TARN: Temporal Attentive Relation Network for Few-Shot and Zero-Shot Action Recognition

Mina Bishay
m.a.t.bishay@qmul.ac.uk
Georgios Zoumpourlis
g.zoumpourlis@qmul.ac.uk
Ioannis Patras
i.patras@qmul.ac.uk

Multimedia and Vision Research Group
Queen Mary University of London
London, UK

Abstract

In this paper we propose a novel Temporal Attentive Relation Network (TARN) for the problems of few-shot and zero-shot action recognition. At the heart of our network is a meta-learning approach that learns to compare representations of variable temporal length, that is, either two videos of different length (in the case of few-shot action recognition) or a video and a semantic representation such as word vector (in the case of zero-shot action recognition). By contrast to other works in few-shot and zero-shot action recognition, we a) utilise attention mechanisms so as to perform temporal alignment, and b) learn a deep-distance measure on the aligned representations at video segment level. We adopt an episode-based training scheme and train our network in an end-to-end manner. The proposed method does not require any fine-tuning in the target domain or maintaining additional representations as is the case of memory networks. Experimental results show that the proposed architecture outperforms the state of the art in few-shot action recognition, and achieves competitive results in zero-shot action recognition.

1 Introduction

Human action recognition has received significant attention in the last decade due to its application in areas like video surveillance, human-computer interaction, and video retrieval [13], as is the trend in most Computer Vision problems, using Deep Neural Networks [4, 26, 35, 39]. However, training deep architectures requires a large amount of annotated data, something that is not easily available for new action classes. By contrast, humans are able to recognize new actions using a few labelled examples or just an action-related description.

For this reason, Few-Shot Learning (FSL) and Zero-Shot Learning (ZSL) have recently received a lot of attention. Most of the FSL works follow the meta-learning approach where a high-level transferable knowledge is learned on a collection of different tasks. This knowledge, that helps to perform classification on the target few-shot task(s), can be good initial network weights [7], embedding functions [36, 40], or an external memory with useful information [25, 34]. Some works treat the few-shot problem as a similarity problem, that is, a

similarity model is trained to classify a query example by comparing it to labelled examples in the training set. These works are simple and need no additional memory or cost, however, most of them used a fixed distance metric to calculate the matching score. In [38], Sung *et al*. proposed a relation network that learns to calculate embeddings and a transferable deep measure of similarity between them. [38] achieved state-of-the-art results in image-based FSL, and shows competitive performance when applied to image-based ZSL.

However, most of the works have focused on image-based FSL and ZSL problems, like object recognition [7, 36, 38, 48], while relatively few works were aimed at video-based FSL and ZSL problems like action recognition [42, 47, 49]. Applying FSL and ZSL in videos is more challenging compared to images, due to the additional temporal dimension in videos and the variations that are introduced. To the best of our knowledge only one work has been proposed for few-shot action recognition [49]. However, this work is based on memory networks that require extra computational and space resources, and use a single embedding vector to represent the entire video. This might not capture well the temporal structure of the action and is challenging due to the amount of information existing in it. Xu *et al*. [45] proposed a method for action recognition from videos in limited data scenarios (i.e. when only a portion of the examples is available), however, training and testing is done on all the classes which is different from the few-shot protocol of [40] that is typically used in such problems. All ZSL approaches for action recognition use a single vector as a visual representation of entire videos, obtained either from handcrafted [50] or from deep [24] features. Working on video level, thus being incapable to explicitly leverage time-specific information, they miss more detailed visual cues that appear on the fine-grained level of the segments that form a video.

Our network (TARN) addresses the few-shot problem by working at video-segment level to calculate the relation scores between a query video and other sample videos – the query video is then assigned with the label of the most related video in the sample set. The relation/similarity is calculated in two stages: the embedding stage and the relation stage. In the embedding stage, a C3D [39] network followed by a layer of bidirectional Gated Recurrent Unit (GRU) [3], extract features from short segments of videos. The GRU learns an embedding function that is general and transferable over different tasks. In the relation stage, a segment-by-segment attention mechanism is used to align segment embeddings for a pair of query and sample videos, and then the aligned segments are compared. That is, we introduce segment-to-segment comparisons and model their temporal evolution, as a prior step towards video-to-video matching. The comparison outputs are fed to a deep neural network that learns a general deep distance measure for video matching, and gives at its output the final relation score for a pair of videos. Our approach can generalise to video-based ZSL, where no videos are available for the training in the sample set, but instead class descriptions (e.g. attribute vectors) are given. In this case, the query and the sample set have different types of data (visual and semantic, respectively) and therefore we use two different embedding modules, one for each domain. TARN (excluding the C3D model) is trained in an end-to-end manner in both FSL and ZSL cases. The main contributions of our work are three-fold:

1. We propose a relation network for few-shot and zero-shot action recognition. The proposed architecture compares either segment-wise visual features from a pair of videos (in FSL), or segment-wise visual features from a video with a class-wise semantic representation (in ZSL), retaining temporal information to finally perform video-wise classification.

