Using Deep Learning for Fine-Grained Action Segmentation

Joaquim Yuste Ramos

Treball final de grau
Universitat de Barcelona
Grau d'Enginyeria Informàtica

Dirigit per Albert Clapés i Sergio Escalera

5 de juliol de 2021

Contents

UNIVERSITAT DE BARCELONA

Introduction

Problem to address

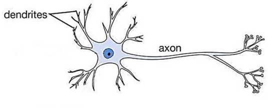
Determine the class label for each contained temporal subparts (actions) of a given video.

Objectives

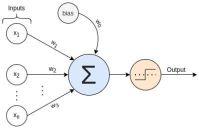
Modify a concrete deep learning (DL) based architecture in an attempt to improve the baseline scores.

Background: Artificial Neural Network

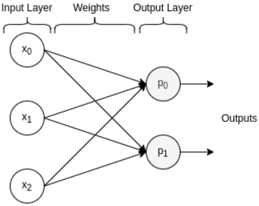
- An Artificial Neural Network (ANN) is a machine learning technique originally inspired by the behavior of biological neurons.



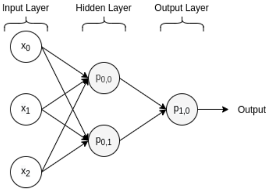
(a) Biological Neuron



(b) Artificial Neuron



(c) Single-Layer Perceptron



(d) Multi-Layer Perceptron

UNIVERSITAT DE BARCELONA

Training

Universal Approximation Theorem

A Multi-Layer Perceptron has the ability to approximate any arbitrary function.

Loss Function

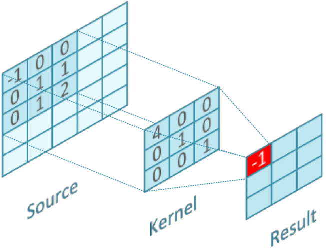
It gives a numerical score that states how good the network's prediction was.

Backpropagation

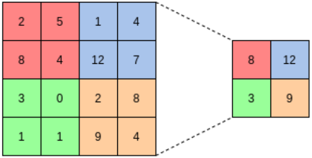
An algorithm that recursively uses the chain rule in order to compute the gradient with respect to every weight.

Convolutional Neural Network

- CNNs can process an input in a manner that the spatial structure of the data remains unchanged.



(a) Convolutional process



(b) MaxPooling process

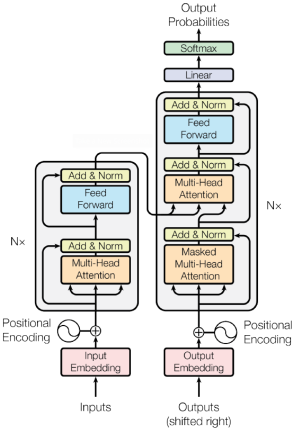
UNIVERSITAT DE BARCELONA

Self-Attention

Attention mechanisms have the capability to extract important relations between two inputs. Thus, they have been widely applied in Neural Machine Translation task.

Self-Attn. equation

$$softmax(W_q(In) \cdot W_k(In)^\top) \cdot W_v(In)$$

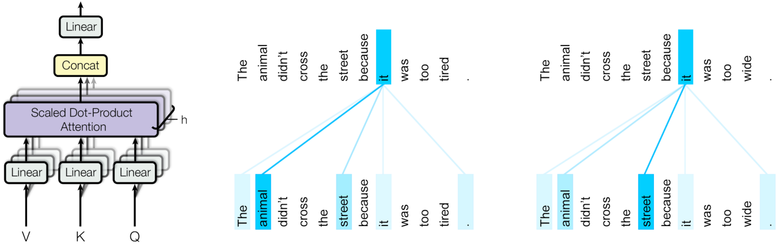


Transformer architecture[1]

[1] Ashish Vaswani et al. "Attention is all you need". In: *arXiv preprint arXiv:1706.03762* (2017).

Attention heads

- The attention mechanism can also be computed in parallel by splitting the input into h equal parts in the feature dimension.



