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

Andrés Villa[1], Juan-Manuel Perez-Rua[2], Victor Escorcia[2], Vladimir Araujo[1,4], Juan Carlos Niebles[3], Alvaro Soto[1]

[1]Pontificia Universidad Católica de Chile, [2]Samsung AI Center Cambridge,

[3]Stanford University, [4]KU Leuven

{afvilla,vgaraujo}@uc.cl, jmpr@fb.com, v.castillo@samsung.com

jniebles@cs.stanford.edu, asoto@ing.puc.cl

Abstract

Recently, few-shot video classification has received an increasing interest. Current approaches mostly focus on effectively exploiting the temporal dimension in videos to improve learning under low data regimes. However, most works have largely ignored that videos are often accompanied by rich textual descriptions that can also be an essential source of information to handle few-shot recognition cases. In this paper, we propose to leverage these human-provided textual descriptions as privileged information when training a few-shot video classification model. Specifically, we formulate a text-based task conditioner to adapt video features to the few-shot learning task. Furthermore, our model follows a transductive setting to improve the task-adaptation ability of the model by using the support textual descriptions and query instances to update a set of class prototypes. Our model achieves state-of-the-art performance on four challenging few-shot video action classification benchmarks. The code can be found at: here.

1. Introduction

Humans use language to guide their learning process [17]. For instance, when teaching how to prepare a cooking recipe, visual samples are often accompanied by detailed or rich language-based instructions (*e.g.*, "*Place aubergine onto pan*"), which are fine-grained and correlated with the visual content. These instructions are a primary cause of the human ability to quickly learn from few examples because they help to transfer learning among tasks, disambiguate and correct error sources [17]. However, modern deep learning approaches in action recognition [9, 14, 30] have mainly focused on a large amount of labeled visual data ignoring the textual descriptions that are usually included along with the videos [7, 10]. These limitations have motivated an increasing interest in Few-Shot Learning (FSL) [31], which consists of learning novel concepts from few labeled instances.

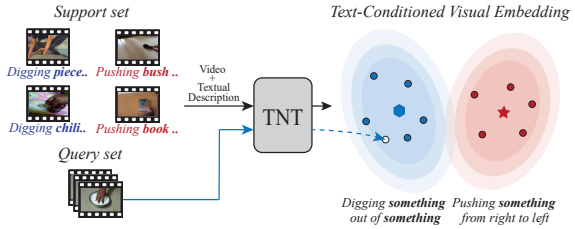

Figure 1. **Outline of our FSL setting.** Our model leverages the rich text descriptions of the support instances (left) to improve class discrimination (right) in two different ways. 1) Modulating the visual feature encoder to alleviate the large intra-class variations of video data. 2) A transductive setting where textual information of the support instances is used alongside visual information of the query set to augment the support set.

The few existing video FSL methods follow one of two approaches: (i) exploiting the temporal and spatial dimensions in videos [2, 34]; or (ii) taking advantage of large amounts of additional video data by using tag retrieval to overcome the labeled data scarcity [32]. However, recent work has not explicitly leveraged the available natural language descriptions that come with videos as an additional information source. These descriptions can be easily obtained without further effort while the dataset is collected, as described by [4]. During the Epic-Kitchens [4] collection, the actors simply narrated their actions using free-form language. We found that these text descriptions are crucial to recognizing actions in a few-shot regime, which agrees with the human ability to compound and exploit multimodal knowledge to learn from few training samples quickly.

