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

Andrés Villa[1]
afvilla@uc.cl
Juan-Manuel Perez-Rua*[2]
jmpr@fb.com
Victor Escorcia†[2]
v.castillo@samsung.com
Vladimir Araujo[1,4]
vgaraujo@uc.cl
Juan Carlos Niebles[3]
jniebles@cs.stanford.edu
Alvaro Soto†[1]
asoto@ing.puc.cl

[1] Pontificia Universidad Católica de Chile
Santiago, Chile
[2] Samsung AI Centre Cambridge
Cambridge, UK
[3] Stanford University
Stanford, CA, USA
[4] KU Leuven
Leuven, Belgium

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 benchmarks commonly used to evaluate few-shot video action classification models.

1 Introduction

Humans use language to guide their learning process [25]. 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 abil-

†Equal advising.

*Current affiliation: Facebook AI, London, UK.

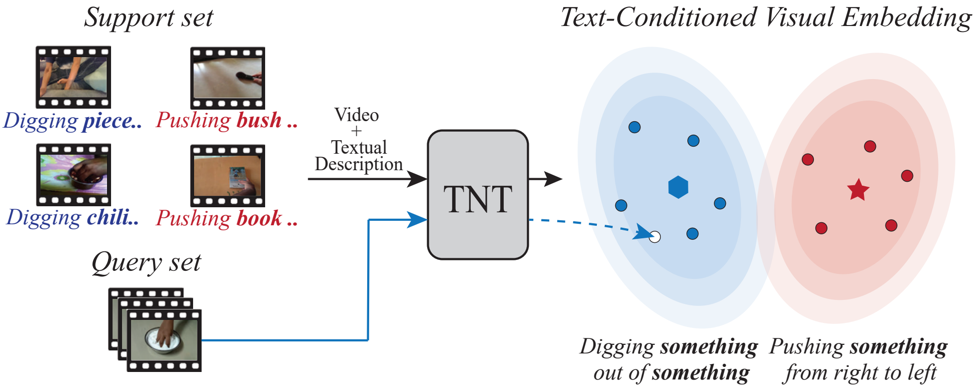
Support set
Digging piece..
Pushing bush ..
Digging chili..
Pushing book ..
Query set
Video + Textual Description
TNT
Text-Conditioned Visual Embedding
Digging something out of something
Pushing something from right to left

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.

ity to quickly learn from few examples because they help to transfer learning among tasks, disambiguate and correct error sources [25]. However, modern deep learning approaches in action recognition [15, 22, 40] have mainly focused on a large amount of labeled visual data ignoring the textual descriptions that are usually included along with the videos [10, 16]. These limitations have motivated an increasing interest in Few-Shot Learning (FSL) [41], which consists of learning novel concepts from few labeled instances.

While most FSL models are focused on image classification [1, 7, 17, 23, 24, 30, 32, 33, 36, 39], few works [3, 19, 26, 44, 45, 47, 48] are dedicated to video classification. Recognizing actions in video with only a few training samples is arguably more challenging than the image classification case. The video content is richer, and action classes exhibit large intra-class variations. For example, the action "Digging something out of something" in Fig. 1 looks significantly different as it involves interactions with two different objects. Therefore, extending existing FSL approaches for image classification to the case of video is not trivial.

The few existing video FSL methods follow one of two approaches: (i) exploiting the temporal and spatial dimensions in videos [3, 45]; or (ii) taking advantage of large amounts of additional video data by using tag retrieval to overcome the labeled data scarcity [42]. 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 [5]. During the Epic-Kitchens [5] 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, for the first time, we introduce a new class of models: **T**ext-conditioned **N**etworks with **T**ransductive inference or TNT. This method exploits the knowledge that is available in text descriptions as a privileged source of information [37] to improve class discrimination in few-shot video classification, see Fig. 1. TNT is built on top of a primary backbone that aims to encode global and extensive knowledge about the visual world. TNT further contains a complementary secondary network trained to extract task-specific knowledge from the support textual descriptions, leveraging the modern language models [31]. This secondary network contextualizes the global knowledge of the primary network according to the semantic information of the task. Moreover, TNT spans a third module, which leverages the detailed textual information of the few support videos to augment them with those unlabeled (query) to obtain more confident class representations (prototypes), following a transductive setup. These prototypes serve as a proxy for the classification of the query instances using a nearest neighbors approach. Overall, the integration of these three networks allows our model to quickly adapt to the challenging data conditions of FSL tasks.

