Temporal Attention-Gated Model for Robust Sequence Classification

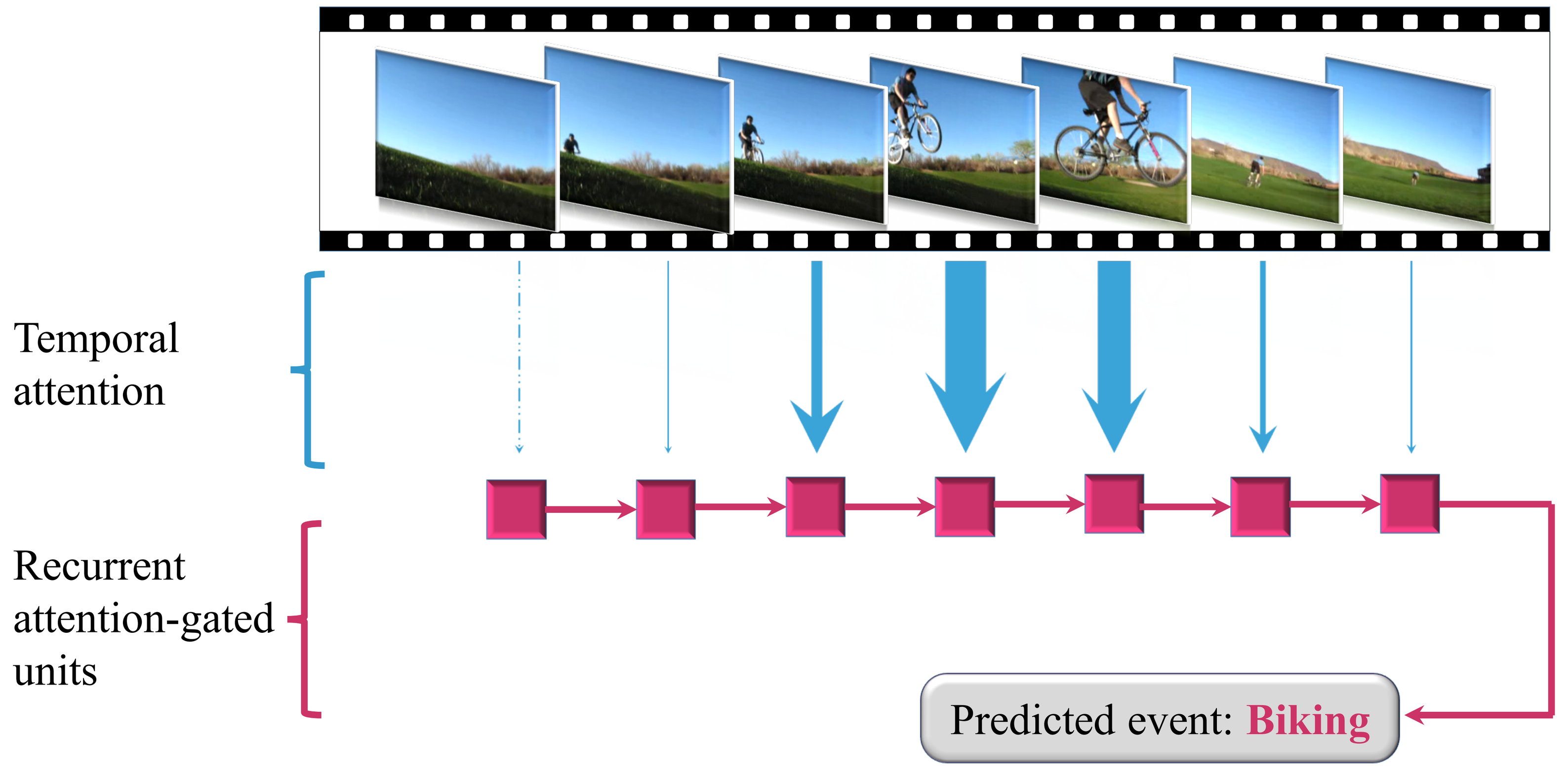
Wenjie Pei[1], Tadas Baltrušaitis[2], David M.J. Tax[1] and Louis-Philippe Morency[2]
[1]Pattern Recognition Laboratory, Delft University of Technology
[2]Language Technologies Institute, Carnegie Mellon University

IEEE 2017 Conference on Computer Vision and Pattern Recognition
CVPR July 21-26 HONOLULU 2017

Motivation

Typical sequence classification models are designed for **well-segmented** sequences, which often need manual and time-consuming pre-processing.