In this paper, we present these main contributions: (**I**) To the best of our knowledge, we propose a new class of models: **T**ext-conditioned **N**etworks with **T**ransductive inference or TNT that leverages the semantic information in textual action descriptions of the support data as a privileged

source of information [27] to improve class discrimination in few-shot video classification, see Fig. 1. (**II**) We show the advantage of using the semantic information in support textual action descriptions to perform transductive learning. We develop a dynamic prototype module that uses textual semantic representations to obtain class prototypes using both labeled and unlabeled samples following an attentive approach. (**III**) We demonstrate that textual embeddings outperform the video ones for task adaptation even when these descriptions are short and class-specific (*e.g.*, class labels: *Headbanging*, *Stretching leg*, etc). (**IV**) We achieve state-of-the-art performance with two families of video action FSL benchmarks, those with detailed or rich textual descriptions such as Something-Something-100 (SS-100) [2] and the new benchmark Epic-Kitchens-92 (EK-92), and those with short class-level textual descriptions such as MetaUCF-101 [18] and Kinetics-100 [35].

2. Related Work

Few-Shot Learning. It is possible to identify three main groups of methods. (i) Gradient based methods: they focus on learning a good parameter initialization that facilitates model adaptation by few-shot fine-tuning [5, 19, 21]. (ii) Metric learning based methods: they aim to learn better metrics for determining similarity of input samples in the semantic embedding space [12, 20, 24, 26, 29]. (iii) More recently methods [1, 23] extend the conditional neural process framework [6] with the goal of effective task-adaptation.

Induction vs Transduction in FSL. There are two types of inference approaches: inductive and transductive. In the inductive setting, only the support or labeled instances are used to guide the inference process [1, 12, 20, 23, 24, 26, 29]. In contrast, in the transductive setting, the model uses extra information from query or unlabeled samples to perform its inference [11, 15, 16, 19].

Few-Shot Video Classification. The shift of action recognition research from coarse [10] to fine-grained categories [4, 8] has intensified the problem of data scarcity. A few works to tackle this issue have appeared recently. However, most of the works focused only on better exploiting visual or temporal information from videos [2, 13, 18, 32–36]. We aim to bridge the gap between the few-shot samples and the nuanced and complex concepts needed for video representation learning by using textual descriptions as privileged information to contextualize the video feature encoder and the classification approach.

3. Method

3.1. Problem Definition

FSL aims to obtain a model that can generalize well to novel classes with few support instances. Therefore, we follow the standard FSL setting [24, 29], wherein a trained model f_θ is evaluated on a significant number of $N-$way $K-$shot tasks sampled from a meta-test set D_{test}. These tasks consist of N novel categories, from which K samples are sampled to form support set $\mathcal{S}$, where K is a small integer, typically, 1 or 5. $\mathcal{S}$ is used as a proxy to classify the B unlabeled instances from the query set $\mathcal{Q}$. The parameters θ of the model f are trained on a meta-training set D_{train}, by applying the episodic training strategy proposed by [29]. This is, $N-$way $K-$shot classification tasks are simulated by sampling from D_{train} during meta-training. $\mathcal{Q}$ is sampled from the same N categories in such a way that the samples in $\mathcal{Q}$ are non-overlapping with $\mathcal{S}$. The set of classes available for meta-training are often referred to as base classes. Note that the model f is evaluated on different categories than it is trained on. In this paper, we assume that a text description is available for each instance in $\mathcal{S}$.

3.2. TNT Model

We strive for action classification in videos within a low-data setting by means of (i) the rich semantic information of textual action descriptions and (ii) exploiting the unlabeled samples at test time. We accomplish this task with our **T**ext-Conditioned **N**etworks with **T**ransductive Inference (TNT), depicted in Fig. 2. Our overall model f is a text-conditioned neural network designed to be flexible and adaptive to novel action labels. Taking inspiration from [1, 23], TNT is composed by three modules: (i) Task-Conditioned Video Encoder g; (ii) Task Conditioner Ψ; and (iii) Task-Conditioned Transductive Classifier h.