In summary, our main contributions are: (**I**) To the best of our knowledge, we propose the first FSL video action classification method that leverages the semantic information in textual action descriptions of the support data to modulate the visual feature encoder. (**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) [3] and the new benchmark Epic-Kitchens-92 (EK-92), and those with short class-level textual descriptions such as MetaUCF-101 [26] and Kinetics-100 [47].

2 Related Work

Few-Shot Learning. It is possible to identify two main groups in the FSL literature: (i) gradient based methods and (ii) metric learning based methods. Gradient-based methods focus on learning a good parameter initialization that facilitates model adaptation by few-shot fine-tuning [7, 27, 30]. On the other hand, metric-based methods aim to learn or design better metrics for determining similarity of input samples in the semantic embedding space [18, 29, 33, 36, 39]. More recently, affine conditional layers are added to the feature extraction backbone in [1, 32] as extension to the conditional neural process framework [9] with the goal of effective task-adaptation. In this work, we extend this framework [9] differently from [1, 32] by adapting the feature extractor and updating the class representations based on the support textual descriptions and query instances. Our goal is to influence the visual backbone with the structured knowledge captured by pre-trained language models.

Induction vs Transduction in FSL. Regarding the inference setup, there are two types of approaches: inductive and transductive FSL. In the inductive setting, only the support instances are used to guide the inference process. In contrast, in the transductive setting, the model uses extra information from query samples to perform its inference [23]. We are motivated by recent work following the transductive setting [17, 23, 24, 27], where the unlabeled query data is exploited to further refine the few-shot classifier. For instance, [23] proposes a prototype rectification approach by label propagation. Departing from previous work, our model proposes a novel transductive approach that takes advantage of the support textual descriptions to augment the support videos with the unlabeled instances, leveraging the cross-attention approach.

Few-Shot Video Classification. With the shift of action recognition research from coarse [16] to fine-grained categories [5, 11], the problem of data scarcity has intensified. A few works to tackle this issue have appeared recently. However, most of works focused only on better exploiting visual or temporal information from videos [3, 19, 26, 42, 44, 45, 47, 48]. Additionally, the approach proposed in [42] uses extra video data and annotations to learn a more suitable representation before meta-training. Although tackling important aspects in video data modeling, none of the previous works offer solutions to the semantic gap between the few-shot samples and the nuanced and complex concepts needed for video representation learning. We aim to bridge this gap by using textual descriptions as privileged information to contextualize the video feature encoder in conjunction with a classification approach based on class prototypes acting under a transductive inference scheme.

Exploiting Text Embeddings in Small Data. Prior works have leveraged multi-modal information to enhance few-shot visual classification [21, 28, 43, 44], where textual description has been widely used for image data. Likewise, Zero-Shot Learning (ZSL) methods for image classification uses text descriptions to classify samples from novel unseen classes [20, 34]. These descriptions focus on nouns that have a structured taxonomy and can be associated with specific regions of the input image. Conversely, actions are defined by verbs that are usually more overloaded and more fine-grained than nouns [6]. Therefore, the extrapolation of these approaches to the case of action recognition is not straightforward. Currently, there are some relevant works in ZSL for video classification [2, 8, 12, 13, 14]. [2, 12, 13] learn a static video encoder to map the videos to an embedding space very close to the semantic representation of their labels. It does not allow these methods to learn to exploit the Spatio-temporal information of the videos specifically, limiting their generalization power in scenarios where some support instances are available. That is why we employ a two steps training process and adaptative method. First, we learn a general video encoder from base classes. Later, the general video encoder is fixed, and our method learns to adapt the video encoder and a transductive classifier to the novel classes using textual descriptions. Likewise, [8, 14] use a static video representation to get the relevant objects in the video and later computes their semantic textual embedding to be the bridge between known and unknown actions labels. In this sense, this method depends on the semantic relation between the action label and its objects, which could be a problem for fine-grained action datasets like [11], where the objects are related to several classes of actions.