(a) Multi-Head strategy (b) Relations generated by the self-attention mechanism

Related Work

Action Detection:

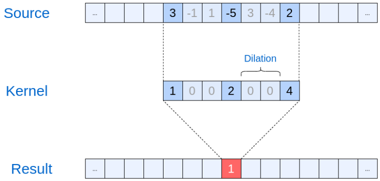
- CDC
- BSN
- Decoupled-SSAD

Action Segmentation:

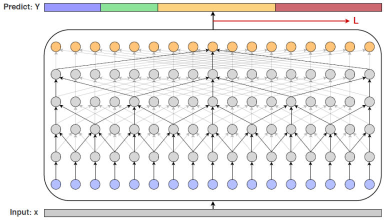
- ED-TCN & D-TCN
- MS-TCN
- ASRF

Others:

- SD-TCN
- Trans-SVNet
- PDAN

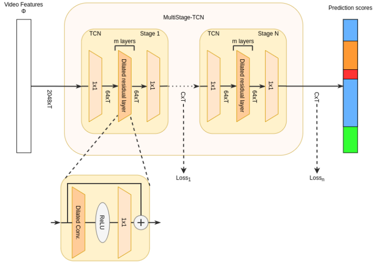


(c) Dilated Convolution

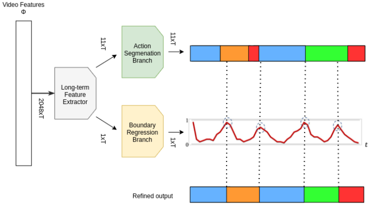


(d) TCN

- The modules proposed in this project are built upon a Multi-Stage Temporal Convolutional Network.



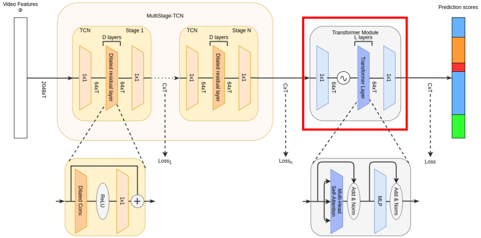
(a) Multi-Stage TCN

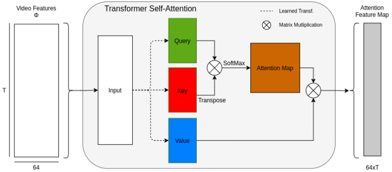


(b) ASRF

Methodology: TCN with Transformer

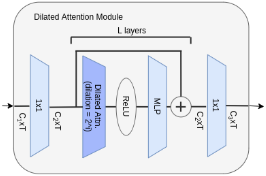
- A transformer has the ability to extract frame-to-frame dependencies.



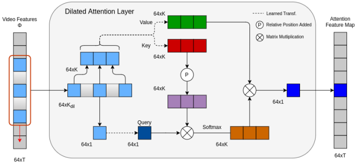


TCN with Conv. Self-Attention

- Instead of computing the attention in a global manner, a convolutional version is implemented.



(a) Convolutional Attention stage

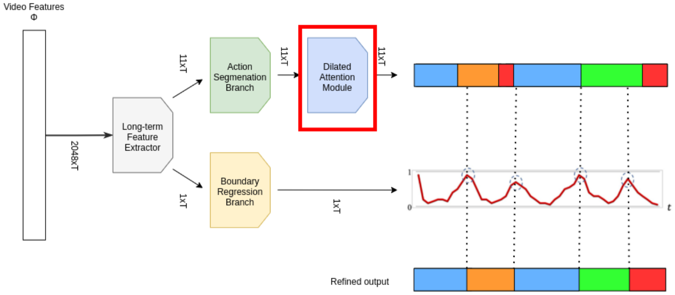


(b) Dilated Attention Layer

UNIVERSITAT DE BARCELONA

ASRF with Conv. Self-Attention

- Then, this last module is introduced in Action Segment Refinement Framework (ASRF)



Results: Dataset[2]

GTEA dataset properties

No. of videos	No. of classes	No. of users	Frame rate	View
28	11	4	15	Egocentric Dynamic

[2] Alireza Fathi, Ali Farhadi, and James M Rehg. "Understanding egocentric activities". In: *2011 international conference on computer vision*. IEEE. 2011, pp. 407–414.

UNIVERSITAT DE BARCELONA

Evaluation Metrics

Frame-wise accuracy

Percentage of frames that were correctly labeled.

Edit distance

Scored based on Levenshtein algorithm in order to emphasize the temporal order of the actions.

F1@IoU% score

Harmonic mean of precision and recall when detecting segments. Parametrized by the threshold on the percentage of IoU overlap required between groundtruth and predicted segments to consider the latter as true positive.