2. The proposed architecture needs no additional resources like memory networks and does not require training or fine-tuning on the target problem like [7, 30]
3. We test the proposed architecture on different benchmark datasets, achieving state-of-the-art results in FSL and very competitive performance in ZSL.

2 Related work

Our approach aims to tackle the problems of few-shot and zero-shot action recognition, that are related to the following research directions:

Few-shot learning: Since the seminal work of Fei-Fei *et al*. in [5], several works have been proposed on learning from few examples, focusing mostly on image classification. In [7], Finn *et al*. addressed the problem of FSL by focusing on fast adaptability through proper initialization conditions. The danger of overfitting on few-shot tasks when adopting fine-tuning [30] has been noted in [40]. A remedy to this has been the episode-based strategy of [40], where at each episode K support examples from each one of C classes (where K and C are typically small), and one query example, are randomly chosen and the network weights are updated according to a loss defined over them. In this way, the generalization is improved without the need to perform weight updates on the support set during inference. Other works [18, 36, 40] treated FSL as a metric-learning problem, where the query samples are classified using either pre-defined or learned distance measures on learned embeddings. The closest work to our FSL method is [38], where query images are classified by comparing them to images from the support set. In [38], a relation module learns a non-linear similarity function to match images. In our work, we propose a relation module that first uses Euclidean distance and cosine similarity for comparing video segments, and then a trainable deep network for modeling the temporal distances/similarities across different segments and inferring a relation score for each pair of videos, is subsequently learned.

Zero-shot learning was initially defined as a problem in the works of Palatucci *et al*. [27] and Larochelle *et al*. [21]. Action recognition in the ZSL scenario typically requires bridging the semantic gap between the distributions of the semantic representations and the visual representations from the unseen classes [19, 42, 46]. The semantic representation of a class is a single vector, which can be either class-related attributes or word2vec embedding of the class label [22, 24]. The visual representations used in existing ZSL approaches are either handcrafted [29, 47, 50] based on the Improved Dense Trajectories (IDT) method [41], or deep [9, 24, 42] features extracted with C3D [39] network. IDT features represent a video with a single vector, by default. C3D features refer to 16-frame video segments, but most of the ZSL methods average them over the entire video to obtain a single visual video representation. In ZSL, the embeddings of the semantic representations are mapped to the embeddings of the visual representations, to establish relationships between the class-related attributes and the visual features. In [8], a ZSL method was proposed, leveraging the inter-class semantic relationships between the known and unknown actions. Other works have investigated using auxiliary data that are relevant to the unseen classes [47], using Wikipedia as an external ontology, to calculate semantic correlations between class labels [10], or constructing universal representations to achieve cross-dataset generalization [50]. However, even in datasets of trimmed videos, factors such as camera motion, viewpoint or action structure complexity can obstruct the extraction of meaningful features at some video parts. Hence, building deep networks that are capable of processing segment-level visual features, is a promising direction for ZSL.

