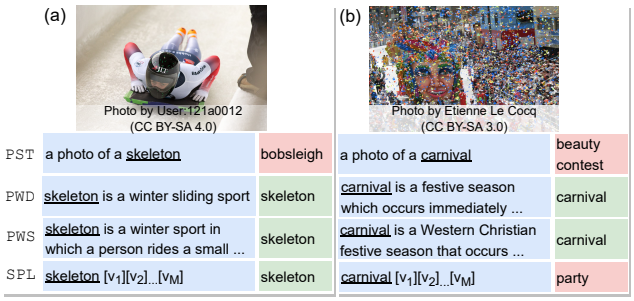

Figure 1. Event classification results using *hard prompts* based on static class labels (PST), *Wikidata* descriptions (PWD), and *Wikipedia* summaries (PWS) as well as soft prompt learning (SPL) where M context tokens $[v_i]$ are learned from labeled training data [36]. Left: prompt for the ground-truth class of the photo. Right: prediction based on the most likely prompt among all classes. In both examples, PST struggles to identify the correct event while PWD and PWS succeed. In (b), SPL does not capture *carnival*, emphasizing the importance of knowledge graph prompts for event contextualization.

Recent approaches for event classification in images mainly fine-tune deep learning models based on labeled datasets (e.g., [4, 20, 30]). But most of these datasets (e.g., [20, 30]) cover only general event types (e.g., sports, types of natural disasters) which limits their applicability in the real-world that typically requires the classification of concrete events (e.g., *2020 U.S. presidential election*). There are only a few datasets that comprise real-world events (e.g., [4]) but they consider a rather limited selection of real-world events with only a few annotated images. Due to the diversity of real-world events and the resulting lack of large-scale training data, zero-shot and few-shot approaches for event classification are of utmost importance. To handle lack of training data, ensemble models based on scene and object descriptors have been used [2, 3, 28]. Ahsan et al. [4] addressed few-shot event classification by training concept classifiers to categorize images into social event types with minimal training samples. More recently, Said et al. [22]

proposed an active learning approach to choose effective training examples for disaster analysis. In recent years, *vision-language models (VLMs)* have shown promising performance in zero-shot and few-shot settings for multiple downstream tasks [14, 21] through prompting [15, 35–37] but have not yet been exploited for event classification of images. In this context, *hard prompts* are based on handcrafted templates, primarily introduced by *CLIP* (Contrastive Language-Image Pretraining [21]), e.g., "This is a photo of a [Class]", and are used to describe a given textual label for zero-shot prediction. However, these *hard prompts* typically do not cover actual class descriptions and can be ambiguous (Fig. 1a) or too unspecific (Fig. 1b). In such cases, prompt learning approaches [25, 35, 36] can automatically learn more meaningful descriptions, denoted as *soft prompts*. Furthermore, external knowledge sources such as *Wikidata* and *Wikipedia* contain summaries and descriptions to contextualize events [20] (Fig. 1b). However, descriptions from knowledge graphs have not yet been leveraged to create and improve prompts for event classification.

In this paper, we suggest and investigate various prompting techniques and supervised approaches for zero-shot and few-shot event classification in images. The main contributions are as follows: (1) We present four hard prompting techniques as well as soft prompt learning to leverage capabilities of *VLMs*; (2) To obtain more distinct descriptions for specific events, we enrich prompts with external knowledge sources such as *Wikidata* and *Wikipedia*, and examine the effectiveness of prompt ensembles; (3) We introduce a novel dataset called *Event Instances* to demonstrate feasibility of our approach for the classification of fine-grained real-world events. It contains 184 event instances for different event types, e.g., *election*, *protest*, and *natural disaster*; (4) To address the lack of labeled training data, we provide several baselines using prompting and supervised learning for zero-shot and few-shot event classification. Experiments on six datasets including our novel *Event Instances* dataset demonstrate the superiority of the proposed approaches to the state-of-the-art in few-shot scenarios. Dataset, source code, and models are publicly available.

The remainder of this paper is organized as follows. Section 2 reviews related work on event classification, and prompting for image classification. Our proposed approach for zero-shot and few-shot event classification is presented in Section 3. In Section 4, we discuss the experimental setup and results. Section 5 concludes the paper and outlines future research directions.

2. Related Work

In this section, we review approaches for event classification in images. Since we aim to combine *VLMs* with prompt learning for event classification, we also provide a brief overview of prompting in image classification.