Task-Conditioned Video Encoder (g). g transforms each video v into a meaningful representation $\mathbf{v}$ focusing the latent information essential for the novel classes. To this end, it uses the TSN [30] with a ResNet backbone that is enhanced by adding Feature-wise Linear Modulation (FiLM) layers after the BatchNorm layer of each ResNet block. FiLM layers adapt the internal representation $\mathbf{v}_i$ at the i^{th} block of g via an affine transformation $FiLM(\mathbf{v}_i; \gamma_i, \beta_i) = \gamma_i \mathbf{v}_i + \beta_i$ where γ_i and β_i are the modulation parameters generated by the Task Conditioner module.

Task Conditioner. The Task Conditioner Ψ is an essential part of our approach that provides high adaptability to our model. Specifically, it leverages the RoBERTa language model [22] to compute conditioning signals that modulate the Task-Conditioned Video Encoder g and the Task-Conditioned Transductive Classifier h based on the textual action descriptions of a set of support instances $\mathcal{S}$. Due to the inherent semantically rich and structured nature of textual action descriptions, we argue that explicitly exploiting text embeddings associated with action labels is crucial to adapt our model on each episode.

Task-Conditioned Transductive Classifier. This module h follows a metric learning approach to classify the unlabeled samples of $\mathcal{Q}$ by matching them to the nearest class

Model	with Rich Textual Descriptions				with Short Class-Level Description			
	EK-92		SS-100		MetaUCF-101		Kinetics-100	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
ARN [34]	-	-	-	-	62.1	84.8	63.7	82.4
TSN++ [2]	39.1*	52.3*	33.6	43.0	76.4*	88.5*	64.5	77.9
CMN++ [2, 35]	-	-	34.4	43.8	-	-	65.4	78.8
TRN++ [2]	-	-	38.6	48.9	-	-	68.4	82.0
TAM [2]	-	-	42.8	52.3	-	-	73.0	**85.8**
TSN++ Transd. [15]	42.33	52.66	39.28	52.63	79.23	90.08	68.0	79.87
TNT	**46.13** $\pm$ 0.27	**59.00** $\pm$ 0.23	**50.44** $\pm$ 0.25	**59.04** $\pm$ 0.23	**86.66** $\pm$ 0.19	**94.14** $\pm$ 0.11	**78.02** $\pm$ 0.24	84.82 $\pm$ 0.19

Table 1. **Results on two families of datasets**. Those with rich textual descriptions: EK-92 and SS-100. Those with class-level textual descriptions: MetaUCF-101 and Kinetics-100. We report top-1 accuracy on the meta-testing sets for the 5-way tasks. *Obtained by us.

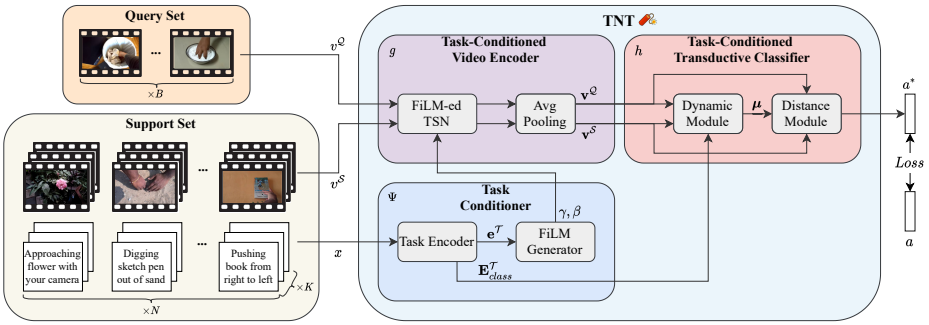

Figure 2. TNT model is composed by three parts. (I) Task-Conditioned Video Encoder g generates representations $\mathbf{v}^{\mathcal{Q}}, \mathbf{v}^{\mathcal{S}}$ of video sequences conditioned on parameters β and γ. (II) Task Conditioner Ψ takes video descriptions x to compute the text embeddings $\mathbf{e}^{\mathcal{T}}$ for generating modulation parameters β and γ, and the semantic class embedding $\mathbf{E}^{\mathcal{T}}_{class}$. (III) Task-Conditioned Transductive Classifier h takes the video representations $\mathbf{v}^{\mathcal{Q}}, \mathbf{v}^{\mathcal{S}}$ and the embedding $\mathbf{E}^{\mathcal{T}}_{class}$ to classify unlabeled samples following a transductive approach.