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 [33, 39], 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. The support set $\mathcal{S}$ 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 [39]. 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 $\mathcal{S}$.

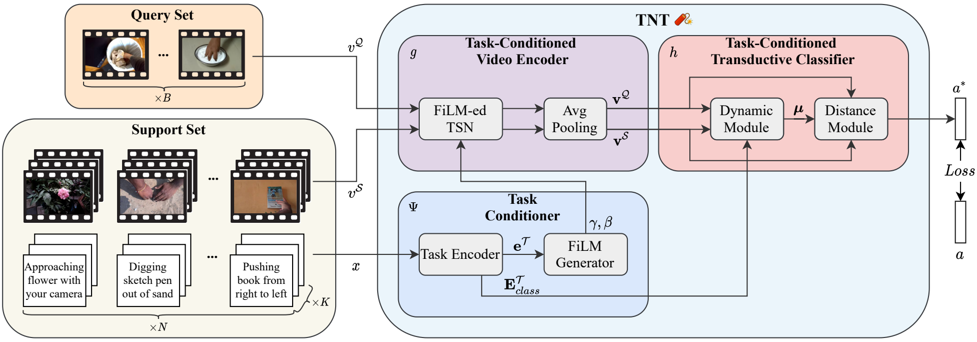
Query Set
×B
Support Set
Approaching flower with your camera
Digging sketch pen out of sand
Pushing book from right to left
×K
×N
$v^{\mathcal{Q}}$
$v^{\mathcal{S}}$
x
TNT
g
Task-Conditioned Video Encoder
FiLM-ed TSN
Avg Pooling
$\mathbf{v}^{\mathcal{Q}}$
$\mathbf{v}^{\mathcal{S}}$
h
Task-Conditioned Transductive Classifier
Dynamic Module
$\boldsymbol{\mu}$
Distance Module
a^*
Loss
a
Ψ
Task Conditioner
γ, β
Task Encoder
$\mathbf{e}^{\mathcal{T}}$
FiLM Generator
$\mathbf{E}^{\mathcal{T}}_{class}$

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.

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, 32], 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. This module g transforms the lower-level visual information of each video v into a more compact and meaningful representation $\mathbf{v}$. To handle novel action classes at test time, it is essential to provide g with a flexible adaptation mechanism that selectively focuses and/or disregards the latent information of its internal representation across different episodes. To achieve this, we employ the TSN video architecture 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. Thereby, this module computes frame-level feature embeddings for each video followed by an adaptive average pooling that summarizes the spatiotemporal information to obtain the video representation $\mathbf{v} = g(v)$ where $v \in \mathbb{R}^{T \times H \times W}$ and $\mathbf{v} \in \mathbb{R}^{G}$.

We use the widely-adopted video frame sampling strategy of temporal segment networks (TSN) [22, 40, 46, 49]. Contrary to the CNAPS strategy [1, 32], we train Task-Conditioned Video Encoder on the base classes within a fully supervised regime rather than on a large dataset. That is due to the variability between the video datasets and their actions. So that, we have to train it for few epochs to avoid overfitting.