Goal: We propose Temporal Attention-Gated Model (TAGM) to better deal with **noisy** or **unsegmented** sequences by automatically locating the salient segments.



Our model can automatically extract the salient frames from the noisy raw input sequences and learns an effective hidden representation for the top classifier. The wider the arrow is, the more salient the frame is and the more the information is taken into account for prediction.

Contributions

- Automatically capture salient parts of the input noisy sequence to achieve better performance.
- Inferred attention scores provide meaningful interpretation for the informativeness of each time step.
- Less parameters leading to faster training and better generalizability with less training data.
- Generalization across different **tasks and modalities**.
- Code available[1].

Model

- **Recurrent Attention-Gated Units**

To learn an effective hidden representation.

$$\mathbf{h}_t = (1 - a_t) \cdot \mathbf{h}_{t-1} + a_t \cdot \mathbf{h}'_t$$

$$\mathbf{h}'_t = g(\mathbf{W} \cdot \mathbf{h}_{t-1} + \mathbf{U} \cdot \mathbf{x}_t + \mathbf{b})$$

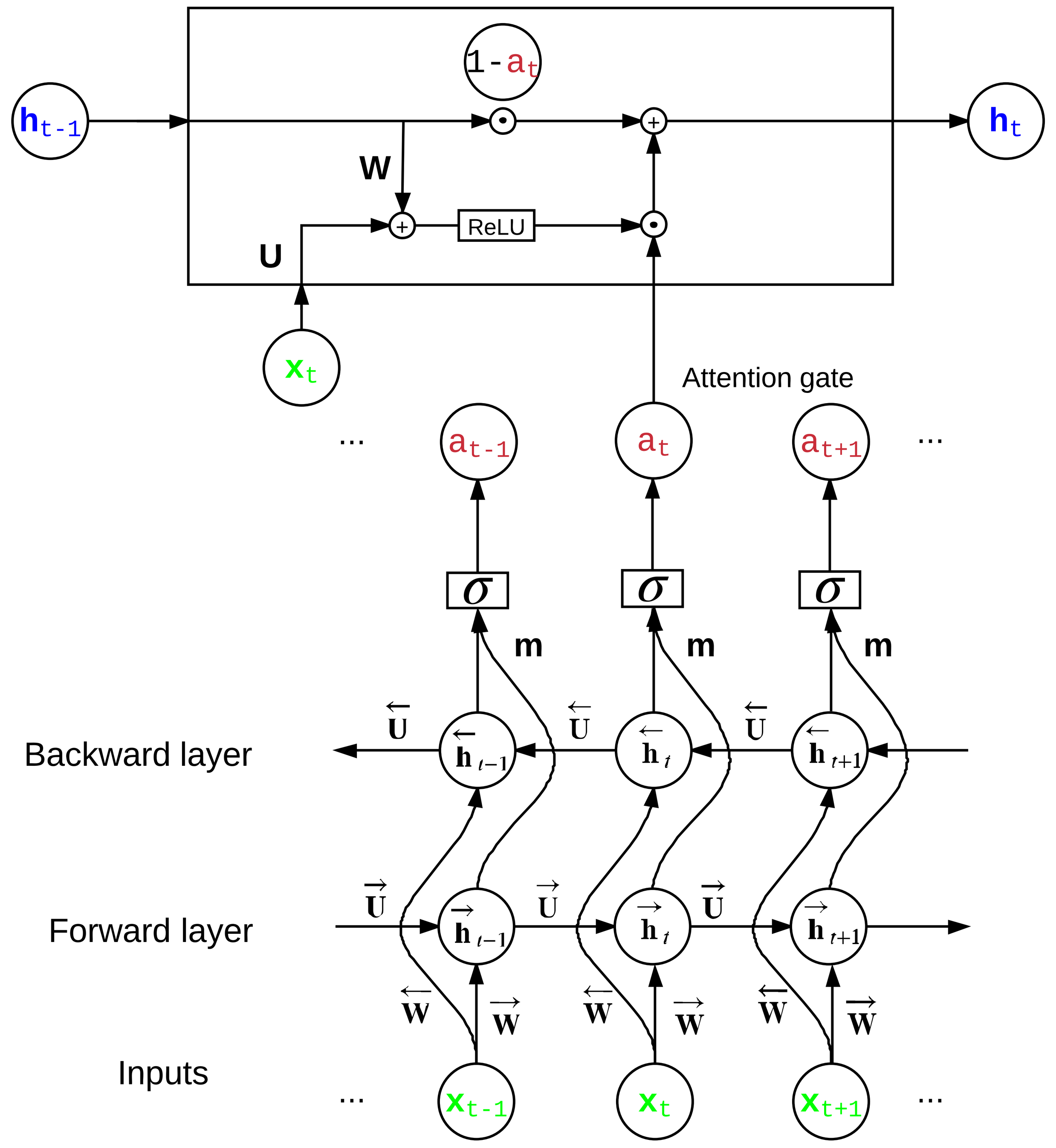
- **Temporal Attention Module**

To extract salient frames.

$$a_t = \sigma(\mathbf{m}^\top(\overrightarrow{h}_t; \overleftarrow{h}_t) + b)$$

$$\overrightarrow{h}_t = g(\overrightarrow{\mathbf{W}}\mathbf{x}_t + \overrightarrow{\mathbf{U}}\overrightarrow{h}_{\mathbf{t-1}} + \overrightarrow{\mathbf{b}})$$

$$\overleftarrow{h}_t = g(\overleftarrow{\mathbf{W}}\mathbf{x}_t + \overleftarrow{\mathbf{U}}\overleftarrow{h}_{\mathbf{t+1}} + \overleftarrow{\mathbf{b}})$$

