Language Development with Deep Learning

From Learning Theory to Computational Experiments

Abdellah Fourtassi

Contents

List of illustrations *page* viii
Notation x
Preface xv

1 The Problem and the Program 1
1.1 Broad Neural Learning and the Physical Learner 2
1.2 Generalization and Competing Solutions 5
1.3 Constraint Matching as an Experimental Program 8
1.4 Positioning and Organization of the Book 11

2 Language Learning as Induction 15
2.1 Three Problems, One Logic 15
2.1.1 Grammatical Generalization 15
2.1.2 Word Meaning 16
2.1.3 Speech Categories 17
2.1.4 The Shared Inductive Problem 18
2.2 Overgeneralization and Inductive Constraints 18
2.2.1 The Problem of Overgeneralization 18
2.2.2 Constraining Generalization 20
2.3 Probability and Inductive Preference 22
2.3.1 Probability Changes the Evidence 22
2.3.2 Evidence Still Needs a Bias 24
2.3.3 Bayesian Models of Language Learning 25
2.4 Contribution to the Research Program 28

3 From Identification to Prediction 30
3.1 From Construction-Specific Problems to Broad Learnability 30
3.1.1 Formal Languages and Generative Capacity 31
3.1.2 The Components of a Learning Problem 32
3.2 Gold: Identification in the Limit 33
3.2.1 Positive Texts and Eventual Identification 33
3.2.2 Superfinite Classes 34
3.2.3 Why Every Learner Can Be Made to Fail 34
3.2.4 What the Negative Result Implies 35

3.3 The Presentation Is Part of the Learning Problem 35
3.3.1 Restricted Target Classes 36
3.3.2 Robustness Over Presentations 36
3.3.3 From a Language to a Probabilistic Source 36
3.4 Solomonoff: Universal Predictive Induction 37
3.4.1 Computable Generative Hypotheses 38
3.4.2 A Universal Simplicity Prior 38
3.4.3 Posterior Weights and Prediction 38
3.5 From Identification to Prediction 39
3.5.1 The Predictive Criterion 39
3.5.2 Predictive Agreement and Internal Structure 40
3.5.3 The Limits of the Ideal Learner 40
3.6 Contribution to the Research Program 40

4 The Developmental Predictive Learner 42
4.1 Defining the Physical Learner 42
4.1.1 A Physical Space of Parameters 42
4.1.2 Learning from Experience 43
4.1.3 Sequential and Concurrent Context 44
4.1.4 From Physical States to Predictive Hypotheses 45
4.1.5 Prediction as a Functional Criterion 46
4.2 Evaluating Predictive Learning 48
4.2.1 Learning from Finite Experience 48
4.2.2 Source-Level Predictive Error 49
4.2.3 Evaluating Generalization 50
4.3 Linguistic Knowledge and Predictive Learning 51
4.3.1 Three Descriptions of One Learner 51
4.3.2 The Linguistic System and its Development 52
4.3.3 How Prediction Favors Linguistic Organization 53
4.3.4 Predictive Learning in Models and Humans 55
4.4 Contribution to the Research Program 56

5 Representational Architecture 57
5.1 Representational Capacity and Depth 57
5.1.1 From Simple Neural Computations to Rich Functions 58
5.1.2 Universal Approximation 62
5.1.3 Network Depth 63
5.2 Examining Internal Representations 64
5.2.1 Decoding Linguistic Distinctions 65
5.2.2 Encoding Neural Responses 67
5.3 Contextual Processing and Linguistic Analysis 69
5.3.1 Self-Attention 69
5.3.2 Self-Attention and Human Sentence Processing 71
5.3.3 A Single Analysis Licensed by Persistent Knowledge 72