Task Conditioner. The Task Conditioner Ψ is an essential part of our approach that provides high adaptability to our model. Specifically, it computes 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

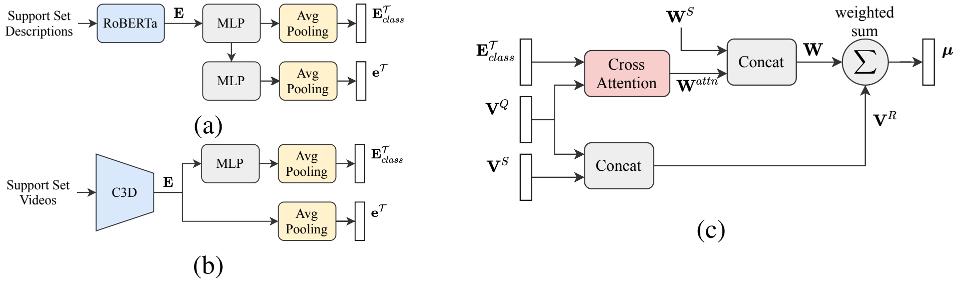

Figure 3: **Task conditioner architectures and Dynamic prototype module.** We propose two kinds of task conditioner: (a) a text-based conditioner based on the RoBERTa model. (b) a video-based conditioner, and (c) a dynamic prototype module based on an attention approach to augment the support set $\mathcal{S}$ samples.

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.

We assume that the instances of the support set $\mathcal{S}$ are a triad (v, x, a) corresponding to a video v, textual action description x, and categorical action label a, respectively. Furthermore, the Task Conditioner subsumes two components:

(a) Task Encoder. This module generates the conditioning signals: (i) the task embedding $\mathbf{e}^{\mathcal{T}}$ to tune the Task-Conditioned Video Encoder g, and (ii) the semantic class embedding $\mathbf{E}^{\mathcal{T}}_{\texttt{class}}$ used to tune the Task-Conditioned Transductive Classifier, given the textual action description x in the support set $\mathcal{S}$. Specifically, our Task Encoder consists of the RoBERTa language model [31] followed by two multilayer perceptrons, as shown in Fig. 3-a. Using RoBERTa, we compute the sample-level text embedding $\mathbf{E}$ of each x. These text representations are projected first through linear layer and average-pooled along the number of shots K, resulting in the class embedding $\mathbf{E}^{\mathcal{T}}_{\texttt{class}} \in \mathbb{R}^{N \times G}$. Additionally, $\mathbf{E}$ is linearly projected a second time to obtain the task embedding $\mathbf{e}^{\mathcal{T}} \in \mathbb{R}^{1 \times L}$.

(b) FiLM Generator. It generates the set of affine parameters γ_i, β_i for every stage i of g to effectively modulate our Task-Conditioned Video Encoder given the task embedding $\mathbf{e}^{\mathcal{T}}$.

In practice, we tune the MLP modules and the FiLM generator parameters in a subsequent training stage after fixing g [1]. The RoBERTa module is initialized from a pre-trained sentence representation and remained unchanged to take advantage of its prior knowledge, avoiding overfitting due to its high number of parameters. Also, note that our Task Conditioner module is conceptually different to the one presented in [1, 32]. While the encoder in [1, 32] is purely a function of the visual instances in the support set $\mathcal{S}$, in our case, we leverage textual descriptions of target categories.

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 prototype. To obtain the class prototypes, a straightforward approach is to compute a class-wise average by considering the K-examples in the support set $\mathcal{S}$ [1, 4, 33]. However, due to the data scarcity, these prototypes are usually biased. To alleviate this problem, we use a transductive classifier that leverages the unlabeled samples to 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:

(a) Dynamic Prototype Module. This module leverages the semantic class embedding $\mathbf{E}^{\mathcal{T}}_{\texttt{class}}$ to get the most relevant unlabeled samples for every class, see Fig. 3-c. Thus, effectively augmenting the support set with unlabeled samples in $\mathcal{Q}$ and subsequently improving the class prototypes. Specifically, we employ a cross-attention layer [38] to compute class-dependent relevance weights for each of the B samples in the query set $\mathbf{W}^{att} \in \mathbb{R}^{N \times B}$ by:

$$\mathbf{W}^{att} = \texttt{softmax}\left(\frac{\mathbf{E}^{\mathcal{T}}_{\texttt{class}}\mathbf{W}^{Q}\left(\mathbf{V}^{\mathcal{Q}}\mathbf{W}^{K}\right)^{T}}{\sqrt{G}}\right), \tag{1}$$