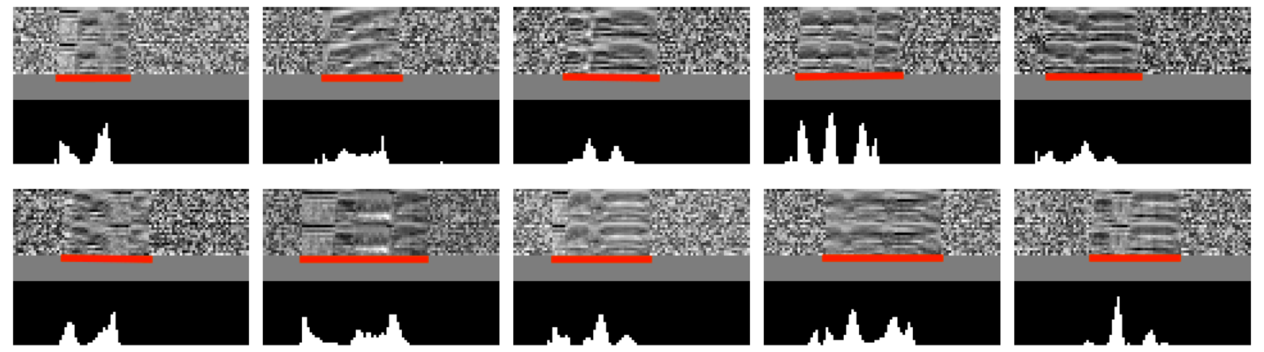


Experiments

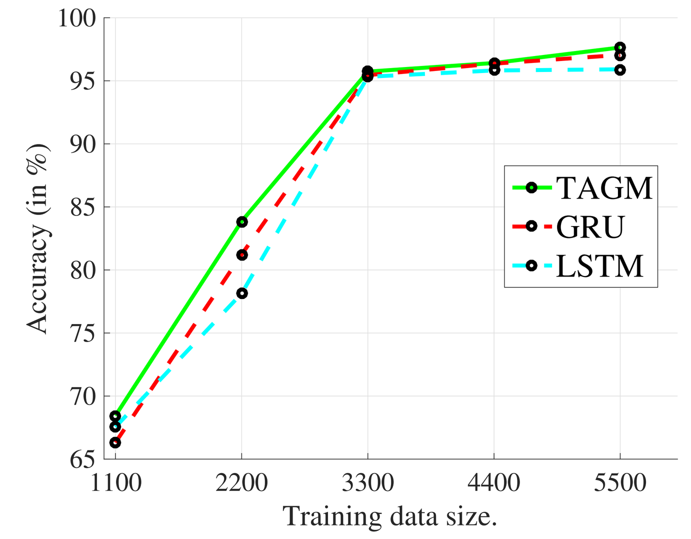
We perform experiments with TAGM on three datasets to show generalization across different tasks and modalities.

- **Speech Recognition**

Dataset: Noisy Arabic spoken digit dataset (8800 utterances,10 digits)
Feature: MFCCs



The visualization of attention weights of TAGM on 10 samples (one sample for each digit). For each subfigure, the top subplot shows the spectrogram of the original sequence data while the bottom subplot shows the attention values over time. The red lines indicate the groundtruth of salient segments.



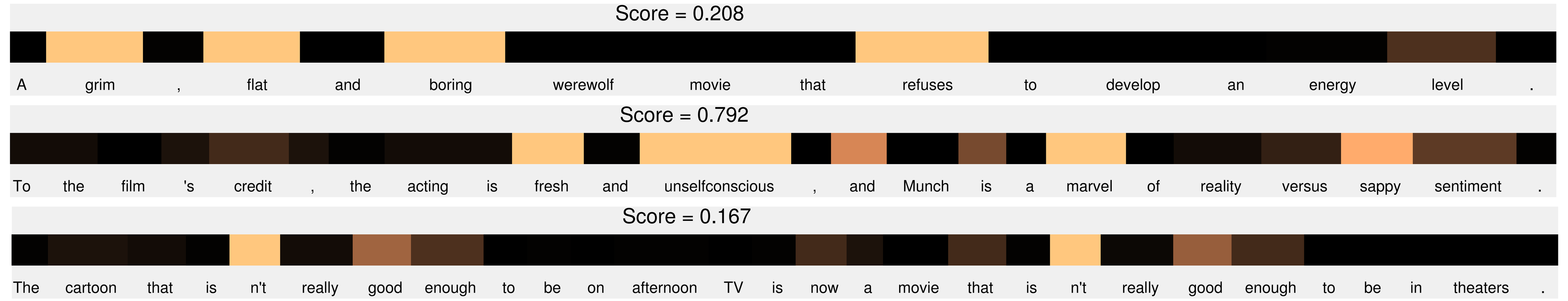
The classification accuracy on the noisy Arabic spoken digit dataset as a function of the size of training data.

Model	#Hidden units	#Parameters	Accuracy
HULM*	–	–	95.32
HCRF*	–	–	96.32
HULM	–	–	88.27
HCRF	–	–	90.41
Plain-RNN*	256	75 K	94.95
Plain-RNN	256	75 K	10.95
GRU	128	61 K	97.05
LSTM	128	81 K	95.91
NN	64	2.4 K	65.50
AM-NN	128-64	43 K	85.59
TAGM	128-64	47 K	**97.64**
Bi-GRU	64	37 K	97.68
Bi-LSTM	256	587 K	97.45
Bi-TAGM	128-128	83 K	**97.91**

Classification accuracy (%) on Arabic spoken digit dataset by different sequence classification models. Asterisked models (*) are trained and evaluated on the clean version of data.