5.4 Contribution to the Research Program 73

6 Learning Dynamics 74
6.1 From Processing to Learning 74
6.2 Updates and the Trajectories They Produce 76
6.2.1 The Learning Rule 76
6.2.2 Gradient Descent 76
6.2.3 Reducing Predictive Error Across Updates 77
6.2.4 Developmental Trajectories 79
6.3 The Learner's Initial State 80
6.3.1 Constructing an Initial State 81
6.3.2 Modeling Prenatal Experience 81
6.3.3 Learning Biases Along the Trajectory 82
6.4 Learning Cycles and Plasticity 83
6.4.1 Incremental Learning and Stochastic Updates 83
6.4.2 Plasticity and Hypothesis Search 84
6.5 Developmental Evidence from Learning Trajectories 84
6.6 Contribution to the Research Program 85

7 Credit Assignment 86
7.1 The credit-assignment problem 86
7.1.1 Locating the Parameters That Need Credit 87
7.2 Backpropagation in a Simple Network 89
7.2.1 Isolating the Calculation 89
7.2.2 From a Weight to a Local Error Signal 90
7.2.3 Passing Credit Backward 90
7.3 From artificial weights to biological synapses 91
7.3.1 Two Local Multiplications 91
7.3.2 Making Backpropagated Credit Locally Available 92
7.4 Predictive coding as internal credit assignment 92
7.4.1 Predictions, Local Errors, and Resettling 93
7.4.2 Deriving the Local Activity Dynamics 94
7.4.3 Representing Errors and Providing a Feedback Pathway 95
7.4.4 From Settled Errors to a Synaptic Update 96
7.5 The Weight-Transport Problem and Feedback Alignment 98
7.5.1 Credit Without Exact Equality 99
7.6 What Credit-Assignment Alternatives Establish 99
7.6.1 Approximating the Backpropagation Benchmark 99
7.6.2 Resettling and Sequential Language Processing 100
7.7 Contribution to the Research Program 102

8 The Effective Learning History 103
8.1 From Learning Events to Developmental Histories 103
8.2 Availability: Which Experience Reaches The Child? 105

8.2.1 How Much Input Is Available? 105
8.2.2 The Distribution of Available Evidence 105
8.2.3 Curriculum and the Order of Experience 106
8.2.4 The Stream Can Be Learner-Contingent 106
8.3 Access: From Available to Effective Experience 107
8.3.1 Processing with Limited Contextual Memory 107
8.3.2 Learning with Limited Context: Is Less More? 108
8.3.3 Developmental Changes in Access 109
8.4 Persistence: Learning After the Event 110
8.4.1 Fast Memory and Slow Learning 111
8.4.2 Episodic Traces and Reconstruction 112
8.4.3 Replay and Consolidation 112
8.4.4 Implications for Learning and Generalization 113
8.5 Contribution to the Research Program 114

9 Learning from Ecological Experience 115
9.1 Processing an Ecological Encounter 115
9.1.1 From Symbolic to Perceptual Experience 115
9.1.2 Unifying the Learnability Challenges 116
9.1.3 Processing by Reciprocal Disambiguation 117
9.1.4 Reciprocal Disambiguation Across Development 119
9.1.5 The Synergistic Learning Hypothesis 120
9.2 Learning From Perceptual Encounters 120
9.2.1 Learning Synergies and Bootstrapping 120
9.2.2 Connections Already Established Between Components 122
9.3 Specialized Components in an Integrated Learner 122
9.3.1 Separate Parameter Groups 123
9.3.2 Modality-Specific Computations 123
9.3.3 From Probabilistic Relations to Component Activity 124
9.4 A Developing Synergetic System 125
9.4.1 From Speech to Sequences and Situations 125
9.4.2 Learning Through the Three Components 127
9.4.3 Evaluating a Coupled Learner 128
9.5 Contribution to the Research Program 131

10 Generalization and Human-Like Solution Selection 132
10.1 From Finite Experience to Generalization 132
10.1.1 Three Requirements for Generalization 132
10.1.2 Isolating the Finite-Sample Estimation Problem 134
10.1.3 Capacity and the Estimation Gap 135
10.1.4 Generalization in Overparameterized Networks 136
10.2 Solution-Dependent Generalization 138
10.2.1 From Class Capacity to Description Length 138
10.2.2 Why a Large Network Can Have a Short Description 139