where $\mathbf{W}^{Q}$, $\mathbf{W}^{K} \in \mathbb{R}^{G \times G}$ are linear projections for the query and keys. In our case, we use $\mathbf{E}^{\mathcal{T}}_{\texttt{class}}$ as query to look up into the sequence of all video representations $\mathbf{V}^{\mathcal{Q}} = [\mathbf{v}^{\mathcal{Q}}_{1}, ..., \mathbf{v}^{\mathcal{Q}}_{B}]$ in $\mathcal{Q}$. Furthermore, we calculate relevance weights $\mathbf{W}^{\mathcal{S}} \in \mathbb{R}^{N \times NK}$ for the support samples in $\mathcal{S}$, we assume that all of them have equal importance for the class i they belong to. Thus, we define the relevance weights $\mathbf{W} = [\mathbf{W}^{att} \ \ \mathbf{W}^{\mathcal{S}}]$ for all the task samples $\mathcal{R} = \mathcal{Q} \cup \mathcal{S}$, as

$$\mathbf{W}_{ij} = \begin{cases} \mathbf{W}^{att}_{ij}, & \mathbf{v}^{\mathcal{R}}_{j} \in \mathcal{Q} \\ 1/K, & i = a_j, \left(\mathbf{v}^{\mathcal{R}}_{j}, a_j\right) \in \mathcal{S}, \\ 0, & i \neq a_j, \left(\mathbf{v}^{\mathcal{R}}_{j}, a_j\right) \in \mathcal{S} \end{cases} \tag{2}$$

where $\mathbf{W} \in \mathbb{R}^{N \times (B+NK)}$. Finally, we calculate the class prototypes through a weighted sum of all samples $\mathbf{V}^{\mathcal{R}}$ attending the relevance weights $\mathbf{W}$:

$$\mu_i = \frac{1}{\sum_{j=1}^{B+NK} \mathbf{W}_{ij}} \sum_{j=1}^{B+NK} \mathbf{W}_{ij}\mathbf{v}^{\mathcal{R}}_{j}, i \in N \tag{3}$$

(b) 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]. We train our model by minimizing: $p(a^{*}_{j} = i | f(v_j), \mathcal{S}) = \texttt{softmax}(-d_i(f(v^{*}_{j}), \mu_i))$, where $j \in B$, a^{*}_{j} is the predicted class for the unlabeled sample v^{*}_{j}, and μ_i is the prototype of class i obtained with the dynamic module. Also, d_i is the distance function that receives the class prototypes explicitly and computes the task prototype by taking the average of these prototypes.

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 [5], Something-Something-V2 [11], and (ii) those with short class-level descriptions: UCF-101 [35] and Kinetics [16]. We propose for the first time to use Epic-Kitchens [5] as a benchmark for few-shot video classification. We coin this new benchmark **Epic-Kitchens-92** (EK-92). [5] features spontaneous actions accompanied with human narrations. Interestingly, a particular action class could encompass diverse narrations, *e.g.*, the action class: "*Put something*" features narrations such as: "*Put plate down*", "*Place aubergine onto pan*". To ensure that action classes are consistent, we use the 97 verb classes defined by [5] and select those with more than 5 instances, yielding 92 action classes. Then, we divide the resulting 92 classes into 58, 11, and 23 for meta-training, meta-validation, and meta-testing, respectively. For the other benchmarks, we follow the evaluation protocol proposed by [3, 26, 47] termed **Something-Something-100** (SS-100), **MetaUCF-101**, and **Kinetics-100**, respectively. The protocols

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 [45]	-	-	-	-	62.1	84.8	63.7	82.4
TSN++ [3]	39.1*	52.3*	33.6	43.0	76.4*	88.5*	64.5	77.9
CMN++ [3, 47]	-	-	34.4	43.8	-	-	65.4	78.8
TRN++ [3]	-	-	38.6	48.9	-	-	68.4	82.0
TAM [3]	-	-	42.8	52.3	-	-	73.0	**85.8**
TSN++ Transd. [23]	42.33	52.66	39.28	52.63	79.23	90.08	68.0	79.87
TNT	**46.13** ± 0.27	**59.00** ± 0.23	**50.44** ± 0.25	**59.04** ± 0.23	**86.66** ± 0.19	**94.14** ± 0.11	**78.02** ± 0.24	84.82 ± 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.