Event Type Classification According to Yang et al. [32], event classification techniques vary depending on the definition of an event and mainly fall into two categories. The first category includes *activities* (e.g., *people celebrating*) either in videos [18, 31], or personal photo albums [6, 7, 10]. The second category covers *real-world news events* from various domains such as *social movements*, and *politics* [4, 11, 20, 30]. In this paper, we focus on the latter definition. The majority of approaches typically use convolutional neural networks (CNNs, e.g., [12, 24]) to extract rich features such as local information from image patches, object regions using object detection techniques, or require place (scene) information [1, 8, 13, 28]. More recently, Müller-Budack et al. [20] proposed an ontology-driven deep learning approach based on 148 unique real-world events extracted from *EventKG* [9]. All of these approaches rely on large labeled image datasets such as *Web Image Dataset for Event Recognition (WIDER)* [30] and the *Visual Event Classification Dataset (VisE-D)* [20]. However, these datasets do not cover real-world event instances (e.g., *2020 U.S. election*) that allow for a broader range of applications. In addition, labeled training data for such event instances is typically sparse. To address this issue, Ahsan et al. [4] propose an event concept learning framework for few-shot event classification along with the *Rare Event Dataset (RED)* that covers 21 event instances. Said et al. [22] leverage an active learning strategy to choose effective training examples for few-shot classification of disaster images. However, novel vision-language models (VLMs) that are successful for many downstream tasks with few training images have not yet been explored for event classification in images.

Prompting in Image Classification Due to the success of *VLMs* (e.g., [5, 14, 21]), several approaches have been proposed recently that use vision and language encoders for zero-shot and few-shot image classification tasks through prompting. While *hard prompts* are based on hand-crafted templates that describe the class labels (e.g. [21]), various approaches are proposed for automatically learning prompts (e.g., [35–37]). Even though prompting techniques demonstrate promising performance in image classification benchmarks, their potential for news data, particularly for event type classification is under-explored. More specifically, prompting can provide a context that could help disambiguate concepts for classification in the photo (e.g., *"skeleton is a winter sliding sport"* in Fig. 1a).

3. Few-shot Event Classification based on Vision-Language Models

This section introduces our approaches for few-shot event classification (Fig. 2). We aim to tackle zero-shot and few-shot event classification using recent vision-language

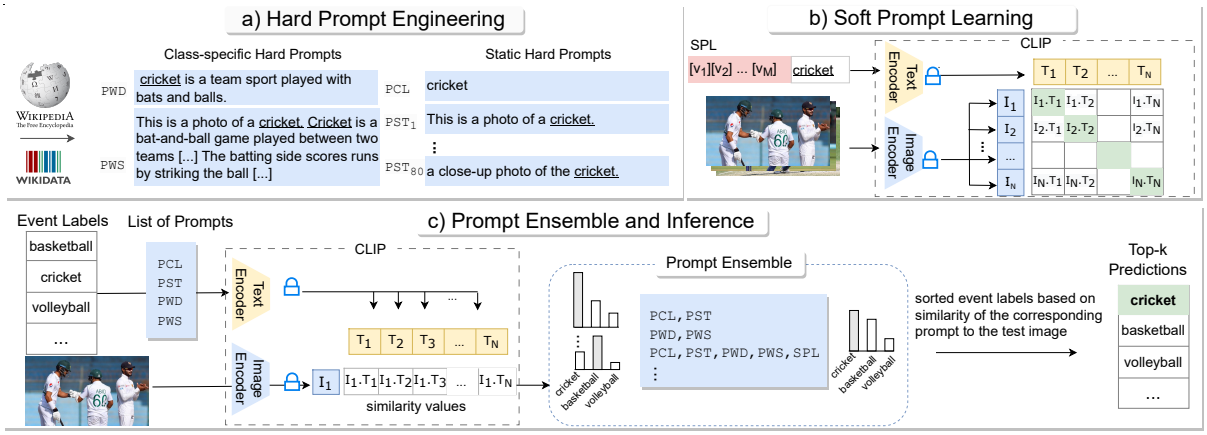

Figure 2. Workflow of the proposed few-shot event classification approach. (a) We propose four strategies to create *hard prompts* based on the class labels (PCL) and a set of static templates (PST) as well as based on knowledge graph information, i.e.,*Wikidata* descriptions (PWD) and *Wikipedia* summaries (PWS). (b) We use *CoOp* [36] for soft prompt learning (SPL) based on few training images. (c) Once the prompts are produced, we propose a prompt ensemble to compute the similarity to a test image during inference. Finally, the sorted list of event labels based on the similarity values are the top-k predictions. Please note that we only use *hard prompts* for zero-shot event classification.

models (VLMs) such as *CLIP*. First, we propose to fine-tune *CLIP* based on multimodal news articles (Section 3.1). In Section 3.2, we propose different prompting techniques for events, i.e., hard prompt engineering, soft prompt learning, and prompt ensemble. Finally, we describe the use of prompts for event classification (Section 3.3).

3.1. CLIP Fine-tuning

The *CLIP* model has been pre-trained on a large dataset composed of 400 million image-text pairs from the Web using a contrastive loss function to embed input pairs to a joint embedding [21]. To obtain a more comprehensive understanding of event-centric documents in different domains including *politics*, *sports*, etc., we fine-tune *CLIP* on a large dataset of image-text pairs extracted from news articles. We use the *Multimodal Geolocation Estimation of News (MMG-NewsPhoto)* dataset [26] which contains image-text pairs in various news domains. For each news image, the dataset provides the corresponding body text and a caption. It consists of 554,768 and 60,893 samples for train and validation. We pair every image with its caption and all sentences in the body text. As proposed by Schuhmann et al. [23], we filter out pairs with cosine similarity values of less than 0.3 based on the image and text encoders of *CLIP*. As a result, we get 436,092 and 48,281 samples for training and validation to fine-tune *CLIP* for the news domain. Training details are provided in Section 4.1.1.

