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

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

The supplemental material includes details of the context vectors for the prompt learning approach (Section A) and reports the number of cases where the prompts PWD and PWS cannot be extracted (Section B). Furthermore, the annotation process for the novel *Event Instances* dataset (Section C) is described. Section D provides a comparison of different aggregation methods for the prompt ensemble technique and Section E compares the supervised approaches. Finally, we report the top-3 and top-5 accuracy results in Section F.

A. Context Vector Details for Prompt Learning

We describe our prompt learning technique [5] for few-shot event classification in Section 3.2.2 of the paper. There are different ways to select proper context vectors for tuning. We experiment different context lengths (4 vs 16), the position of the class label (front, middle, end), and the initialization method (PWD, PWS, random). As shown in Table 1, a longer context length obtains better performance for all different initialization prompts. Regarding the position of the class label, we achieved almost identical results and use front for the experiments in the paper.

B. Missing Knowledge Graph Information for Prompt Creation

As mentioned in Section 3.2.1 of the paper, in some cases either the *Wikidata* description for PWD or *Wikipedia* summary for PWS cannot be extracted. While only 2/61 *Wikipedia* summaries are missing for the classes *laborer* and *car driving* in *WIDER* [1], *Wikidata* descriptions are missing for 1/21 class (*2015 Russian air raids in Syria*) in the Rare Event Instances (RED) dataset [1] and for 43/184 classes in *Event Instances*:

- 16/26 descriptions for *Election* instances are missing
- 2/5 descriptions for *Referendum* instances are missing
- 12/79 descriptions for *Protest* instances are missing
- 13/63 descriptions for *Political Campaign* instances are missing

As mentioned in the paper, the PCL prompts are used when *Wikidata* descriptions are missing and the prompt "This is a photo of a [class]." is used for classes with missing *Wikipedia* summaries.

C. Annotation Process of the *Event Instances* Dataset

We introduced the novel *Event Instances* dataset of images for evaluation on fine-grained events in Section 4.2.1. In this context, to verify whether the downloaded images represent the corresponding instance, we performed a manual verification process. To this end, one of the co-authors had to validate whether an image corresponds to the provided event label. Since it is challenging to verify the relevance of an image to the corresponding fine-grained event (e.g., *2018 European drought and heat wave*) the co-author was instructed to employ two different methods. In the first method, for every individual event instance, the meta information (e.g., date and location) of images of the corresponding *Wikidata* Web page were considered to verify if they match the provided event instance label. In the second approach, the co-author had to search for similar images using external resources such as news Web pages and determine their visual similarity to the images in the dataset. The final data statistics from the manual annotation process can be found in Table 1 of the paper.

D. Comparison of the Aggregation Methods for the Prompt Ensemble Approach

As mentioned in Section 3.2.3 of the paper, one of the techniques that we use for prompting event labels is an ensemble approach, where we combine the similarity scores

Table 1. Comparison of different context lengths, initialization methods, and positions of the class label for prompt learning. Results are reported for the best prompt ensemble SPL,PST,PWD,PWS on the *Event Instances* test set.

Len.	Class Pos.	Random	PWD	PWS
4	front	64.55	64.48	64.9
16	front	**66.00**	65.81	66.01
16	middle	65.99	65.09	**66.12**
16	end	65.87	**66.03**	65.87

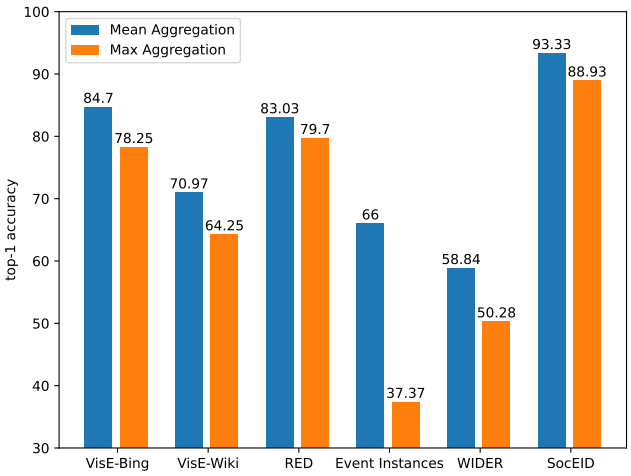

Figure 1. Comparison between max and mean aggregation methods for the prompt ensemble technique (SPL,PST,PWD,PWS) introduced in Section 3.2.3 of the paper.