in [3, 47] define a set of 64 classes for meta-training, 12 classes for meta-validation, and 24 classes for meta-testing, which are sampled randomly from Something-Something-V2 and Kinetics, respectively. In terms of the protocol in [26], it samples randomly 70 classes from UCF-101 for meta-training, 10 classes for meta-validation, and 21 for meta-testing. To facilitate the comparison, we use the same partitions proposed by the original authors.

Additionally, we make use of the provided text-based action descriptions from every meta-dataset. In Kinetics-100 and MetaUCF-101, we directly employ the class labels (*e.g.*, *Headbanging*, *Stretching leg*, etc), which are the same for all samples that belong to the same class. Alternatively, EK-92 and SS-100 provide a fine-grained textual action description per instance based on the action and objects depicted in the video. Further details about these datasets can be found in the Supplementary Material.

Implementation Details. We train our model following the episodic learning approach to mimic the meta-testing conditions [36]. For this purpose, we assemble N-way and K-shot tasks, selecting N classes randomly with K samples for the support set and B unlabeled samples for the query set. Thus, each episode has $NK + B$ videos. We report results for the 5-shot and 1-shot tasks, each with 5-ways and 50 elements of these classes in the query set (10 elements per class). Our model was trained during 15×10^3 episodes with the same data augmentation proposed in [40], using $T = 8$ frames per video. We calculate the mean accuracy by sampling 10^4 episodes (for a total of 5×10^5 queries) to test our model. In regards to the FiLM generator and the distance module, we follow the design choices in [1]. We optimize our model using task batch size of 16 and Adam with a learning rate of 5×10^{-4} for EK-92, SS-100 and MetaUCF-101, and 1×10^{-4} for Kinetics-100.

For our feature encoder, we use a TSN Network with a ResNet-50 backbone pre-trained on ImageNet and augmented with FiLM layers. For the first training stage of the TSN backbone, we tune it during 12 epochs using the setting proposed in [40].

Baselines. We compare the performance of our TNT model against state-of-the-art methods for few-shot video classification, namely TAM [3] and ARN [45]. We also consider additional stronger baselines, namely TSN++, TRN++ and CMN++ which are proposed by [3], following the practices from [4, 47]. 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 [23] which adopts a pseudo-labeling strategy to augment the support set. Conversely, the method proposed by [42] is not considered because it relies on a large amount of additional data, and its evaluation protocol is different from ours and from the one used in the baselines. Specifically, it uses the whole video instead of a segment of it.

Impact of rich textual descriptions. As it can be observed in Table 1, we achieve state-of-the-art-results in all standard benchmark metrics across the two tested datasets with rich

Model	**Inference Type**	**Task Encoder**		**SS-100 (Val)**	
		Video	**Text**	**1-shot**	**5-shot**
TSN++	Inductive	✗	✗	35.72 ± 1.37	49.4 ± 1.24
T-TSN+CNAPS	Transd	✓	✗	42.20 ± 1.43	57.68 ± 1.37
VNI	Inductive	✓	✗	33.53 ± 1.07	50.51 ± 1.30
TNI	Inductive	✗	✓	34.27 ± 1.24	56.21 ± 1.22
VNT	Transd.	✓	✗	41.5 ± 1.53	56.60 ± 1.23
TNT	Transd.	✗	✓	**47.98** ± 1.41	**60.18** ± 1.21

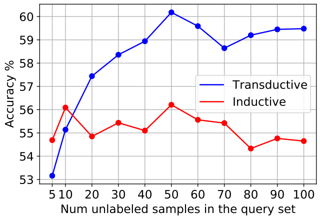

Table 2: (Left) **Ablation study results in the validation set.** We consider: a Video-conditioned Networks with Inductive (VNI) and Transductive (VNT) inference, Text-conditioned Networks with Inductive (TNI) and Transductive (TNT) inference, the TSN++ baseline, and the T-TSN+CNAPS baseline based on [1, 23].