- **Sentiment Analysis**

Dataset: Stanford Sentiment Treebank (11,855 review sentences, binary-classification or fine-grained task)
Feature: 300-d glove word vectors



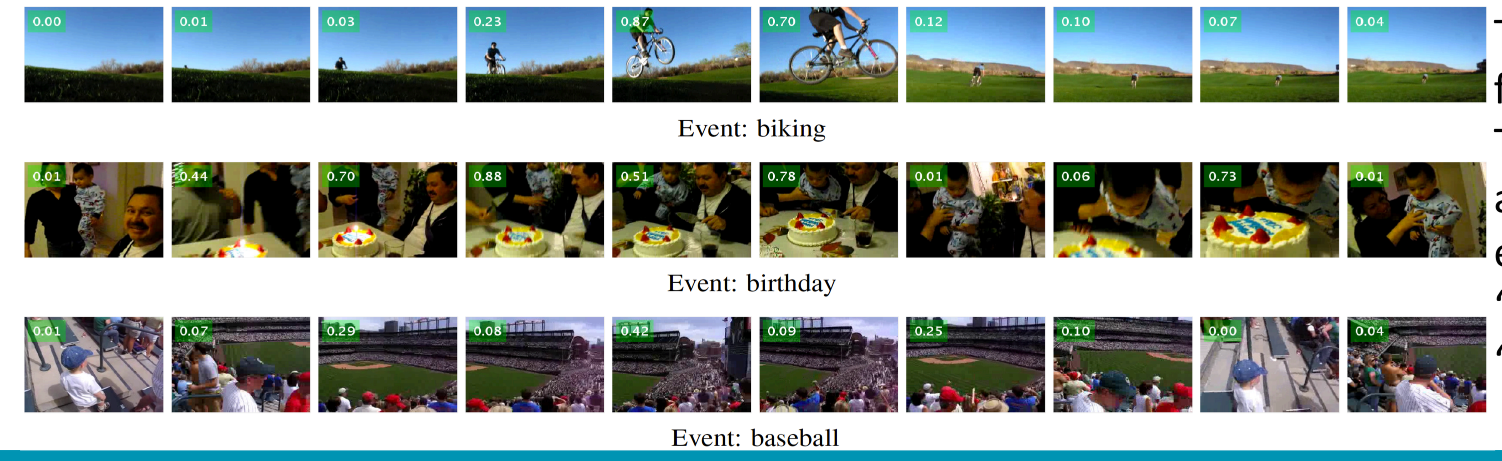
The visualization of attention weights of TAGM. The scores displayed are the groundtruth label indicating the sentiment for this review. Darker color indicates smaller scores.

	Model	Binary	Fine-grained	Overall Performance
Unordered compositions	NBOW-RAND	81.4	42.3	123.7
	NBOW	83.6	43.6	127.2
	BiNB	83.1	41.9	125.0
Syntactic compositions	RecNN	82.4	43.2	125.6
	RecNTN	85.4	45.7	131.1
	DRecNN	86.6	49.8	136.4
	DAN	86.3	47.7	134.0
	TreeLSTM	86.9	**50.6**	137.5
	CNN-MC	**88.1**	47.4	135.5
	PVEC	87.8	48.7	136.5
Our model	TAGM	87.6	50.1	**137.7**

Classification accuracy (%) on Stanford Sentiment Treebank dataset for both binary classification and 5-level fine-grained classification task.

- **Event Recognition**

Dataset: Columbia Consumer Video Database (9317 videos, 20 events)
Feature: CNN features from pre-trained AlexNet model



The attention weight is indicated for representative frames. Our TAGM is able to capture the action of ‘**riding bike**’ for the event ‘biking’, ‘**cake**’ for the event ‘birthday’ and ‘**infield zone**’ for ‘baseball’.

Model	Training strategy	Feature	mAP
BOW+SVM +late average fusion	Separately (one-vs-all)	SIFT	0.52
		STIP	0.45
		SIFT+STIP	0.55
		CNN	0.67
Plain-RNN	Jointly	CNN	0.45
GRU	Jointly	CNN	0.56
LSTM	Jointly	CNN	0.55
TAGM	Jointly	CNN	0.63

Mean Average Precision (mAP) of our TAGM model and baseline models on CCV dataset.

[1]https://github.com/wenjiepei/TAGM.