obtained from the individual prompts. To aggregate the similarity scores, we use max and mean operations. A comparison of these two approaches is presented in Fig. 1. As shown for all the six test datasets, the mean operation is superior to the max operation considerably. The *Event Instances* dataset particularly demonstrates this, where the *hard prompts* like PWD and PWS have difficulty differentiating event instances, even if they yield high similarity scores (refer to Table 2 of the paper). Thus, it is more advantageous to give equal importance to all prompts when making the final prediction, rather than solely depending on the prompt with the highest similarity score. For future direction, we could explore improved weighting methods or investigate automatic learning of such schemes.

E. Supervised Learning based on SVM

We introduced the supervised baseline Linear probe in Section 4.2.3 of the paper. We also experimented with an SVM approach. In general, we follow the same training procedure as described in the paper. We perform a grid search to find the best regularization and gamma values and select the best model based on the top-1 accuracy on the validation set. We repeat the training using the same three training and validation sets. The evaluation scores on the test sets are averaged for the three resulting models.

As shown in Table 2, the Linear probe approach outperforms SVM regardless of the number of training samples per class. Same as Linear probe, the SVM approach also achieves better results than state-of-the-art approaches trained on large dataset by using only 30 number of images.

F. Results based on Top-3 and Top-5 Accuracy

The top-3 and top-5 accuracy are presented in Tables 3 and 4 for different test sets and approaches. The results confirmed the findings reported in Section 4.3 of the paper and show that our proposed approaches perform significantly better than the state-of-the-art while using much fewer images for training.

References

[1] 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, 3, 4

[2] 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. 3, 4

[3] 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. 3, 4

[4] 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, 4

[5] 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

Table 2. Comparison of supervised methods `SVM` and `Linear probe` 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 [3]; (2) The `CLIP-MMG` fine-tuned on the *MMG-News* dataset [4].

Approach	n	Backbone	VisE-Bing	VisE-Wiki	RED	WIDER	SocEID	Event Instances
`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
`SVM`	5	`CLIP-WIT`	67.68	55.28	67.72	45.47	88.74	53.66
`SVM`	5	`CLIP-MMG`	66.88	55.29	68.78	45.38	87.05	53.60
`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
`SVM`	30	`CLIP-WIT`	82.60	65.55	81.50	59.14	92.21	71.58
`SVM`	30	`CLIP-MMG`	82.08	65.52	81.38	58.50	92.49	71.35
CO_γ^{cos} [2]	all		81.90	63.50	**80.90**	49.70	**91.50**	–
Event concepts [1]	all		–	–	77.60	**78.60**	85.40	–

Table 3. Comparison of different approaches based on top-3 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 [3]; (2) The `CLIP-MMG` fine-tuned on the *MMG-News* dataset [4].