UNIVERSITAT DE BARCELONA

Implementation Details

Baseline parameters

Feature dimension	Number of Layers	Kernel size
64	12	3

Training configuration

Validation technique	Optimizer	Learning rate	Loss function
Four-fold cross-validation	ADAM	5e−4	Categorical Cross-Entropy loss Gaussian similarity-weighted loss (Binary Logistic Regression loss)

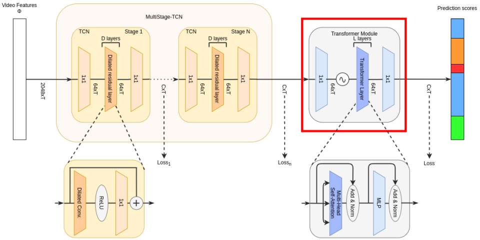
UNIVERSITAT DE BARCELONA

Transformer Module: Effect of the feature dimension

- First we evaluate the behaviour of the Transformer module when it is applied in top of a SingleStage-TCN

	F1@{10,25,50}			Edit	Acc
$D_{\text{Transf}} = 11$	46.13	43.69	**35.42**	37.17	**73.26**
$D_{\text{Transf}} = 64$	**48.40**	**43.79**	34.74	**38.26**	69.81

Table: Impact of the feature dimension (D_{Transf}) that the transformer works with.



UNIVERSITAT DE BARCELONA

Adding a Positional Encoding

- Injecting order information into the input data seems to be beneficial for the transformer module since it achieves better results.

	F1@{10,25,50}			Edit	Acc
w/o PE	48.40	43.79	34.74	38.26	69.81
with PE	**52.32**	**48.68**	**38.26**	**44.10**	**70.07**

UNIVERSITAT DE BARCELONA

Effect of the number of attention heads

- In this case, there is no significant difference between results when the number of heads are changed.

	F1@{10,25,50}			Edit	Acc
h=1	52.32	**48.48**	38.26	**44.10**	70.07
h=2	46.63	42.91	35.01	41.26	71.36
h=4	48.94	44.89	35.54	40.29	70.39
h=8	**52.72**	48.57	**39.15**	43.44	**72.16**
h=16	45.07	41.97	32.50	34.99	70.72

UNIVERSITAT DE BARCELONA

Effect of the number of layers

- The last experiment over the transformer module is made to evaluate the performance when its depth is increased.

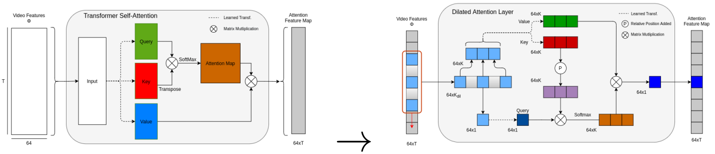
	F1@{10,25,50}			Edit	Acc
L=1	**52.32**	**48.48**	**38.26**	**44.10**	**70.07**
L=2	46.55	42.09	32.42	42.44	60.45

UNIVERSITAT DE BARCELONA

Conv. Self-Attention Module: Effect of the kernel size

- Now, the convolutional version of the attention is introduced.

	F1@{10,25,50}			Edit	Acc
K=7	79.04	74.82	59.69	71.97	74.09
K=9	79.52	75.27	62.66	**73.82**	74.36
K=11	**80.33**	**76.97**	**62.82**	72.76	**75.48**
K=13	77.41	72.86	61.28	70.47	72.96
K=15	80.20	75.13	61.41	72.98	75.44



(c) Transformer Layer

(d) Conv. Self-Attn.

Effect of the number of heads

- In this case, grouped convolutions take place in order to imitate the attention's heads mechanism.

	F1@{10,25,50}			Edit	Acc
$G = 1$	73.45	71.09	58.88	66.86	**75.63**
$G = 2$	75.10	71.38	57.76	67.79	74.76
$G = 4$	78.35	74.47	60.14	70.10	74.32
$G = 8$	**80.33**	**76.97**	**62.82**	**72.76**	75.48
$G = 16$	78.40	74.82	61.39	72.73	74.68

Table: Comparison choosing different number of groups G, whereas the number of layers and kernel sizes are $L = 1$ and $K = 11$ respectively

UNIVERSITAT DE BARCELONA

Effect of the number of layers

- There are two possible methodologies to work with when the depth is increased.

Version	Layers	F1@{10,25,50}			Edit	Acc
Standard	$L = 1$	80.33	76.97	62.82	72.76	**75.48**
Attn. Convolution	$L = 2$	80.51	76.99	63.69	74.53	73.27
	$L = 3$	82.09	79.93	66.36	76.81	74.40
Dilated	$L = 1$	-	-	-	-	-
Attn. Convolution	$L = 2$	81.56	78.64	64.40	75.89	74.89
	$L = 3$	**82.99**	**80.75**	**67.14**	**77.71**	73.95

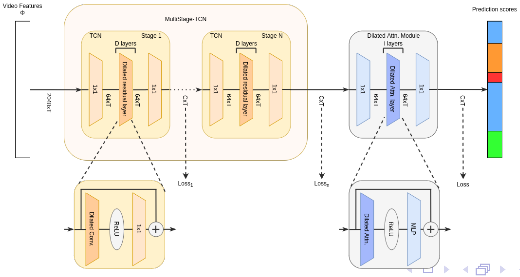
Table: Comparison between each tested strategies

UNIVERSITAT DE BARCELONA

Effect of the number of stages

- Until now, every experiment was carried out over a single stage TCN.