prototype. To obtain the class prototypes, we use a transductive approach that leverages the unlabeled samples from $\mathcal{Q}$ to augment the support set and subsequently improve the class prototypes based on the semantic class embedding $\mathbf{E}^{\mathcal{T}}_{\texttt{class}}$. Specifically, the Task-Conditioned Transductive Classifier consists of two components. (i) Dynamic Prototype Module. This module employs a cross-attention layer [28] to compute class-dependent relevance weights for each of the B samples in $\mathcal{Q}$ to augment $\mathcal{S}$. (ii) Distance Module. This module classifies the unlabeled instances of the query set by matching them to the nearest class prototype. To compute the distance between each instance and prototypes, we use a class-covariance-based distance (Mahalanobis) as in [1].

4. Experiments

Datasets. We evaluate our approach using two families of datasets: (i) those with rich and detailed textual descriptions of actions per video: Epic-Kitchens [4], Something-Something-V2 [8], and (ii) those with short class-level descriptions: UCF-101 [25] and Kinetics [10]. We propose for the first time to use Epic-Kitchens [4] as a benchmark for few-shot video classification.

Baselines. We compare the performance of our TNT model against state-of-the-art methods for few-shot video classification, namely TAM [2] and ARN [34]. We also consider additional stronger baselines, namely TSN++, TRN++ and CMN++ which are proposed by [2], following the practices from [3, 35]. Because our model makes use of a transductive setting, we also consider a transductive baseline named TSN++ Transd. This baseline is an extension of the image-based method [15].

Results. As it can be observed in Table 1, we achieve state-of-the-art-results in all standard benchmark metrics across the two tested families of datasets. Notably, our model achieves outstanding results when both spontaneous, unstructured and fine-grained descriptions and short and general descriptions are available, although it is designed to leverage the rich semantic information in fine-grained textual descriptions.

5. Conclusions

In this paper, we propose the Text-Conditioned Network with Transductive Inference (TNT), a novel few-shot model that leverages the fine-grained textual descriptions of the support instances to improve video understanding under a low-data regime. Unlike previous works, TNT uses text representations from a pre-trained language model to adapt and contextualize the feature encoder to each FSL task and improve class prototypes in a transductive setting. Our experiments show that our model outperforms a wide range of state-of-the-art models in four challenging datasets. Furthermore, our ablation study shows that the dynamic prototype module plays an important role in improving the 1-shot task. As an important finding, we verify that textual conditioning provides a more helpful signal than video-based conditioning to enhance the video feature encoder.

References

[1] Peyman Bateni, Raghav Goyal, Vaden Masrani, Frank Wood, and Leonid Sigal. Improved few-shot visual classification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2020. 2, 3

[2] 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. 1, 2, 3

[3] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *Int. Conf. Learn. Represent.*, 2019. 3

[4] Dima Damen, Hazel Doughty, Giovanni Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Computer Architecture Letters*, (01):1–1, 2020. 1, 2, 3

[5] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. 2

[6] Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and S. M. Ali Eslami. Conditional neural processes. In *Int. Conf. Machine learning*, volume 80, pages 1704–1713. PMLR, 2018. 2

[7] R. Girdhar, J. Carreira, C. Doersch, and A. Zisserman. Video action transformer network. *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 1

[8] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. In *Int. Conf. Comput. Vis.*, Oct 2017. 2, 3

[9] J. Ji, S. Buch, JC. Niebles, and A. Soto. End-to-end joint semantic segmentation of actors and actions in video. In *Eur. Conf. Comput. Vis.*, 2018. 1