Approach	n	Backbone	VisE-Bing	VisE-Wiki	RED	WIDER	SocEID	Event Instances
`PCL`	0	`CLIP-WIT`	90.72	76.60	89.91	68.85	97.92	52.13
`PST`	0	`CLIP-WIT`	92.84	79.60	89.77	70.79	98.33	53.66
`PWD`	0	`CLIP-WIT`	91.29	78.89	91.22	67.54	96.02	57.36
`PWS`	0	`CLIP-WIT`	90.18	78.84	89.58	65.00	97.59	53.84
`PWD,PWS`	0	`CLIP-WIT`	92.98	81.15	91.08	69.12	98.01	**57.84**
`PST,PWD,PWS`	0	`CLIP-WIT`	93.59	81.90	91.13	70.95	98.24	57.48
`PCL,PST,PWD,PWS`	0	`CLIP-WIT`	**93.59**	82.77	91.13	**71.68**	**98.38**	56.52
`PCL,PST,PWD,PWS`	0	`CLIP-MMG`	92.77	81.76	**91.74**	71.00	97.59	54.02
`SPL`	5	`CLIP-WIT`	82.32	70.66	82.85	61.37	95.95	67.53
`SPL,PST`	5	`CLIP-WIT`	91.11	79.12	90.27	71.49	98.09	**74.82**
`SPL,PST,PWD,PWS`	5	`CLIP-WIT`	**94.00**	**82.92**	91.86	**74.30**	**98.58**	73.93
`SPL,PST,PWD,PWS`	5	`CLIP-MMG`	93.43	82.83	**92.60**	73.22	98.10	69.34
`Linear probe`	5	`CLIP-WIT`	85.03	73.94	87.18	65.60	97.42	71.23
`Linear probe`	5	`CLIP-MMG`	84.44	72.83	87.23	65.29	97.02	71.41
`SPL`	30	`CLIP-WIT`	92.16	77.18	92.77	73.84	98.07	83.60
`SPL,PST`	30	`CLIP-WIT`	94.58	81.82	**94.49**	**77.95**	98.72	**86.09**
`SPL,PST,PWD,PWS`	30	`CLIP-WIT`	**95.30**	**83.84**	94.27	77.58	**98.96**	82.98
`SPL,PST,PWD,PWS`	30	`CLIP-MMG`	94.34	83.40	**94.41**	75.59	98.58	76.96
`Linear probe`	30	`CLIP-WIT`	94.34	81.2	94.04	78.11	98.80	84.54
`Linear probe`	30	`CLIP-MMG`	93.76	80.74	93.84	77.71	98.49	84.48
CO_γ^{cos} [2]			90.80	74.30	–	–	–	–
Event concepts [1]			–	–	–	–	–	–

Table 4. Comparison of different approaches based on top-5 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 [3]; (2) The `CLIP-MMG` fine-tuned on the *MMG-News* dataset [4].

Approach	n	Backbone	VisE-Bing	VisE-Wiki	RED	WIDER	SocEID	Event Instances
PCL	0	CLIP-WIT	94.53	82.26	94.27	76.00	99.23	61.08
PST	0	CLIP-WIT	95.72	84.48	93.66	77.46	99.63	61.65
PWD	0	CLIP-WIT	95.11	84.79	**95.54**	74.64	98.72	67.00
PWS	0	CLIP-WIT	93.99	83.66	94.13	72.44	99.42	63.09
PWD,PWS	0	CLIP-WIT	96.15	85.92	95.45	75.97	99.46	**67.90**
PST,PWD,PWS	0	CLIP-WIT	96.33	**86.41**	95.21	77.84	99.61	67.03
PCL,PST,PWD,PWS	0	CLIP-WIT	**96.37**	86.19	95.21	**78.55**	**99.62**	66.43
PCL,PST,PWD,PWS	0	CLIP-MMG	96.22	86.31	95.49	77.50	99.31	64.20
SPL	5	CLIP-WIT	87.14	76.04	88.95	68.72	98.68	74.24
SPL,PST	5	CLIP-WIT	94.32	83.74	94.63	78.03	99.45	81.57
SPL,PST,PWD,PWS	5	CLIP-WIT	96.76	86.99	96.06	**80.84**	**99.65**	**81.61**
SPL,PST,PWD,PWS	5	CLIP-MMG	**96.91**	**87.13**	**96.18**	79.89	99.50	77.67
Linear probe	5	CLIP-WIT	89.74	79.44	92.46	73.17	99.23	78.27
Linear probe	5	CLIP-MMG	89.31	78.70	92.07	72.75	98.97	78.76
SPL	30	CLIP-WIT	95.32	81.47	96.09	80.25	99.40	87.72
SPL,PST	30	CLIP-WIT	97.10	85.77	**97.33**	**83.84**	99.63	**90.38**
SPL,PST,PWD,PWS	30	CLIP-WIT	**97.97**	**87.85**	97.39	83.70	**99.72**	88.20
SPL,PST,PWD,PWS	30	CLIP-MMG	97.39	87.77	96.99	81.89	99.62	83.36
Linear probe	30	CLIP-WIT	97.07	85.36	97.21	84.15	99.68	89.05
Linear probe	30	CLIP-MMG	96.80	85.34	97.06	83.74	99.64	88.86
CO_{γ}^{cos} [2]			93.20	78.80	–	–	–	–
Event concepts [1]			–	–	–	–	–	–