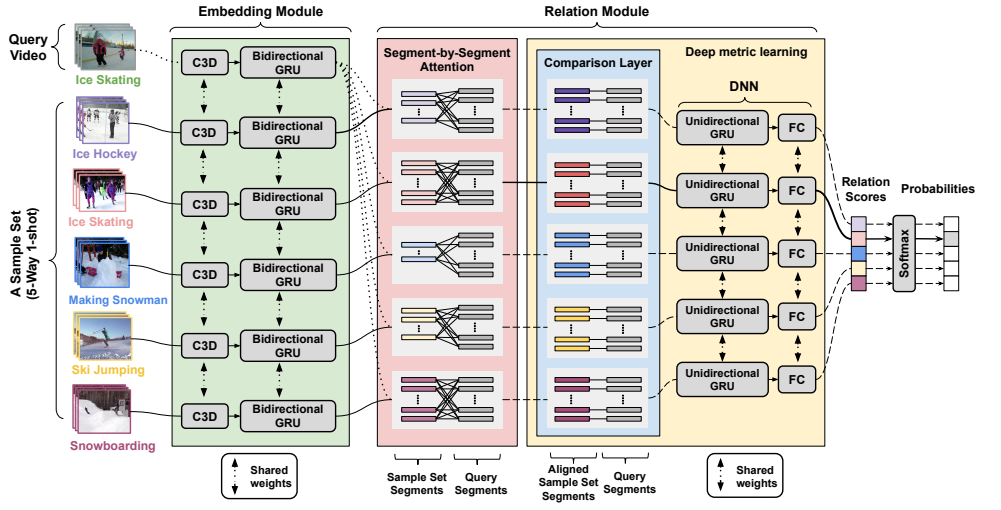

Figure 1: The proposed TARN architecture, consisting of the embedding module and the relation module. In an C-way K-shot task (where $K > 1$), the relation score of the query video to each class of the support set, is the average of the sample relation scores of that class.

Sequence matching: We claim that matching video sequences can benefit from comparing video segments as a first step. Previously, Fernando *et al*. [6] have proposed a method to match video segments from a pair of videos, that share similar temporal evolutions. We also draw our inspiration from works on text matching, where representations of sequence parts are compared and aggregated to match an entire text sequence [42]. An attention mechanism similar to that of [2, 31] is used to align the segments of each video in the support set, with respect to the query video. This is done to semantically align the features of the support set videos to the features of the query video. It also transforms the number of segments for each video of the support set to be equal to that of the query video.

3 Proposed Architecture

In this section we introduce a novel deep architecture, named Temporal Attentive Relation Network (TARN) for the problems of Few-Shot Learning (FSL) and Zero-Shot Learning (ZSL) for video-based tasks. Figure 1 shows an overview of the network. TARN learns to compare a query video against a sample set of videos in FSL, or semantic attributes in ZSL, representing a group of actions. In the FSL case, the inputs are segment-wise visual representations of the query and sample videos, and in the ZSL case the inputs are segment-wise visual representations of a query video on the one hand and semantic representations of the unseen classes on the other. The output is a relation score, either for each pair of videos, or for each pair of video and semantic attributes. More specifically, TARN consists of two modules: the embedding module and the relation module. First, the embedding module processes the visual or semantic representations, retaining the temporal structure of visual features, and produces embeddings that are later compared. Second, the relation module initially applies segment-by-segment attention. By doing so, it either transforms the

visual representations of the sample set to have the same number of segments as the query video (FSL), or transforms the semantic representations of the unseen classes to allow a segment-wise comparison between them and the visual representations (ZSL). Afterwards, a per-segment comparison is performed. Finally the relation module produces the matching score by taking into account the variations of the comparisons across all segments of the query video. These modules are explained in detail in the following subsections.

3.1 Embedding Module

TARN uses a single embedding module to embed both the query and sample videos in FSL, while two different embedding modules are used in ZSL. That is, one is used to embed the visual data while the other is used to embed the semantic data.

Video embedding: A pre-trained C3D network is used to extract spatio-temporal features across short segments of videos. This network acts as a local feature extractor in the embedding module. Then, a bi-directional GRU uses the local features to learn more globally-aware features, allowing each time step to access both backward and forward information across the whole video. Moreover, it reduces the dimension of the C3D features. This visual embedding module is applied in the same way to embed video segments in both the FSL and ZSL cases.

Semantic embedding: ZSL first compares visual representations extracted from segments of the query video to semantic representations of the sample set classes. The original semantic information needs to be encoded into a representation that allows a feature-rich comparison of video segments to class-related attributes, and therefore needs to have the same dimensions with the visual representation of the video segments. The semantic embedding module consists of two stacked fully-connected (FC) layers.

3.2 Relation Module

In this section we will explain the relation module from the FSL perspective, and then we will show how it is extended for ZSL. In FSL, in order to match a query video to a sample set of videos, firstly we pair the query video with each video in the sample set. Given pairs of videos, secondly we align segments in the videos using a segment-by-segment attention layer. The attention layer maps the sample video to have the same number of segment embeddings as the query. Third, each segment in the query is compared to the corresponding aligned sample segment. Fourth, the comparison outputs of the different segments are fed to a deep neural network, that learns a deep metric for video matching, and gives at its output a relation score for each pair. Finally, a softmax layer is used to map the relation scores to a probability distribution over the sample classes.