[10] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017. 1, 2, 3

[11] Jongmin Kim, Taesup Kim, Sungwoong Kim, and Chang D. Yoo. Edge-labeling graph neural network for few-shot learning, 2019. 2

[12] Gregory Koch. Siamese neural networks for one-shot image recognition. In *Int. Conf. Machine learning*, 2015. 2

[13] Sai Kumar Dwivedi, Vikram Gupta, Rahul Mitra, Shuaib Ahmed, and Arjun Jain. Protogan: Towards few shot learning for action recognition. In *Int. Conf. Comput. Vis. Worksh.*, pages 0–0, 2019. 2

[14] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Int. Conf. Comput. Vis.*, October 2019. 1

[15] Jinlu Liu, Liang Song, and Yongqiang Qin. Prototype rectification for few-shot learning. In *Eur. Conf. Comput. Vis.*, August 2020. 2, 3

[16] Yanbin Liu, Juho Lee, Minseop Park, Saehoon Kim, Eunho Yang, Sungju Hwang, and Yi Yang. Learning to propagate labels: Transductive propagation network for few-shot learning. In *Int. Conf. Learn. Represent.*, 2019. 2

[17] Gary Lupyan and Benjamin Bergen. How language programs the mind. *Topics in Cognitive Science*, 8(2):408–424, 2016. 1

[18] 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. 2

[19] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms, 2018. 2

[20] Hang Qi, Matthew Brown, and David G Lowe. Low-shot learning with imprinted weights. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5822–5830, 2018. 2

[21] Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. Meta-learning with implicit gradients. In *Adv. Neural Inform. Process. Syst.*, pages 113–124, 2019. 2

[22] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3982–3992, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. 2

[23] James Requeima, Jonathan Gordon, John Bronskill, Sebastian Nowozin, and Richard E Turner. Fast and flexible multi-task classification using conditional neural adaptive processes. In *Adv. Neural Inform. Process. Syst.*, 2019. 2

[24] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Adv. Neural Inform. Process. Syst.*, pages 4077–4087. Curran Associates, Inc., 2017. 2

[25] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. 3

[26] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H. S. Torr, and Timothy M. Hospedales. Learning to compare: Relation network for few-shot learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1199–1208, 2018. 2

[27] Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6):544–557, 2009. 2

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Adv. Neural Inform. Process. Syst.*, volume 30, pages 5998–6008. Curran Associates, Inc., 2017. 3

[29] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In

Adv. Neural Inform. Process. Syst., pages 3630–3638, 2016. 2

[30] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(11):2740–2755, 2019. 1, 2

[31] Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM Comput. Surv.*, 53(3), June 2020. 1

[32] Yongqin Xian, Bruno Korbar, M. Douze, B. Schiele, Zeynep Akata, and L. Torresani. Generalized many-way few-shot video classification. In *Eur. Conf. Comput. Vis. Worksh.*, 2020. 1, 2

[33] Baohan Xu, Hao Ye, Yingbin Zheng, Heng Wang, Tianyu Luwang, and Yu-Gang Jiang. Dense dilated network for few shot action recognition. ICMR '18, page 379–387, New York, NY, USA, 2018. Association for Computing Machinery. 2

[34] Hongguang Zhang, Li Zhang, Xiaojuan Qi, Hongdong Li, Philip H. S. Torr, and Piotr Koniusz. Few-shot action recognition with permutation-invariant attention. In *Eur. Conf. Comput. Vis.*, 2020. 1, 2, 3

[35] Linchao Zhu and Yi Yang. Compound memory networks for few-shot video classification. In *Eur. Conf. Comput. Vis.*, September 2018. 2, 3

[36] Xiatian Zhu, Antoine Toisoul, Juan-Manuel Perez-Rua, Li Zhang, Brais Martinez, and Tao Xiang. Few-shot action recognition with prototype-centered attentive learning. *arXiv preprint arXiv:2101.08085*, 2021. 2