Contents

Preface **1**

Introduction **3**

1 Deep Learning **5**

1.1 Fundamentals of Deep Learning 5

1.1.1 What is Machine Learning? 5

1.1.2 Perceptron Algorithm 6

1.1.3 Multi-Layer Networks 7

1.1.4 Error Back-Propagation 9

1.1.5 Representation Learning 10

1.2 Deep Learning Techniques in Computer Vision 10

1.2.1 Convolutional Networks 10

1.2.2 AlexNet and Image Classification on the ImageNet Challenge . . . 12

1.2.3 Convolutional Networks after AlexNet 14

1.3 Deep Learning Techniques in Natural Language Processing 16

1.3.1 Word Embeddings 17

1.3.2 Architectures for Sequence Processing 20

1.3.3 Generating Output 27

1.4 Conclusion 34

2 Notable Models **35**

2.1 Word2Vec and the Others 35

2.2 Attention and Machine Translation with Recurrent Neural Networks 36

2.3 Transformer for Machine Translation 37

2.4 CoVe: Contextual Embeddings are Born 38

2.5 ELMo: Sesame Street Begins 39

2.6 BERT: Pre-trained Transformers 40
2.7 GPT and GPT-2 43
2.8 Conclusion 44

3 Interpretation of Neural Networks 45
3.1 Supervised Methods: Probing 47
3.2 Unsupervised Methods: Clustering and Component Analysis 47
3.3 Network Layers and Linguistic Units 48
3.3.1 Words versus States 49
3.3.2 Words versus Subwords 51
3.4 Conclusion 52

4 Emblems of the Embeddings 55
4.1 Word Analogies 55
4.1.1 Word2Vec and Semantic Arithmetic 55
4.1.2 Glove Word Analogies 56
4.1.3 FastText Subword Correspondence 59
4.2 Positioning Words 59
4.3 Embedding Bands 63
4.4 Visualising Word Embeddings with T-SNE 63
4.5 Emoji Embeddings 63
4.6 Principal Component Analysis 66
4.6.1 Visualisation 66
4.6.2 Correlations with Principal Components 66
4.6.3 Histograms of Principal Components 69
4.6.4 Sentiment Analysis 69
4.7 Independent Component Analysis 72
4.8 Word Derivations 73
4.9 Mapping Embedding Spaces 75
4.10 Debiasing: Interpretation as Manipulation 76
4.11 Conclusion 77

5 May I Have Your Attention? 79
5.1 Cross-Lingual Attentions and Word Alignment 80

5.2 Self-Attentions and Syntactic Relations 84
5.2.1 Categorization of Self-Attention Heads 85
5.2.2 Highly Redundant Attention Heads 91
5.2.3 Syntactic Features of Self-Attention Heads 93
5.2.4 Dependency Trees 95
5.2.5 Constituency Trees 96
5.2.6 Syntactic Information across Layers 98
5.2.7 Other Relations between Words 99
5.3 Interpretability of Attentions Not as Easy as Expected 100
5.3.1 Eliminate the Highest Attention Weight 100
5.3.2 Change the Whole Attention Distribution 102
5.3.3 Do Not Attend to Useful Tokens 106
5.4 Conclusion 107

6 Contextual Embeddings as Un-hidden States 109

6.1 How Contextual Embeddings Came to Be 110
6.2 What do Hidden States Hide? 111
6.2.1 Morphology 112
6.2.2 Syntax 113
6.2.3 Coreference 116
6.2.4 Semantics 116
6.2.5 Context 118
6.2.6 Word Senses 119
6.2.7 World Knowledge and Common Sense 121
6.3 What is Hidden Where? 121
6.3.1 Comparison of Architectures and Models 121
6.3.2 Distribution of Linguistic Features across Layers 123
6.3.3 Effect of Pre-training Task 126
6.4 Multilinguality 130
6.5 Conclusion 132

Afterword 135

Summary 137

CONTENTS

List of Figures 139

List of Tables 141

List of Abbreviations 144

Bibliography 145

Index 167