Segment-by-segment attention: Several recent works in text sequence matching and textual entailment use an attention mechanism, named word-by-word attention, to align the words of two given sentences [2, 28, 31, 44]. Similarly, as shown in the corresponding block of Fig. 1, we adopt the word-by-word attention in our architecture to align the sample and query segment-embeddings (i.e. segment-by-segment attention). Given a sample video $\boldsymbol{S} \in \mathbb{R}^{N \times d}$ and a query video $\boldsymbol{Q} \in \mathbb{R}^{M \times d}$, where each row in $\boldsymbol{S}$ and $\boldsymbol{Q}$ represents a segment-embedding vector of dimension d, and where N and M denote the number of segments in videos $\boldsymbol{S}$ and $\boldsymbol{Q}$ respectively. The segment-by-segment attention is calculated as follows:

$$\boldsymbol{A} = softmax((\boldsymbol{S}\boldsymbol{W} + \boldsymbol{b} \otimes \boldsymbol{e}_N)\boldsymbol{Q}^T), \qquad \boldsymbol{H} = \boldsymbol{A}^T\boldsymbol{S}, \tag{1}$$

where $\boldsymbol{W} \in \mathbb{R}^{d \times d}$ and $\boldsymbol{b} \in \mathbb{R}^{d}$ are parameters to be learned, and the operator "$\otimes \boldsymbol{e}_N$" repeats the bias vector $\boldsymbol{b}$, N times to form a matrix of dimension $N \times d$. $\boldsymbol{A} \in \mathbb{R}^{N \times M}$ is the attention weight matrix and $\boldsymbol{H}$ is the aligned version of $\boldsymbol{S}$. Each row vector in $\boldsymbol{H}$ is a weighted sum of the $\boldsymbol{S}$ segment-embeddings, and represents the parts of $\boldsymbol{S}$ that are most similar to the corresponding row vector (segment-embedding) of $\boldsymbol{Q}$. The row vectors of $\boldsymbol{Q}$ and $\boldsymbol{H}$ are used as inputs to a comparison layer.

Deep metric learning: The relation module performs deep similarity/distance metric learning by using a comparison layer and a non-linear classifier on the top of it. The comparison layer calculates a similarity measure between each of the M segments of the row vectors of $\boldsymbol{Q} \in \mathbb{R}^{M \times d}$ and $\boldsymbol{H} \in \mathbb{R}^{M \times d}$. This measure, as described in [44], can be based on one of the following operations: multiplication (Mult), subtraction (Subt), neural network (NN), subtraction and multiplication followed by a neural network (SubMultNN), or Euclidean distance and cosine similarity (EucCos). Since the measure is estimated at each of the M pairs of segments, the output of this layer has also M dimensions. This layer acts as an intermediate stage that produces low-level representations of the comparisons between the sample and the query segments. As we will show in the experimental section, decomposing the query-sample matching problem into several comparisons across segments performs better than just a single comparison of two vectors representing the sample and query videos – this coincides with the findings in text sequence matching problems [11, 28, 43].

Unlike other works in FSL which used a linear classifier or a fixed metric to match query and sample examples [18, 36], we follow [38] and use a deep neural network for deep metric learning. That is, the outputs of the comparison layer are passed to the deep network that learns a global deep metric over the entire videos. For FSL, we use a uni-directional GRU to learn temporal information across different segment comparisons and a Fully-Connected (FC) layer for giving the final relation score. The final relation scores coming from different sample examples are passed to a softmax layer, so that they can be mapped to a probability distribution over the sample classes. The query video is assigned with the label of the most related video in the sample set. In the case of multiple shots ($K > 1$) per class in the sample set, the mean of the relation scores over the shots of each class is taken as the relation score of the query video with that class.

The full architecture (excluding the C3D model) is trained in an end-to-end fashion. We use episode-based training scheme proposed in [36, 40] to train our architecture, and binary cross-entropy as the cost function. The total batch/episode cost is:

$$L(t,q) = -\frac{1}{KC} \sum_{k=1}^{K} \sum_{c=1}^{C} (t_{kc} \log q_{kc} + (1 - t_{kc}) \log(1 - q_{kc})) \tag{2}$$

where K denotes the number of shots, C the number of classes in each episode, t the target relation score, and q the predicted relation value.

