

# TNT: Text-Conditioned Network with Transductive Inference for Few-Shot Video Classification - Supplementary Material

Andrés Villa<sup>1</sup>

afvilla@uc.cl

Juan-Manuel Perez-Rua<sup>\*2</sup>

jmpr@fb.com

Victor Escorcía<sup>†2</sup>

v.castillo@samsung.com

Vladimir Araujo<sup>1,4</sup>

vgaraujo@uc.cl

Juan Carlos Niebles<sup>3</sup>

jniebles@cs.stanford.edu

Alvaro Soto<sup>†1</sup>

asoto@ing.puc.cl

<sup>1</sup> Pontificia Universidad Católica de Chile  
Santiago, Chile

<sup>2</sup> Samsung AI Center Cambridge  
Cambridge, UK

<sup>3</sup> Stanford University  
Stanford, CA, USA

<sup>4</sup> KU Leuven  
Leuven, Belgium

In the main paper, we introduce a novel Few-Shot Learning (FSL) model for video action classification: **Text-Conditioned Networks with Transductive inference (TNT)**. This model leverages the semantic information in the textual descriptions of support instances as a privileged source of information to improve class discrimination in a scarce data regime. Specifically, we leverage the textual descriptions to modulate the visual encoder and compute transductive inference, augmenting the support instances with those unlabeled through an attention-based and multimodal approach. Overall, the integration of these abilities allows our model to adapt to the FSL tasks quickly. This supplementary material contains the following: **(1)** An overview of the main attributes of the four challenging video datasets used to evaluate our model; **(2)** Qualitative evaluations to demonstrate the relevance of using textual descriptions to modulate our network. Moreover, this document is accompanied by a video that presents the principal motivation and contributions of our approach.

## 1 Datasets

We evaluate our model on four challenging video action FSL benchmarks, those that contain rich textual descriptions (such as EK-92 and SS-100 [1]), and short class-level descriptions (such as MetaUCF-101 [2] and Kinetics-100 [3]). Interestingly, in EK-92 and SS-100 [1], the descriptions are specific per each video instance and can have more than 8 and 15 words, as is shown in Fig. 1-a and Fig. 1-b, respectively. Thus, its descriptions are detailed and very correlated with the actions and objects in the videos. Conversely, in Kinetics-100 [3]

Dataset	Num Classes			Num Instances			Rich Textual Descriptions
	Train	Val	Test	Train	Val	Test	
EK-92	58	11	23	49621	8352	18370	✓
SS-100	64	12	24	67013	1926	2857	✓
MetaUCF-101	70	10	21	9154	1421	2745	✗
Kinetics-100	64	12	24	6400	1200	2400	✗

Table 1: **Statistics** from the datasets for FSL: SS-100 [1], MetaUCF-101 [1], Kinetics-100 [1] and the introduced EK-92.

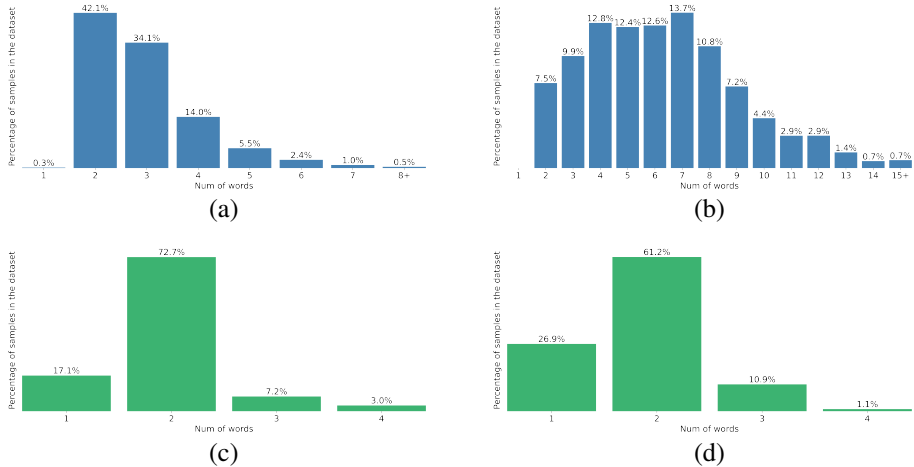


Figure 1: **Length of action descriptions.** Distribution of the number of words per instance description in both families of datasets. (Blue figures) Those with rich textual descriptions: (a) EK-92 and (b) SS-100 [1]. (Green figures) Those with short class-level descriptions: (c) Kinetics-100 [1] and (d) MetaUCF-101 [1].

and MetaUCF-101 [1], the descriptions correspond to the class label. Therefore, they are the same for the videos that belong to the same class. Fig. 1-c and Fig. 1-d show that Kinetics-100 [1] and MetaUCF-101 [1], respectively, have short descriptions with no more than 2 words for most of them. Notably, our model achieves outstanding results on the four benchmarks, even on those with short class-level descriptions. Table 1 summarizes the main attributes of these benchmarks, which are presented in the Dataset section of the main paper.