3.2. Prompting Techniques

The *CLIP* model has proven to be effective in zero-shot and few-shot settings for many downstream tasks [21]. As explained in Section 3.3, classification is conducted by measuring the similarity between the textual prompts that describe a set of pre-defined concepts, in our case events $\mathbb{E}$, to a test image. The performance of the zero-shot classification relies on quality of the textual prompts. We present four strategies to automatically create hard prompts for zero-shot classification (Section 3.2.1) as well as a prompt-learning technique to create *soft prompts* for few-shot classification (Section 3.2.2). Furthermore, we suggest to combine these prompts according to Section 3.2.3.

3.2.1 Hard Prompts for Zero-Shot Classification

Given a vision-language model such as *CLIP*, the automatic creation of hard prompts (descriptions) for a set of pre-defined events $\mathbb{E}$ allows for zero-shot classification. We suggest four strategies to create hard prompts, including novel approaches that use *Wikidata* and *Wikipedia*.

Prompts based on Class Labels (PCL) The most basic hard prompt PCL solely consists of the event name, i.e., class label. Since the event name often comprises a single or few words without much context (see Fig. 2), more descriptive prompts are required to exploit the *CLIP* model.

Prompts based on Static Templates (PST) Typically, zero-shot classification approaches based on *CLIP* make use of *static prompts* that are employed to create a sentence for the set of concepts to be classified. Therefore, the second type of hard prompt is based on templates shared between all classes denoted as PST. These templates are based on 80 different context prompts introduced by Radford et al. [21], e.g., "This is a photo of a [Class]", "a close-up photo of the [Class]", etc. Although these templates add some content to the concepts, they still can be ambiguous. For instance,

the sports event *skeleton* (Fig. 1) can be confused with the human skeleton without providing further context.

Prompts based on *Wikidata* Descriptions (PWD) To add more context to the prompts, we incorporate *class-specific prompts* that provide detailed descriptions for each event label. For this purpose, we employ external knowledge graphs that contain structured data for many real-world events. The *Event Knowledge Graph* (*EventKG* [9]) has extracted events from several large-scale knowledge graphs including, but not limited to, *Wikidata* [27] and *Wikipedia*. Müller-Budack et al. [20] used *EventKG* to link event classes of several benchmark datasets [4,20,30] to *Wikidata*. Using the links, we extract *Wikidata* description to create a prompt, denoted as PWD, as follows: "[Class] is a [Wikidata description of the class]". We limit the prompt to *CLIP*'s maximum context length of 77 tokens, which is rarely exceeded since *Wikidata* descriptions are typically quite short. However, they may lack important details or can be ambiguous, e.g., the description "a team sport played with bats and balls" of the sport *cricket* also applies to *baseball*. Also, for some events the descriptions are not available (see supplementary material) in which case we use PCL as prompt.

Prompts based on *Wikipedia* Summaries (PWS) To address the limitations of PWD, we aim to create more comprehensive event definitions using *Wikipedia*. Therefore, we extract English summaries for all events using the *Wikipedia* URLs provided by *Wikidata*. If the English summary is not available, we automatically select a summary from a sorted list of languages such as 'German', 'Spanish', etc., and translate the text to English using *Google Translate* API. Based on the collected summaries, we define PWS as "This is a photo of a [class]. [Wikipedia summary]". If no summary is available (only in rare cases, see supplementary material), we only use the first part of the prompt. Compared to the previous prompts, these summaries can comprise long texts. To not exceed the maximum input context length for *CLIP*, we limit PWS to the first 77 tokens that typically contain the most important information of the summary.

3.2.2 Prompt Learning for Few-shot Classification

Soft Prompt Learning (SPL) We apply Context Optimization (*CoOp* [36]) to automatically learn *soft prompts* (denoted as SPL) using few training samples. This method is aimed at learning m learnable context vectors per class. In other words, *CoOp* optimizes the context vectors $p = \{\boldsymbol{f}_i^e, \boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_m\}$ where $\boldsymbol{f}_i^e$ denotes token embedding for i-th event label $e \in \mathbb{E}$, and $\boldsymbol{v}_i$ denotes a learnable token with same number of d dimensions as embeddings of the text encoder $\psi_T(.)$. The position of event token embedding $\boldsymbol{f}_i^e$ can be at the front (as in the notation above), middle, or end of the context vectors. To assess the impact of knowledge graph information in prompt learning, we initialize the learnable prompts with embeddings of the PWD and PWS. Training details are provided in Section 4.1.2.

3.2.3 Prompt Ensemble

The prompts proposed in the previous section can have synergies as they have different strengths and weaknesses. For example, while PWD provides a concise description, PWS offers a more in-depth, contextual understanding for the event labels. Sometimes, *Wikidata* or *Wikipedia* might not provide sufficient information for certain events. In such cases, PCL and PST can act as reliable alternatives to fill the information gap. When training data becomes available, *soft prompts* (SPL) can be learned to create suitable descriptions to differentiate events. To leverage synergies between prompts, we combine them to generate a set of rich queries. For this purpose, we compare each prompt to an image, which results in a set of similarity scores based on the inference strategy in Section 3.3. We suggest to aggregate similarities of different prompts in the ensemble approach using mean of all scores since it provides better results than max operation as detailed in the supplemental material.