In **zero-shot learning**, each of the sample set classes is expressed by a semantic attribute/word vector. In this case, the attention mechanism estimates the similarity between the semantic vector and each segment of the query video, instead of aligning video segments as in the FSL case. The semantic embedding module encodes class attributes in representations that allow finding relations between the underlying class attributes and query video segments. We perform several segment-to-attribute comparisons, leveraging this fine-grained information to improve the training process and reduce overfitting. The comparison outputs are aggregated over all query video segments using two FC layers and an average pooling layer that produces the final relation score. The comparison, FC, and pooling layers

represent in this case the deep network for metric learning. Note that the uni-directional GRU of the FSL case has been replaced, as there is no temporal alignment information to learn in the deep network.

4 Experimental results

4.1 Few-shot action recognition

Implementation details. In [49], Zhu and Yang introduced a dataset for few-shot video classification, that is a modification of the original Kinetics dataset [16]. In this work, we follow [49] and use their dataset and evaluation protocol. The dataset videos are sliced into fixed length segments of 16 frames each, and then these segments are fed to a C3D network pre-trained on Sports-1M [15]. Visual features are extracted from the last FC layer (i.e. FC7) of the C3D network, and used as input to the embedding bidirectional GRU. The dimension of the C3D features is 4096, and the GRU has a hidden state of size 256. In the relation module, the output of the comparison layer is used as input to a unidirectional GRU layer of size 256, and the GRU output at the last time step is fed to an FC layer with a single neuron for predicting the final relation score. We use $20,000$ episodes for training, 500 for validation, and $1,000$ for testing. Each episode has a sample set of 5 classes. We evaluate the performance of our architecture on the validation set every 500 training episodes. The best-performing model on the validation set is used for testing. We train our architecture using Stochastic Gradient Descent (SGD) with momentum $m = 0.9$, and learning rate equal to 10^{-3} for the first $10k$ training episodes, and 10^{-4} for the last $10k$ episodes.

Results. In our first experiment, we investigate the impact of the various functions that can be used as distance/similarity measure in the comparison layer, on the TARN performance. Following [44], we compare five different distance measures (Mult, Subt, NN, SubMultNN, EucCos). Table 1 shows the accuracy obtained by the TARN model over the different measures. EucCos leads to the best accuracy over all shots. Although EucCos is a fixed measure with no learnable parameters, the following layers in the relation module are trainable and non-linear.

In the next experiment, we investigate the benefits of using segment-by-segment attention and comparing segment-wise representations in our architecture. To do so, we compare the TARN model to another model that has no attention layer and performs a single comparison for each video pair. Specifically, we modify the embedding module by replacing the bidirectional GRU with a unidirectional GRU of size 256, and the relation module by replacing the unidirectional GRU with a FC layer of size 256 for deep metric learning. In this case, the embedding GRU output at the last time step summarizes the entire video into a single vector. We call this model "TARN-single". Table 2 shows the obtained accuracies

Method	Measure	1-shot	2-shot	3-shot	4-shot	5-shot
TARN	Mult	63.10	71.16	74.08	76.36	75.64
TARN	Subt	64.82	70.70	73.90	76.26	77.54
TARN	NN	63.26	70.46	72.70	75.62	75.58
TARN	SubMultNN	66.10	73.74	75.44	77.08	78.20
TARN	EucCos	**66.55**	**74.56**	**77.33**	**78.89**	**80.66**

Table 1: TARN model accuracy when using different similarity/distance measures in the comparison layer.

Method	Features	1-shot	2-shot	3-shot	4-shot	5-shot
CMN [49] - ECCV 2018	ResNet-50	60.5	70.0	75.6	77.3	78.9
TARN-f-single		57.92	63.96	65.90	68.52	70.64
TARN-f		64.83	72.94	76.22	78.02	78.52
TARN-single	C3D	62.84	69.80	73.96	74.88	76.88
TARN		**66.55**	**74.56**	**77.33**	**78.89**	**80.66**

Table 2: Accuracies of the the state-of-the-art method [49], as well as the TARN model at different settings.

by both the TARN and TARN-single models. Aligning video segments through attention and comparing segment-wise embeddings leads to consistent gains over the different shots. Furthermore, TARN model achieves better accuracy than the state-of-the-art method [49]. The accuracy gains hold for all shots, with significant boosts in the more difficult one-shot case, showing that even with a single sample per class in the sample set, the TARN model can still perform well.

