

# Contents

<b>List of Tables</b>	<b>viii</b>
<b>List of Figures</b>	<b>x</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation . . . . .	1
1.2 Objectives . . . . .	2
1.2.1 Main objective . . . . .	2
1.2.2 Specific objectives . . . . .	2
1.3 Methodology . . . . .	2
1.4 Structure of this Research . . . . .	3
<b>2 Theoretical background</b>	<b>4</b>
2.1 Sentence Vectorization . . . . .	4
2.1.1 Bag of Words . . . . .	4
2.1.2 Word2Vec . . . . .	5
2.1.3 Recurrent Neural Networks . . . . .	5
2.1.4 InferSent . . . . .	7
2.2 Video vectorization . . . . .	8
2.2.1 Optical embeddings . . . . .	8
2.2.2 Dense Trajectory . . . . .	8
2.2.3 Convolutional networks . . . . .	9
2.3 Video-to-Text Matching & Ranking . . . . .	10
<b>3 Baseline system overview</b>	<b>12</b>
3.1 Video multi-level side . . . . .	13
3.1.1 Level 1. Global encoding by Mean Pooling . . . . .	13
3.1.2 Level 2. Temporal-Aware Encoding by biGRU . . . . .	13
3.1.3 Level 3. Local-Enhanced Encoding by biGRU-CNN . . . . .	14
3.2 Text multi-level side . . . . .	14
3.3 Common vector space . . . . .	15
<b>4 Enhanced Video-To-Text model</b>	<b>17</b>
4.1 Video representation . . . . .	18
4.1.1 Dense Trajectories embedding . . . . .	19
4.1.2 ResNeXt-101 embedding . . . . .	21

4.2	Sentence representation . . . . .	22
4.3	Distance functions for the common vector space . . . . .	24
4.3.1	Optimization . . . . .	24
4.4	Parametric complexity . . . . .	25
<b>5</b>	<b>Data and pre-processing</b>	<b>26</b>
5.1	Training dataset . . . . .	26
5.1.1	MSVD . . . . .	26
5.1.2	MSR-VTT . . . . .	27
5.1.3	TGIF . . . . .	28
5.1.4	Framerate analysis . . . . .	29
5.2	Validation and testing datasets . . . . .	30
5.2.1	TRECVID 2016 and 2018 VTT official evaluation dataset . . . . .	30
5.3	Testing dataset representation . . . . .	32
5.4	Data pre-processing . . . . .	34
5.4.1	Video and GIF feature extraction . . . . .	35
5.4.2	Sentence vectorization . . . . .	36
5.5	Dense Trajectories with CUDA . . . . .	38
5.6	Hardware . . . . .	38
<b>6</b>	<b>Experimental evaluation</b>	<b>39</b>
6.1	Evaluation metric: Mean Inverted Rank . . . . .	39
6.2	Base model validation . . . . .	40
6.3	Training methodology . . . . .	41
6.3.1	Training validation metrics . . . . .	41
6.4	Training results . . . . .	42
6.4.1	Baseline Li <i>et al.</i> 2018 ResNeXt-101 comparison . . . . .	42
6.4.2	Baseline Li <i>et al.</i> 2018 distance function comparison . . . . .	43
6.4.3	DT-ALL variation . . . . .	44
6.4.4	DT-ONLY variation . . . . .	47
6.4.5	INFERSENT-ALL variation . . . . .	50
6.4.6	INFERSENT-ONLY variation . . . . .	51
6.5	Testing results . . . . .	53
6.5.1	Video-sentence matching examples . . . . .	56
6.6	Discussion . . . . .	61
6.6.1	Training and testing results . . . . .	61
6.6.2	Models performance in real use-case scenarios . . . . .	62
6.6.3	Training and testing dataset selection . . . . .	63
6.6.4	ResNeXt-101: Shufflenet vs MXnet . . . . .	65
6.6.5	Training methodology . . . . .	65
<b>7</b>	<b>TRECVID 2019 participation</b>	<b>67</b>
7.1	System Detail . . . . .	67
7.2	Datasets . . . . .	69
7.3	Submissions . . . . .	69
7.4	Result examples . . . . .	70

<b>8</b>	<b>Conclusions and future work</b>	<b>73</b>
8.1	Conclusions . . . . .	73
8.2	Source Code . . . . .	74
8.3	Baseline validation obstacles . . . . .	75
8.4	Future work . . . . .	75
	<b>Bibliography</b>	<b>77</b>
	<b>Appendix A Dong <i>et al.</i> 2017 model validation</b>	<b>85</b>