3.3. Inference

During inference, we predict the event label for a given test image I based on a set of pre-defined event $\mathbb{E}$. For this purpose, we employ the prompts introduced in Section 3.2 to describe each event. For each prompt T, we extract the textual embedding $\mathbf{e}_T = \psi_T(T)$ from *CLIP*'s text encoder ψ_T. The test image I is fed to the image encoder ψ_I to obtain a respective image embedding $\mathbf{e}_I = \psi_I(I)$. The textual embeddings of all prompts are compared to the image embedding of the test image using the dot product to extract image-text similarities. The *softmax* function is applied over these similarities to represent the predicted class probabilities $\hat{\mathbf{y}}$. Finally, the probabilities are sorted to predict the top-k events for the given query image.

4. Experimental Setup and Results

Here, we present implementation details (Section 4.1), experimental setup (Section 4.2) and results (Section 4.3).

4.1. Implementation Details

Details on fine-tuning *CLIP* for news (see Section 3.1) are provided in Section 4.1.1. Furthermore, we present the training details for soft prompt learning in Section 4.1.2.

4.1.1 Fine-tuning CLIP for News

For fine-tuning the *CLIP* model [21] for news articles (Section 3.1), we initialize the text and vision transformers using

Table 1. Dataset statistics including the number of event instances, test and training images for various event types (from left to right: Election, Referendum, Epidemic, Protest, Political Campaign, Natural Disaster) for the *Event Instances* dataset.

	Elec.	Ref.	Epid.	Prot.	Pol. Camp.	Nat. Dis.	Total
Event Inst.	26	5	5	79	63	6	184
Test Im.	462	60	232	1459	974	143	3330
Train Im.	396	61	176	1311	859	123	2926
Total Im.	858	121	408	2770	1833	266	6256

the "ViT-B/32" model pre-trained on OpenAI's *WebImage-Text (WIT)* dataset. We set the batch size to 128 with a learning rate of $1e^{-6}$ with a linearly decaying schedule. We train the model for 100 epochs using *Adam* optimizer [16]. The model with the lowest mean rank for cross-modal retrieval on the validation set is used in the experiments.

4.1.2 Prompt Learning

For automatic prompt learning, we rely on few-shot learning scenario where the model is expected to learn effective representations and predictions from a limited amount of data. For the training process, we rely on the setup provided by *CoOp* [36]. We set learning rate to $1e^{-6}$, batch size to 32, and number of epochs to 200. The best model is chosen based on the accuracy on validation splits provided by the benchmark datasets (Section 4.2.1). In all experiments, the best context length is 16, position of class label is front, the initialization is random. A comparison of different context lengths (4 vs. 16), position of the class label (front, middle, end), and initialization method (`PWD`, `PWS`, and random) is included as supplemental material. For a fair comparison, we use vision transformer "ViT-B/32" for all experiments.

4.2. Experimental Setup

Section 4.2.1 introduces the benchmark datasets including the novel *Event Instances* dataset and a data sampling strategy for few-shot classification. The evaluation metrics and baselines are presented in Sections 4.2.2 and 4.2.3.

4.2.1 Benchmark Datasets

Public Benchmarks We use four public benchmark datasets for the evaluation. (1) The *Visual Event Classification Dataset (VisE-D)* [20] covers 148 diverse event types with 570,540 images. The test sets *VisE-Bing* and *VisE-Wiki* include 2,779 and 8,138 samples, respectively. (2) The *Web Image Dataset for Event Recognition (WIDER)* [30] comprises 25,275 images for training plus 25,299 samples for test on 61 events. (3) The *Rare Events Dataset (RED)* [4] contains 21 real-world events with 7,000 images where 30% is used for test. For a fair comparison, we use the splits provided by Müller-Budack et al. [20]. (4) The *Social Event Dataset (SocEID)* [4] contains eight social events with 27,687 images for train and 9,237 for test.

Event Instances Dataset Since *RED* [4] covers only a small set of 21 real-world event instances, we introduce a new dataset called *Event Instances*. To collect this dataset, we select different event types (see Table 1) from *Wikidata* that have high societal and environmental impact, e.g., *election (Q40231)* and *natural disaster (Q8065)*. Then, we query *Wikidata* for all event instances based on the *"instance Of" (P31)* relation and choose the most popular instances that are accessed on average more than 100 times per day on *Wikipedia* since 2015. We downloaded the corresponding images from the associated *Wikimedia Commons* category *(P373)* and manually verified that images depict the respective event instance. The annotation details are provided as supplementary material. The dataset contains 6,256 images for 184 event instances that are randomly assigned to the training and test set. It covers much more real-world events from a broader spectrum of six event types than *RED*. The statisics are presented in Table 1.