## 2 Qualitative results

**Visualization of Class Activation Maps.** An important aspect of our proposed method is its ability to adapt the feature backbone to a particular task. We are interested in verifying the effect of our proposed text-conditioned module in video samples. To study the influence of FiLM layers to modulate the TSN network using semantic information from textual action descriptions, we analyze the class activation mapping (CAM) for the cases with and without employing the FiLM layers. We use our TNT model trained on SS-100 for the 5-shot 5-way task. Fig. 2-a shows the visualization obtained for one test sample after applying our TNT model with and without FiLM layers following the Grad-CAM [1] approach.

As it can be seen, the TNT model manages to exploit relevant cues from the support textual descriptions. In this example, our FiLM-ed model activations are located in appropriate

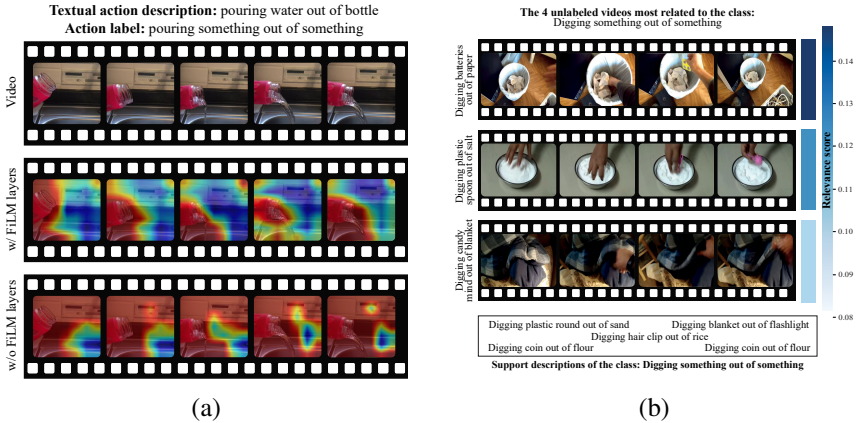


Figure 2: **Visualizations.** (a) We take our model pre-trained on SS-100 in the 5-shot and 5-way task and analyze the changes of its CAMs with and without using the FiLM layers before the last average pooling layer of the ResNet. (b) Top-3 relevant unlabeled samples for the class in a 5-way 5-shot task. The darkest color indicates the most important sample.

regions for the action “Pouring water out of bottle”. Of particular interest, it seems that the effect of our text-conditioned features is to allow the underlying model to focus on the scene elements that matter the most for the particular query sample. This is, in this sample, the bottle and the pouring water. Meanwhile, the model with no FiLM layers tends to attend irrelevant regions. This arguably has effects in a diminished capacity for fine-grained classification. In general, we found that the conditioning process, based on the textual action descriptions, enables the network to obtain better video features for specific tasks.

**Dynamic Module Inspection.** To assess the transductive capabilities of our model, we closely study the proposed dynamic module. We exploit our SS-100-trained TNT model in the 5-way 5-shot setting. For this experiment, we directly observe the values of the relevance weight matrix  $\mathbf{W}$ , which can be easily interpreted. Indeed, they encode how relevant is each one of the query set videos to the semantic class embedding obtained by textual descriptions of the support set  $\mathbf{E}_{class}^T$ . Fig. 2-b shows the three most related videos from the query set  $\mathcal{Q}$  to the “digging something out of something” class. The text-driven class representation is obtained by averaging the semantic embedding of all sampled descriptions for this class in the support set. These descriptions are shown at the bottom of the figure. Furthermore, in this figure, we present a heatmap in which darker color means greater relevance of the query video sample to the actual support set description. Interestingly, the top three most related videos from the query set are the ones that also belong to the class “digging something out of something”. This example demonstrates that our dynamic module can leverage the semantic information of textual action descriptions to augment the support set samples with unlabeled samples from the query set. This behavior of our model is essential to improve class prototypes and alleviate the problem of low training data.

## References

- [1] Kaidi Cao, Jingwei Ji, Zhangjie Cao, Chien-Yi Chang, and Juan Carlos Niebles. Few-shot video classification via temporal alignment. In *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2020.

- 
- [2] Ashish Mishra, Vinay Kumar Verma, M Shiva Krishna Reddy, Arulkumar S, Piyush Rai, and Anurag Mittal. A generative approach to zero-shot and few-shot action recognition. In *2018 IEEE Winter Conference on Applications of Computer Vision*, pages 372–380, Los Alamitos, CA, USA, mar 2018. IEEE Computer Society. doi: 10.1109/WACV.2018.00047. URL <https://doi.ieeecomputersociety.org/10.1109/WACV.2018.00047>.
  - [3] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Int. Conf. Comput. Vis.*, Oct 2017.
  - [4] Linchao Zhu and Yi Yang. Compound memory networks for few-shot video classification. In *Eur. Conf. Comput. Vis.*, September 2018. doi: 10.1007/978-3-030-01234-2\_46.