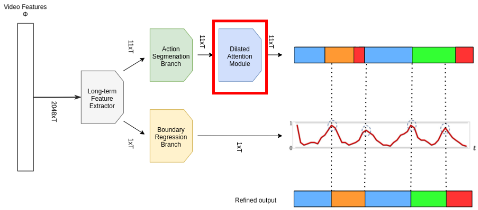
	F1@{10,25,50}			Edit	Acc
$S = 1$	82.99	80.75	67.14	77.71	73.95
$S = 2$	**86.04**	83.66	70.27	81.72	76.34
$S = 3$	85.16	81.29	64.15	80.18	74.17
$S = 4$	86.04	**84.10**	**70.57**	**82.41**	**76.58**



Adding Conv. Self-Attention to ASRF: Influence of the module on each branch

- For the last evaluation, the best attention module configuration is used on top of an ASRF architecture.

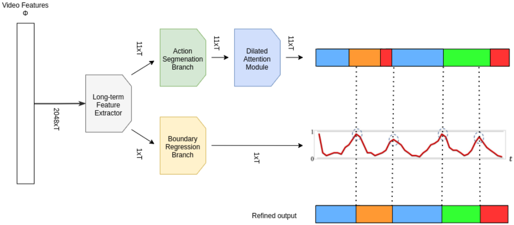
Branch	F1@{10,25,50}			Edit	Acc
Classification	**82.95**	**81.31**	**71.23**	76.42	**74.50**
Boundary	81.10	76.36	63.00	75.32	72.37
Both	82.56	79.73	67.11	**76.85**	73.21



Effect of the number of stages

- Finally, as made with MultiStage-TCN architecture, an addition of stages was performed.

	F1@{10,25,50}			Edit	Acc
$S = 1$	82.95	81.31	71.23	76.42	74.50
$S = 2$	83.42	81.98	69.15	**78.61**	75.16
$S = 3$	**86.09**	**85.06**	**74.58**	78.45	75.91
$S = 4$	85.48	84.07	71.43	78.28	**76.57**



Comparisson with the State-of-the-Art

Model	Stages	F1@{10,25,50}			Edit	Acc
MS-TCN	$S = 1$	60.97	57.20	47.09	52.32	74.32
	$S = 2$	80.53	77.53	64.08	76.51	74.31
	$S = 3$	85.30	82.45	70.36	80.38	75.70
	$S = 4$	**86.63**	84.15	69.44	82.00	74.91
ASRF	$S = 1$	84.36	81.26	67.86	79.09	74.34
	$S = 2$	82.67	80.10	68.78	75.55	73.09
	$S = 3$	84.60	83.27	72.22	76.35	75.04
	$S = 4$	85.40	83.95	72.96	77.84	74.62
MS-TCN+Transf	$S = 1$	52.32	48.48	38.26	44.10	70.07
MS-TCN+DAM	$S = 1$	82.99	80.75	67.14	77.71	73.95
	$S = 2$	86.04	83.66	70.27	81.72	76.34
	$S = 3$	85.16	81.29	64.15	80.18	74.17
	$S = 4$	86.04	84.10	70.57	**82.41**	**76.58**
ASRF+DAM	$S = 1$	82.95	81.31	71.23	76.42	74.50
	$S = 2$	83.42	81.98	69.15	78.61	75.16
	$S = 3$	86.09	**85.06**	**74.58**	78.45	75.91
	$S = 4$	85.48	84.07	71.43	78.28	76.57

Table: Comparison with the state-of-the-art on GTEA dataset where DAM refers to Dilated Attention Module

UNIVERSITAT DE BARCELONA

Conclusions

Two different attention modules have been analyzed in this project in order to tackle the problem of temporal action segmentation task .

- The transformer module is able to process the whole video in a non-local manner.
- The convolutional self-attention module slides its kernel through the temporal domain.
- The addition of a dilated convolutional self-attention module in a MultiStage-TCN displays significant improvements.

We aim to evaluate the proposed method in additional datasets (already working on that).

UNIVERSITAT DE BARCELONA