10.2.3 Sample Size and Generalization Regimes 140
10.2.4 Where Solution Preferences Come From 141
10.3 From Generalization to Human-Like Alignment 141
10.3.1 The Second Induction Problem 141
10.3.2 Memorization vs. Abstraction 142
10.3.3 Short-cut and Overgeneralization 143
10.3.4 Behavioral Success and Representational Change 143
10.4 Constraint Matching and Human-Like Solution Selection 144
10.4.1 Architecture 145
10.4.2 Initial State 146
10.4.3 Objective 146
10.4.4 Learning Dynamics 147
10.4.5 Credit Assignment 147
10.4.6 Availability 148
10.4.7 Access 148
10.4.8 Persistence 149
10.4.9 Ecological Learning 149
10.5 General Conclusions 150

Bibliography 153
Subject index 173

Illustrations

1.1	Prediction can make linguistic distinctions useful.	3
1.2	Scope and mechanistic specification of learning.	4
1.3	Generalizing and generalizing like a human.	7
1.4	Competing generalizations of subject–verb agreement.	8
1.5	Constraint matching and knowledge alignment.	9
2.1	Language learning as inference from sensory input to linguistic knowledge.	16
2.2	Overlapping /æ/ and /ɛ/ tokens in F1–F2 space.	17
2.3	Positive and negative evidence constrain different errors.	20
2.4	A broader predictor divides a finite probability budget.	23
2.5	Two candidate program descriptions of the dative alternation.	25
2.6	Accumulating evidence can outweigh a simplicity preference.	28
3.1	A failure presentation for an infinite target.	35
3.2	Gold and Solomonoff formulate broad learnability differently.	39
4.1	From a physical learner to predictive linguistic behavior.	44
4.2	Sequential and concurrent information constrain one prediction.	45
4.3	Predictive behavior across domains and across learning.	47
4.4	Learning connects three probability descriptions.	50
4.5	Predictions and the underlying linguistic organization.	52
4.6	A developing system licenses increasingly differentiated analyses.	53
5.1	Illustration of representational inadequacy.	58
5.2	A one-neuron physical state and its probabilistic readout.	59
5.3	ReLU transformation of a single neural unit.	60
5.4	Approximation of a one-dimensional target function by one-hidden-layer ReLU networks of increasing width.	61
5.5	Successive transformations of the shared visual and linguistic context.	65
5.6	Dependency structure.	67
5.7	Decoding a distinction and encoding a neural response.	68
5.8	Schematic model-to-brain alignment across processing hierarchies.	69
5.9	The need for retrieval and a self-attention solution.	70
5.10	Retrieval and transformation in a Transformer.	71
6.1	Processing supplies the error used by learning.	75
6.2	Convex and non-convex empirical cross-entropy.	79
6.3	The same endpoint can conceal developmental mismatch.	80
7.1	Local connections contribute to a network-level prediction.	88

7.2	The forward computation and its training procedure.	93
7.3	A missing credit pathway and a recurrent circuit that provides local prediction error signals.	96
7.4	Fixed and evolving activity during credit assignment.	98
8.1	Three aspects of the effective learning history.	104
8.2	Developing access changes the effective evidence.	110
8.3	Persistence must be supplied somewhere.	111
8.4	Replay supplies further learning experience.	113
9.1	Three learnability challenges constrain one analysis.	118
9.2	Visual disambiguation and developmental precision.	121
9.3	Modality-adapted computations in a coupled learner.	126
9.4	From separate learning challenges to one coupled learning problem.	129
9.5	Evaluating three learnability challenges in one learner.	130
10.1	Empirical and source-level success in parameter space.	134
10.2	The classical capacity–estimation trade-off.	136
10.3	The same capacity permits memorization and generalization.	137
10.4	Solution selection within the empirical–source intersection.	142
10.5	From developmental inductive constraints to aligned knowledge.	145