Data Sampling For Few-shot Classification Few-shot learning is an inherently challenging problem due to the risk of limited generalization. Thus, we randomly sample sets for training and validation for the datasets mentioned in Section 4.2.1. The sampling is repeated three times to create three different models and alleviate potential artifacts from overfitting. For benchmark datasets, we use official train and validation splits while for *Event Instances* dataset, we skip validation sampling because of limited images per class. For each set, we randomly select up to n images (see Fig. 4) from the train and validation sets to assess the impact of number of training images on classification performance. For the evaluation, we average scores for all three models on the test set using the metrics mentioned below.

4.2.2 Evaluation Metric

For the evaluation, we use the top-1 accuracy to compare with the state-of-the-art approaches. The top-3 and top-5 accuracy values are reported as supplementary material.

4.2.3 Baseline Systems

Supervised Learning To create a baseline for few-shot learning that does not rely on prompting, we use the `Linear probe` approach on top of visual features from images using *CLIP*'s image encoder. We follow the same training method as Radford et al. [21]. We use validation set to find the best regularization value and model based on

the top-1 accuracy. As a second approach, we train a support vector machine (SVM) for which the results are reported in the supplementary material as it performs slightly worse.

State-of-the-Art We compare the proposed approaches with two best-performing state-of-the-art baselines. (1) The *Event concepts* approach [4] learns concept classifiers, for a combination of objects, scenes, actions, and attributes for social events to address few-shot event classification. (2) The ontology-driven (CO_{γ}^{cos}) approach [20] is based on a *ResNet-50* [12] that uses an ontology based on various events covered in *Wikidata* for optimization.

4.3. Experimental Results

Here, we report and discuss results for zero-shot (Section 4.3.1) and few-shot classification (Section 4.3.2) on five benchmarks and our novel *Event Instances* datasets. We compare our approach to baselines (Section 4.3.3) and analyze the results for *Event Instances* dataset (Section 4.3.4).

4.3.1 Zero-shot Classification

Comparison of Hard Prompt Techniques As shown in Table 2, static templates (PST) outperform prompts based on class labels (PCL) in the zero-shot setting confirming previous studies [21]. Although both of these prompts perform quite well in datasets containing event types, prompts based on Wikidata descriptions (PWD) provide the best results for datasets with fine-grained events, i.e., *Event Instances* and *RED*. However, both individual prompts based on knowledge graphs (PWD and PWS) do not provide significant improvements compared to the static prompts but we argue that they still provide contextual information that can improve event classification as the next paragraph reveals.

Impact of Prompt Ensemble We note that combination of knowledge graph prompts (PWD,PWS) outperforms individual prompts. More importantly, the ensemble of all *hard prompts* (PCL,PST,PWD,PWS) achieves promising results. This shows how the prompt ensemble technique harnesses synergies between knowledge graph and *static prompts* to achieve a more comprehensive and enriched set of queries for event classification.

4.3.2 Few-shot Classification

For few-shot classification, we experiment with 1-50 training samples per class. The results are displayed in Table 2.

Impact of Prompt Learning If the number of training samples is low (ca. 5) the prompt learning technique SPL is not able to outperform the zero-shot approaches for event type classification. In contrast, considering the *Event Instances* dataset individual *hard prompts* such as PCL and PST obtain very low accuracy compared to the *soft prompts* such as SPL. This confirms the fact that the *CLIP* text encoders are less effective in distinguishing more fine-grained events compared to broader event types. The fine-tuned backbone CLIP-MMG sometimes outperforms CLIP-WIT when the number of training samples is low (ca. 5).

Impact of Prompt Ensemble As Table 2 shows, the combination of *soft prompts* with *static prompts* SPL,PST results in a considerable improvement. Also it is observed that use of knowledge graph prompts SPL,PST,PWD,PWS considerably boosts the performance and outperforms all prompting techniques including the zero-shot approaches. One interesting insight is that, as shown in Fig. 4, ensemble of all prompts SPL,PST,PWD,PWS requires only around two images per class to outperform zero-shot methods, whereas the SPL prompt requires at least ten images. Thus, with fewer samples per class, it is beneficial to use an ensemble of *soft prompts* with knowledge graph prompts.

4.3.3 Comparison to the Baselines

In this section, we evaluate the *CLIP* fine-tuning and compare our approach to state-of-the-art supervised baselines.

Impact of *CLIP* Fine-tuning The results of CLIP-MMG are comparable but mostly slightly inferior to CLIP-WIT. Performance especially lacks for *Event Instances*. One reason might be the sampling of image-text pairs for fine-tuning where it is not ensured that body text correlates with the image or it mentions an event. More advanced sampling approaches as well as other fine-tuning strategies, e.g., only training the text encoder [33] could resolve this issue.

Comparison to the State-of-the-Art As Table 2 shows, the ensemble of *hard prompts* (PCL,PST,PWD,PWS) achieves better or comparative results to the state-of-the-art without the need for training data. This highlights the potential of zero-shot classification over supervised methods that require large training data. Considering few-shot approaches, the ensemble of all prompts is superior on most datasets. When using more training samples (up to 30), we observe that SPL outperforms the zero-shot approaches. Thus, if we have sufficient training data, prompt learning techniques achieve enhanced results. However, if we have a low number of training samples, the combination of *soft prompts* with knowledge graph prompts yield improved results compared to other baselines and zero-shot approaches.