In the last experiment, we compare the TARN model to the architecture proposed in [49] when using the same visual features. In the comparison, we use features extracted from a ResNet-50 [12] model pretrained on ImageNet [33]. ResNet features are extracted every 3 frames. We modify the TARN model to use the frame-level ResNet-50 features instead of the C3D ones, by replacing the bidirectional GRU in the embedding module by a unidirectional GRU, that is applied over video segments and gives an output at the last time step of each segment. The rest of the network remains the same to that of the original TARN model. This model is called "TARN-f". In this comparison, we also show the effect of not using attention and segment-level comparisons. To do so, the unidirectional GRU is applied over the entire video, and then the GRU output at the last time step is fed to the comparison layer. This model is called "TARN-f-single". Table 2 shows the results obtained by both the TARN-f and TARN-f-single models. First, we can see that TARN-f performs better than CMN [49] when using the same visual features (ResNet-50). Second, using attention and segment-wise comparisons improves the performance of our architecture. Finally, the C3D features perform better than the ResNet features, when trying to compare video segments in a few-shot scenario. This is probably due to the fact that C3D features are spatiotemporal, while ResNet features are static.

4.2 Zero-shot action recognition

Datasets and settings: We use two action recognition datasets to evaluate our architecture, namely UCF-101 [37] and HMDB51 [20]. UCF-101 has 13320 video clips from 101 classes, while HMDB51 has 6766 clips from 51 classes. Following [42], we divide the 101 actions in UCF-101 into 51/50 and 81/20 (seen/unseen) classes. For HMDB51, we divide the 51 actions into 26/25 classes. In the literature, different numbers of splits (ranging between 5-50) are randomly generated for the seen/unseen classes, and the mean accuracy and standard deviation are reported over them. In this work, we follow [42] and use 30 random splits used by [46] for the 51/50 and 81/20 cases in UCF-101, and for the 26/25 case in HMDB51.

Video/class representations: In our experiments, we use two types of semantic representations, both of which are widely used in the literature. First, we use the 115 binary semantic attributes (denoted as "Attr") that are manually annotated by [14] for UCF-101. To the best of our knowledge, there are no semantic attributes available for HMDB51. Second, we use 300-dimensional Word Vectors (mentioned as "WV") generated by the skip-gram

model of [23], that is trained on the Google News dataset. While there are other ways of representing semantic information, an extensive analysis of their influence is beyond the scope of this work. For example, [29] showed better performance using Error-Correcting Output Codes (ECOC). Here, we follow the majority of the works and use attributes and/or word vectors. Similar to FSL, we use C3D for extracting visual features.

Implementation details: The C3D features are embedded using a bidirectional GRU layer with a hidden state of size 256. Hence, the output dimension of the GRU at each time step (i.e. video segment) is of size 512. The semantic information is embedded using two FC layers with 4096 and 512 nodes. The deep network used for metric learning has two FC layers of size 256 and 1. We train our architecture in an end-to-end fashion using episode-based training strategy [36, 40]. Adam optimizer [17] with an initial learning rate set to 10^{-4} and gradient clipping to 0.5 is used in the training. The architecture is trained for $3,000$ episodes and tested for 100 episodes. During training, episodes/batches of size 16 for UCF-101 and 8 for HMDB51 are used, while in testing episodes are formed from all the unseen classes in the target split. We evaluate our architecture every 50 training episodes.

Ablation studies: In Table 3 we show the results obtained by our architecture on ZSL using different settings. The first and second settings (first two rows in Table 3) show the performance of our architecture when having a single representation for an entire video, instead of multiple segment representations. In the first setting, we use a unidirectional GRU to summarize the video segments into a single vector, obtained from the last time step of the GRU, and perform a single comparison between the semantic and the video vectors. In the second setting, we use a bidirectional GRU that gives an output at each time step. The attention mechanism is applied to the visual and semantic embedding, so as to map the multiple segment representations of the query video into a single representation. The single aligned query representation is then compared to the semantic embedding. The third setting (third row in Table 3) is a network that performs per-segment comparisons between the visual and semantic features. The attention mechanism maps the single semantic embedding of each class into multiple representations, as many, as the number of segments in the query video. In this way, each segment of the query video is compared to a semantic representation, allowing us to find relations between class-related semantic attributes and the visual features of each segment. Regarding the single comparison cases, the results show that the attention mechanism (second row in Table 3) effectively encodes multiple segment features into a single representation. The best performance is achieved when performing multiple segment-to-attribute comparisons.