Figure 4: (Right) **Sensitivity analysis to the number of query set samples.** Model performance in the 5-way, 5-shot task for different B size.

textual descriptions per instance. Notably, our model achieves outstanding results in EK-92, where it must handle spontaneous and unstructured descriptions. Likewise, our model improves over the TSN++ transductive baseline by around 7% and 4% in the 5-shot and 1-shot tasks, respectively, which shows the relevance of using the textual descriptions to modulate the network and make a transductive inference. It is worth noting that TSN-based backbone does not have a strong temporal modeling capacity. This is in sharp contrast to TAM [3], which is specially designed to capture temporal information. Despite these disadvantages, our method is able to outperform this strong alternative by around 8% in the 5-shot and 1-shot tasks on SS-100.

Impact of short class-level textual descriptions. Table 1 shows the performance of our model in datasets without rich textual description. Notably, our model outperforms the state-of-the-art baselines on MetaUCF-101 by a large margin. On Kinetics-100, TNT beats the TAM model, which is the best baseline in this benchmark, in the 1-shot task by 3.5%, while in the 5-shot, the result remains competitive. A possible reason for such results is that, unlike MetaUCF-101, Kinetics-100 has significantly fewer training instances. These results suggest that TNT achieves outstanding performance with short class-level descriptions, although it is designed to leverage the rich semantic information in fine-grained textual descriptions.

Ablation Study. We analyze the impact of our proposed text encoder and transductive inference approach. To this end, we train and evaluate our model with two different task encoders (text and video) and two inference approaches (Inductive and Transductive). Specifically, we consider video-conditioned networks with inductive (**VNI**) and transductive (**VNT**) inference; and text-conditioned networks with inductive (**TNI**) and transductive (**TNT**) inference. The video-based conditioner module consists of a C3D model with a Conv-4-64 backbone to extract significant spatio-temporal information at low computational cost [45], see Fig. 3-b. Moreover, we consider the TSN++ [3] as our primary baseline, which is purely based on [40] and inductive prototypical learning. For a fair comparison, we also implement T-TSN+CNAPS, a variant of TSN++ with modulation [1] and transductive inference based on label propagation [23]. Table 2 shows our results. All the results are computed in the validation set using the same visual backbone model (ResNet-34).

It is important to note that feature encoder adaptation generates a substantial increase in model performance with regard to the TSN++ baseline across all evaluation modalities. This is generally true for both the VNI and TNI models. Crucially, the TNI model yields a performance gain of 6% and 1% in 5-shot and 1-shot tasks on the SS-100 over VNI, respectively.

Likewise, there is a further performance increase, especially in the 1-shot task when the dynamic module is included to perform a transductive inference (TNT and VNT). Interestingly, the 1-shot task is the most data-deprived testing set-up, which speaks positively about the effectiveness of our transductive model. It should be noted that our TNT model outperforms the T-TSN+CNAPS and VNT models by a large margin on both 5-shot and 1-shot tasks. This proves the relevance of using textual descriptions to modulate or contextualize the video feature encoder and improve the class prototypes in a transductive approach.

Effect of Query Set Size. We assess the TNT model sensitivity to the number of instances in the query set. This study can be observed in Fig. 4. We evaluate our model trained on SS-100 in the 5-way, 5-shot task with $B = 50$, increasing the value of B from 5 to 100. Model performance increases until the number of query samples $B = 50$ after which it remains almost constant. We hypothesize that this is due to a saturation point on the amount of extra information that can be extracted from query samples.

We also conduct qualitative evaluations to demonstrate how our model works and the relevance of using textual descriptions to modulate the visual feature encoder and perform a transductive inference. They are shown in the supplementary material.

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.

Acknowledgements

This work was supported in part by the Millennium Institute for Foundational Research on Data (IMFD).