Comparison to the `Linear probe` Regarding the zero-shot classification, as shown in Fig. 4, the Linear

Table 2. Comparison of different approaches based on top-1 accuracy using different number of samples per class (n). Two types of backbones are used: (1) The `CLIP-WIT` pre-trained on the *WIT* dataset [21]; (2) The `CLIP-MMG` fine-tuned on the *MMG-News* dataset [26].

Approach	n	Backbone	VisE-Bing	VisE-Wiki	RED	WIDER	SocEID	Event Instances
Zero-shot Event Classification								
`PCL`	0	`CLIP-WIT`	75.06	60.09	77.61	50.03	**90.70**	32.76
`PST`	0	`CLIP-WIT`	78.95	64.50	79.06	51.30	89.78	34.50
`PWD`	0	`CLIP-WIT`	75.60	62.68	79.39	49.52	83.98	35.44
`PWS`	0	`CLIP-WIT`	75.64	62.02	77.00	46.16	88.72	33.03
`PWD,PWS`	0	`CLIP-WIT`	79.20	65.79	78.87	50.86	88.63	**36.79**
`PST,PWD,PWS`	0	`CLIP-WIT`	80.60	**67.07**	79.48	52.39	89.72	36.76
`PCL,PST,PWD,PWS`	0	`CLIP-WIT`	**80.89**	67.02	**80.00**	**52.97**	90.51	36.37
`PCL,PST,PWD,PWS`	0	`CLIP-MMG`	79.92	66.85	79.53	52.61	87.04	33.48
Few-shot Event Classification								
`SPL`	5	`CLIP-WIT`	66.01	54.96	66.04	43.97	85.87	52.08
`SPL,PST`	5	`CLIP-WIT`	77.05	64.12	76.73	53.12	90.71	**57.05**
`SPL,PST,PWD,PWS`	5	`CLIP-WIT`	**81.96**	68.59	79.89	**55.78**	**91.95**	54.57
`SPL,PST,PWD,PWS`	5	`CLIP-MMG`	81.27	**68.67**	**80.53**	54.65	89.13	49.53
`Linear probe`	5	`CLIP-WIT`	68.73	57.04	69.70	46.91	89.18	54.49
`Linear probe`	5	`CLIP-MMG`	68.27	55.96	70.42	46.52	88.04	54.01
`SPL`	30	`CLIP-WIT`	79.80	63.40	81.10	55.63	91.86	71.14
`SPL,PST`	30	`CLIP-WIT`	84.37	68.72	**83.41**	**59.91**	93.31	**72.28**
`SPL,PST,PWD,PWS`	30	`CLIP-WIT`	**84.70**	**70.97**	83.03	58.84	**93.33**	66.00
`SPL,PST,PWD,PWS`	30	`CLIP-MMG`	83.76	69.96	82.68	56.83	91.22	58.71
`Linear probe`	30	`CLIP-WIT`	82.78	66.75	81.78	59.48	92.83	72.26
`Linear probe`	30	`CLIP-MMG`	81.91	66.27	81.63	58.88	92.14	71.57
Fully-supervised Baselines using *all* training images								
CO_γ^{cos} [20]	all		81.90	63.50	**80.90**	49.70	**91.50**	–
Event concepts [4]	all		–	–	77.60	**78.60**	85.40	–

Table 3. Comparison of different prompts per event type on the *Event Instances* dataset based on n number of samples per class.

Approach	n	Election	Referendum	Epidemic	Protest	Political Campaign	Natural Disaster
`PWD,PWS`	0	27.98	26.55	50.62	33.88	46.98	70.66
`SPL`	30	54.38	68.90	75.84	**58.72**	51.82	82.46
`SPL,PST`	30	**60.61**	**72.90**	**80.75**	58.37	57.55	82.46
`SPL,PST,PWD,PWS`	30	55.09	62.28	75.07	55.99	**57.87**	**83.38**

`probe` (`CLIP-WIT`) requires at least around ten images per class to outperform the zero-shot approaches. However, when the number of training samples is low (less than 5), the zero-shot approaches considerably outperform the `Linear probe` (`CLIP-WIT`). Regarding few-shot approaches the `SPL,PST,PWD,PWS` prompting considerably outperforms the supervised methods for a smaller number of training samples (e.g., 5). However, a training dataset containing more than 30 images per class allows for supervised methods, i.e., `Linear probe` (`CLIP-WIT`), to leverage their capacity for learning and slowly start to outperform prompting approaches in about 50 images As Table 2 shows, `Linear probe` (`CLIP-MMG`) maintains competitive performance compared to the `CLIP-WIT` version. Therefore, fine-tuning *CLIP* does not necessarily lead to superior performance. In summary, prompt learning aids image event classification, particularly with fewer training samples per class. However, if there are sufficient training samples supervised learning yields improved results.