Comparison to state of the art in ZSL: Methods proposed in the literature for ZSL have used a wide range of testing settings. In order to have a common setting across the majority of works, we do not compare with works that: (1) use auxiliary data to augment the training set [46, 47, 50]; (2) fuse different semantic or visual features [19, 42]; or (3) require access to the testing (unseen) classes during training (also known as "transductive" setting) [19, 42, 46, 47]. This allows us to have a clear and model-based comparison to

Method	UCF-101 (51/50)	UCF-101 (81/20)	HMDB51 (26/25)
TARN (w/o attention, single comparison)	16.7±4.0	35.8±5.9	16.6±3.4
TARN (with attention, single comparison)	20.0±2.5	38.1±5.6	17.4±2.8
TARN (with attention, multi comparison)	**23.2±2.9**	**42.7±5.4**	**19.5±4.2**

Table 3: Accuracies of the TARN model at different settings on zero-shot action recognition on the UCF-101 and HMDB51 datasets.

Method	Visual Repr.	Semantic Repr.	UCF-101 (51/50)	UCF-101 (81/20)	HMDB51 (26/25)
ESZSL† [32] - ICML 2015	IDT	WV	15.0±1.3	-	18.5±2.0
SJE† [1] - CVPR 2015	IDT	Attr	12.0 ± 1.2	-	-
		WV	9.9 ± 1.4	-	13.3 ± 2.4
UDICA [9] - CVPR 2016	C3D	Attr	-	29.6±1.2	-
KDICA [9] - CVPR 2016			-	31.1±0.8	-
MTE [47] - ECCV 2016	IDT	Attr	18.3±1.7	-	-
		WV	15.8±1.3	-	19.7±1.6
ZSECOC [29] - CVPR 2017	IDT	Attr	3.2±0.7	-	-
		WV	13.7±0.5	-	16.5±3.9
		ECOC	15.1±1.7	-	22.6±1.2
BiDiLEL [42] - IJCV 2017	C3D	Attr	20.5±0.5	39.2±1.0	-
		WV	18.9±0.4	38.3±1.2	18.6±0.7
GMM [24] - WACV 2018	C3D	Attr	22.7±1.2	-	-
		WV	17.3±1.1	-	19.3±2.1
UAR [50] - CVPR 2018	IDT	WV	17.5±1.6	-	**24.4±1.6**
TARN	C3D	Attr	**23.2±2.9**	**42.7±5.4**	-
		WV	19.0±2.3	36.0±5.3	19.5±4.2

Table 4: Accuracies of the TARN model as well as other state-of-the-art methods on zero-shot action recognition on the UCF-101 and HMDB51 datasets. Results marked with "†" are reported as reproduced by [47].

the literature. Table 4 summarizes the comparison results over the UCF-101 (51/50), UCF-101 (81/20), and HMDB51 (26/25) splits. We state in Table 4 the type of semantic and visual representation used in each method. Our architecture achieves the best results over the UCF-101 51/50 and 81/20 splits with almost 0.5% and 3%, respectively, in comparison to the second best performing methods. For HMDB51, we only get lower results than works that use either Improved Dense Trajectories (IDT) as visual features [29, 50], or ECOC as semantic representation [29].

5 Conclusion

In this work, we propose a deep network (called TARN) for addressing the problem of few-shot and zero-shot action recognition. TARN includes an embedding module for encoding the query and sample set examples, and a relation module that utilize attention for performing temporal alignment and a deep network for learning deep distance measure on the aligned representations at video segment level. The proposed network requires no additional resources or fine-tuning on the target problem. Our experimental results show that using attention and comparing segment-wise representations benefit the video-to-video or video-to-vector matching, compared to using video-wise representations. Furthermore, our method achieves the state-of-the-art results in FSL and very competitive performance in ZSL.

6 Acknowledgments

The work of Mina Bishay is supported by the Newton-Mosharafa PhD scholarship, which is jointly funded by the Egyptian Ministry of Higher Education and the British Council. This research has also been supported by EPSRC under grant No. EP/R026424/1. We gratefully acknowledge NVIDIA for the donation of the GTX Titan X GPU used for this research.

References

[1] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2927–2936, 2015.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[3] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734. Association for Computational Linguistics, October 2014.

[4] Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):677–691, April 2017.

[5] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594–611, 2006.

[6] Basura Fernando, Sareh Shirazi, and Stephen Gould. Unsupervised human action detection by action matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–9, 2017.

[7] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 1126–1135, 2017.

[8] Chuang Gan, Ming Lin, Yi Yang, Yueting Zhuang, and Alexander G. Hauptmann. Exploring semantic inter-class relationships (SIR) for zero-shot action recognition. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA.*, pages 3769–3775, 2015.