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] Biagio Brattoli, Joseph Tighe, Fedor Zhdanov, Pietro Perona, and Krzysztof Chalupka. Rethinking zero-shot video classification: End-to-end training for realistic applications. In *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2020.

[3] 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.

[4] 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. URL https://openreview.net/forum?id=HkxLXnAcFQ.

[5] 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.

[6] B. Ghanem F. Caba, V. Escorcia and J.C. Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015.

[7] 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. URL http://proceedings.mlr.press/v70/finn17a.html.

[8] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. I know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs. *AAAI*, 33:8303–8311, 07 2019. doi: 10.1609/aaai.v33i01.33018303.

[9] 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.

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

[11] 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.

[12] Meera Hahn, Andrew Silva, and James M. Rehg. Action2vec: A crossmodal embedding approach to action learning, 2019.

[13] Sheng-Hung Hu, Yikang Li, and Baoxin Li. Video2vec: Learning semantic spatio-temporal embeddings for video representation. pages 811–816, 12 2016. doi: 10.1109/ICPR.2016.7899735.

[14] Mihir Jain, Jan C. van Gemert, Thomas Mensink, and Cees G. M. Snoek. Objects2action: Classifying and localizing actions without any video example. In *Int. Conf. Comput. Vis.*, December 2015.

[15] 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.

[16] 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. URL http://arxiv.org/abs/1705.06950.

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

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

[19] 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.

[20] Chung-Wei Lee, Wei Fang, Chih-Kuan Yeh, and Yu-Chiang Frank Wang. Multi-label zero-shot learning with structured knowledge graphs. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1576–1585, 2018.

[21] Jieya Lian, Haojie Wang, and ShengWu Xiong. Learning class prototypes via anisotropic combination of aligned modalities for few-shot learning. In *Int. Conf. Multimedia and Expo*, pages 1–6, 2020. doi: 10.1109/ICME46284.2020.9102883.

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

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

[24] 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. URL https://openreview.net/forum?id=SyVuRiC5K7.

[25] Gary Lupyan and Benjamin Bergen. How language programs the mind. *Topics in Cognitive Science*, 8(2):408–424, 2016. doi: https://doi.org/10.1111/tops.12155. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/tops.12155.

[26] 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.

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

[28] Frederik Pahde, Moin Nabi, T. Klein, and P. Jähnichen. Discriminative hallucination for multi-modal few-shot learning. *IEEE Int. Conf. Image Process.*, pages 156–160, 2018.

[29] 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.

[30] 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.

[31] 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, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://www.aclweb.org/anthology/D19-1410.

[32] 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.

[33] 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. URL http://papers.nips.cc/paper/6996-prototypical-networks-for-few-shot-learning.pdf.

[34] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Adv. Neural Inform. Process. Syst.*, pages 935–943. Curran Associates, Inc., 2013. URL http://papers.nips.cc/paper/5027-zero-shot-learning-through-cross-modal-transfer.pdf.

[35] 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. URL http://arxiv.org/abs/1212.0402.

[36] 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. doi: 10.1109/CVPR.2018.00131.

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

[38] 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.

[39] 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.

[40] 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.

[41] 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. ISSN 0360-0300. doi: 10.1145/3386252. URL https://doi.org/10.1145/3386252.

[42] 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.

[43] Chen Xing, Negar Rostamzadeh, Boris N. Oreshkin, and Pedro O. Pinheiro. Adaptive cross-modal few-shot learning. In *Adv. Neural Inform. Process. Syst.* Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/8731-adaptive-cross-modal-few-shot-learning.pdf.

[44] 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. ISBN 9781450350464. doi: 10.1145/3206025.3206028. URL https://doi.org/10.1145/3206025.3206028.

[45] 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.

[46] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Eur. Conf. Comput. Vis.*, September 2018.

[47] 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.

[48] 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.

[49] Mohammadreza Zolfaghari, Kamaljeet Singh, and Thomas Brox. Eco: Efficient convolutional network for online video understanding. In *Eur. Conf. Comput. Vis.*, September 2018.