4.3.4 In-depth Analysis of *Event Instances* Dataset

Table 3 presents results on *Event Instances* dataset per event type. The results are significantly better in the few-shot setting compared to the zero-shot setting indicating that the task of event instance classification is very challenging and requires prompt learning. Furthermore, it is shown that a prompt ensemble is more effective compared to soft prompts individually. Overall, the best results are achieved for *natural disasters* (e.g., Fig. 3d) and significant improvements can be made for *political campaigns*.

Figure 3. Qualitative examples based on soft prompts (SPL) and a prompt ensemble (SPL,PST,PWD,PWS) that includes knowledge graph information on the *Event Instances* dataset. The correct prediction is colored green.

Event instances typically share similar visual attributes and knowledge graph information can provide geographical and temporal context to differentiate similar instances. Examples are *climate change denial* (Fig. 3c) and *2018 European drought and heat wave* (Fig. 3d). Nonetheless, sometimes *hard prompts* lack context or are too short (e.g., Fig. 3a, 3b), in these cases SPL is more effective. For example, *Wikidata* descriptions for 16 out of 26 *election* instances are missing (see supplementary material) which might explain the lower performance of in the ensemble of prompts.

5. Conclusions and Future Work

In this paper, we have proposed novel approaches for zero-shot and few-shot event classification of images based on novel prompting techniques for vision-language models. For zero-shot classification, we have suggested four different *hard prompts* that include knowledge graph information from *Wikidata* and *Wikipedia*. We have combined these prompts with *soft prompts* learned for specific events in an ensemble approach for few-shot event classification. Furthermore, we have introduced a novel dataset that encompasses fine-grained events from various types such as *protests* and *natural disasters* to assess our approach for real-world scenarios. Experimental results have demonstrated that *hard prompts* based on event descriptions from knowledge graphs yield significantly improved results compared to simple *hard prompts* (e.g., "This is a photo of a [Class]") in zero-shot settings. Also, an ensemble of *hard prompts* with *soft prompts* greatly reduces the need for a large amount of training data. The proposed approaches outperform state-of-the-art on six benchmark test datasets using much fewer images for training.

In future work, more advanced prompt learning techniques [15, 25] can be investigated. Another interesting direction is to explore ways to add information from longer descriptions (e.g., Wikipedia summaries) and to integrate ontology information from knowledge graphs into prompts.

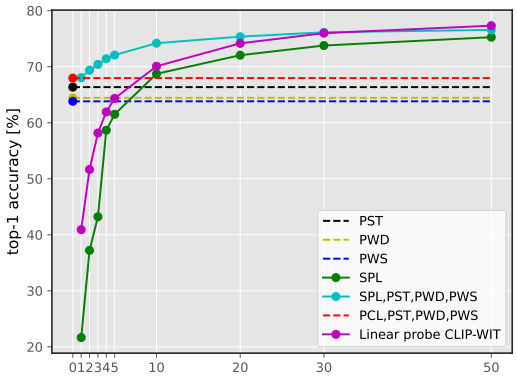

Figure 4. Comparison of different prompting techniques based on number of training samples per class. The values are averaged over all the datasets introduced in Section 4.2.1. As illustrated, the ensemble of all prompts (SPL,PST,PWD,PWS) requires considerably less number of samples (about two) to outperform other approaches compared to using only the *soft prompt* SPL.

Acknowledgements

This work was funded by the Ministry of Lower Saxony for Science and Culture (Responsible AI in digital society, project no. 51171145), and the German Federal Ministry of Education and Research (BMBF, FakeNarratives, project no. 16KIS1517).

References

[1] Kashif Ahmad, Nicola Conci, and Francesco G. B. De Natale. A saliency-based approach to event recognition. *Signal Processing: Image Communication*, pages 42–51, 2018. 2

[2] Kashif Ahmad, Mohamed Lamine Mekhalfi, Nicola Conci, Giulia Boato, Farid Melgani, and Francesco GB De Natale. A pool of deep models for event recognition. In *IEEE International Conference on Image Processing, ICIP 2017*, pages 2886–2890. IEEE, 2017. 1

[3] Kashif Ahmad, Mohamed Lamine Mekhalfi, Nicola Conci, Farid Melgani, and Francesco G. B. De Natale. Ensemble of deep models for event recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications, TOMM*, (2):51:1–51:20, 2018. 1

[4] Unaiza Ahsan, Chen Sun, James Hays, and Irfan A. Essa. Complex event recognition from images with few training examples. In *Winter Conference on Applications of Computer Vision, WACV 2017, Santa Rosa, CA, USA, March 24-31, 2017*, pages 669–678. IEEE Computer Society, 2017. 1, 2, 4, 5, 6, 7

[5] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems, NeurIPS 2022*, 2022. 2

[6] 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, 2013. 2

[7] Tamar Glaser, Emanuel Ben Baruch, Gilad Sharir, Nadav Zamir, Asaf Noy, and Lihi Zelnik-Manor. PETA: photo albums event recognition using transformers attention. In *International Conference on Pattern Recognition, ICPR 2022, Montreal, QC, Canada, August 21-25, 2022*, pages 2532–2538. IEEE, 2022. 2