[9] Chuang Gan, Tianbao Yang, and Boqing Gong. Learning attributes equals multi-source domain generalization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 87–97, 2016.

[10] Chuang Gan, Yi Yang, Linchao Zhu, Deli Zhao, and Yueting Zhuang. Recognizing an action using its name: A knowledge-based approach. *International Journal of Computer Vision*, 120(1):61–77, 2016.

[11] Hua He and Jimmy Lin. Pairwise word interaction modeling with deep neural networks for semantic similarity measurement. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 937–948, 2016.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, pages 630–645, 2016.

[13] Samitha Herath, Mehrtash Harandi, and Fatih Porikli. Going deeper into action recognition: A survey. *Image and vision computing*, 60:4–21, 2017.

[14] Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes, 2014.

[15] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of International Computer Vision and Pattern Recognition*, 2014.

[16] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[18] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, 2015.

[19] Elyor Kodirov, Tao Xiang, Zhenyong Fu, and Shaogang Gong. Unsupervised domain adaptation for zero-shot learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2452–2460, 2015.

[20] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso A. Poggio, and Thomas Serre. HMDB: A large video database for human motion recognition. In *IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011*, pages 2556–2563, 2011.

[21] Hugo Larochelle, Dumitru Erhan, and Yoshua Bengio. Zero-data learning of new tasks. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, AAAI'08, pages 646–651. AAAI Press, 2008. ISBN 978-1-57735-368-3.

[22] Jingen Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pages 3337–3344. IEEE Computer Society, 2011. ISBN 978-1-4577-0394-2.

[23] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[24] Ashish Mishra, Vinay Kumar Verma, M Shiva Krishna Reddy, S Arulkumar, 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. IEEE, 2018.

[25] Tsendsuren Munkhdalai and Hong Yu. Meta networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2554–2563, 2017.

[26] Petar Palasek and Ioannis Patras. Discriminative convolutional fisher vector network for action recognition. *arXiv preprint arXiv:1707.06119*, 2017.

[27] Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. Zero-shot learning with semantic output codes. In *Advances in Neural Information Processing Systems 22*, pages 1410–1418. Curran Associates, Inc., 2009.

[28] Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255, November 2016.

[29] Jie Qin, Li Liu, Ling Shao, Fumin Shen, Bingbing Ni, Jiaxin Chen, and Yunhong Wang. Zero-shot action recognition with error-correcting output codes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2833–2842, 2017.

[30] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

[31] Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomas Kocisky, and Phil Blunsom. Reasoning about entailment with neural attention. In *International Conference on Learning Representations (ICLR)*, 2016.

[32] Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *International Conference on Machine Learning*, pages 2152–2161, 2015.

[33] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[34] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy P. Lillicrap. Meta-learning with memory-augmented neural networks. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 1842–1850, 2016.

[35] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, pages 568–576, 2014.

[36] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, 2017.

[37] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[38] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1199–1208, 2018.

[39] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4489–4497, 2015.

[40] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 3637–3645, 2016.

[41] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision*, pages 3551–3558, 2013.

[42] Qian Wang and Ke Chen. Zero-shot visual recognition via bidirectional latent embedding. *Int. J. Comput. Vision*, 124(3):356–383, September 2017. ISSN 0920-5691.

[43] Shuohang Wang and Jing Jiang. Machine comprehension using match-lstm and answer pointer. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

[44] Shuohang Wang and Jing Jiang. A compare-aggregate model for matching text sequences. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

[45] Baohan Xu, Hao Ye, Yingbin Zheng, Heng Wang, Tianyu Luwang, and Yu-Gang Jiang. Dense dilated network for few shot action recognition. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pages 379–387. ACM, 2018.

[46] Xun Xu, Timothy Hospedales, and Shaogang Gong. Semantic embedding space for zero-shot action recognition. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 63–67. IEEE, 2015.

[47] Xun Xu, Timothy M Hospedales, and Shaogang Gong. Multi-task zero-shot action recognition with prioritised data augmentation. In *European Conference on Computer Vision*, pages 343–359. Springer, 2016.

[48] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2021–2030, 2017.

[49] Linchao Zhu and Yi Yang. Compound memory networks for few-shot video classification. In *Proceedings of the European Conference on Computer Vision*, pages 751–766, 2018.

[50] Yi Zhu, Yang Long, Yu Guan, Shawn Newsam, and Ling Shao. Towards universal representation for unseen action recognition. In *2018 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2018.