[8] Michael Goebel, Arjuna Flenner, Lakshmanan Nataraj, and Bangalore S. Manjunath. Deep learning methods for event verification and image repurposing detection. In *Media Watermarking, Security, and Forensics 2019, Burlingame, CA, USA, 13-17 January 2019*. Ingenta, 2019. 2

[9] Simon Gottschalk and Elena Demidova. Eventkg: A multilingual event-centric temporal knowledge graph. In *Extended Semantic Web Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings*, pages 272–287. Springer, 2018. 2, 4

[10] Cong Guo and Xinmei Tian. Event recognition in personal photo collections using hierarchical model and multiple features. In *International Workshop on Multimedia Signal Processing, MMSP 2015, Xiamen, China, October 19-21, 2015*, pages 1–6. IEEE, 2015. 2

[11] Xin Guo, Luisa F. Polanía, Bin Zhu, Charles Boncelet, and Kenneth E. Barner. Graph neural networks for image understanding based on multiple cues: Group emotion recognition and event recognition as use cases. In *Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, USA, March 1-5, 2020*, pages 2910–2919. IEEE, 2020. 2

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 2, 6

[13] Vidit Jain, Amit Singhal, and Jiebo Luo. Selective hidden random fields: Exploiting domain-specific saliency for event classification. In *Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA*. IEEE Computer Society, 2008. 2

[14] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 4904–4916. PMLR, 2021. 2

[15] Muhammad Uzair Khattak, Hanoona Abdul Rasheed, Muhammad Maaz, Salman H. Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 19113–19122. IEEE, 2023. 2, 8

[16] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015*, 2015. 5

[17] Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. Clip-event: Connecting text and images with event structures. In *Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 16399–16408. IEEE, 2022. 1

[18] Zhihui Li, Xiaojun Chang, Lina Yao, Shirui Pan, Zongyuan Ge, and Huaxiang Zhang. Grounding visual concepts for zero-shot event detection and event captioning. In *Knowledge Discovery and Data Mining SIGKDD, Virtual Event, CA, USA, August 23-27, 2020*, pages 297–305. ACM, 2020. 2

[19] Eric Müller-Budack, Kader Pustu-Iren, and Ralph Ewerth. Geolocation estimation of photos using a hierarchical model and scene classification. In *European Conference on Computer Vision, ECCV 2018, Munich, Germany, September 8-14, 2018*, pages 575–592, 2018. 1

[20] Eric Müller-Budack, Matthias Springstein, Sherzod Hakimov, Kevin Mrutzek, and Ralph Ewerth. Ontology-driven event type classification in images. In *Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 2927–2937. IEEE, 2021. 1, 2, 4, 5, 6, 7

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 8748–8763, 2021. 2, 3, 4, 5, 6, 7

[22] Naina Said, Kashif Ahmad, Nicola Conci, and Ala I. Al-Fuqaha. Active learning for event detection in support of disaster analysis applications. *Signal, Image and Video Processing*, (6):1081–1088, 2021. 1, 2

[23] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: open dataset of clip-filtered 400 million image-text pairs. *CoRR*, 2021. 3

[24] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015*, 2015. 2

[25] Ximeng Sun, Ping Hu, and Kate Saenko. Dualcoop: Fast adaptation to multi-label recognition with limited annotations. In *Advances in Neural Information Processing Systems, NeurIPS 2022*, 2022. 2, 8

[26] Golsa Tahmasebzadeh, Sherzod Hakimov, Ralph Ewerth, and Eric Müller-Budack. Multimodal geolocation estimation of news photos. In *European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part II*, pages 204–220. Springer, 2023. 3, 7

[27] Denny Vrandecic and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, pages 78–85, 2014. 4

[28] Limin Wang, Zhe Wang, Yu Qiao, and Luc Van Gool. Transferring deep object and scene representations for event recognition in still images. *International Journal of Computer Vision*, (2-4):390–409, 2018. 1, 2

[29] Tobias Weyand, Ilya Kostrikov, and James Philbin. Planet - photo geolocation with convolutional neural networks. In *European Conference on Computer Vision, ECCV, Amsterdam, The Netherlands, October 11-14, 2016*, pages 37–55. Springer, 2016. 1

[30] Yuanjun Xiong, Kai Zhu, Dahua Lin, and Xiaoou Tang. Recognize complex events from static images by fusing deep channels. In *Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1600–1609. IEEE Computer Society, 2015. 1, 2, 4, 5

[31] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 3333–3343, 2022. 2

[32] Zhenguo Yang, Zhuopan Yang, Zhiwei Guo, Zehang Lin, Haizhong Zhu, Qing Li, and Wenyin Liu. Towards temporal event detection: A dataset, benchmarks and challenges. *IEEE Transactions on Multimedia*, 2023. 2

[33] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18102–18112. IEEE, 2022. 6

[34] Bolei Zhou, Àgata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):1452–1464, 2018. 1

[35] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 16795–16804. IEEE, 2022. 2

[36] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, (9):2337–2348, 2022. 1, 2, 3, 4, 5

[37] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. *CoRR*, 2022. 2