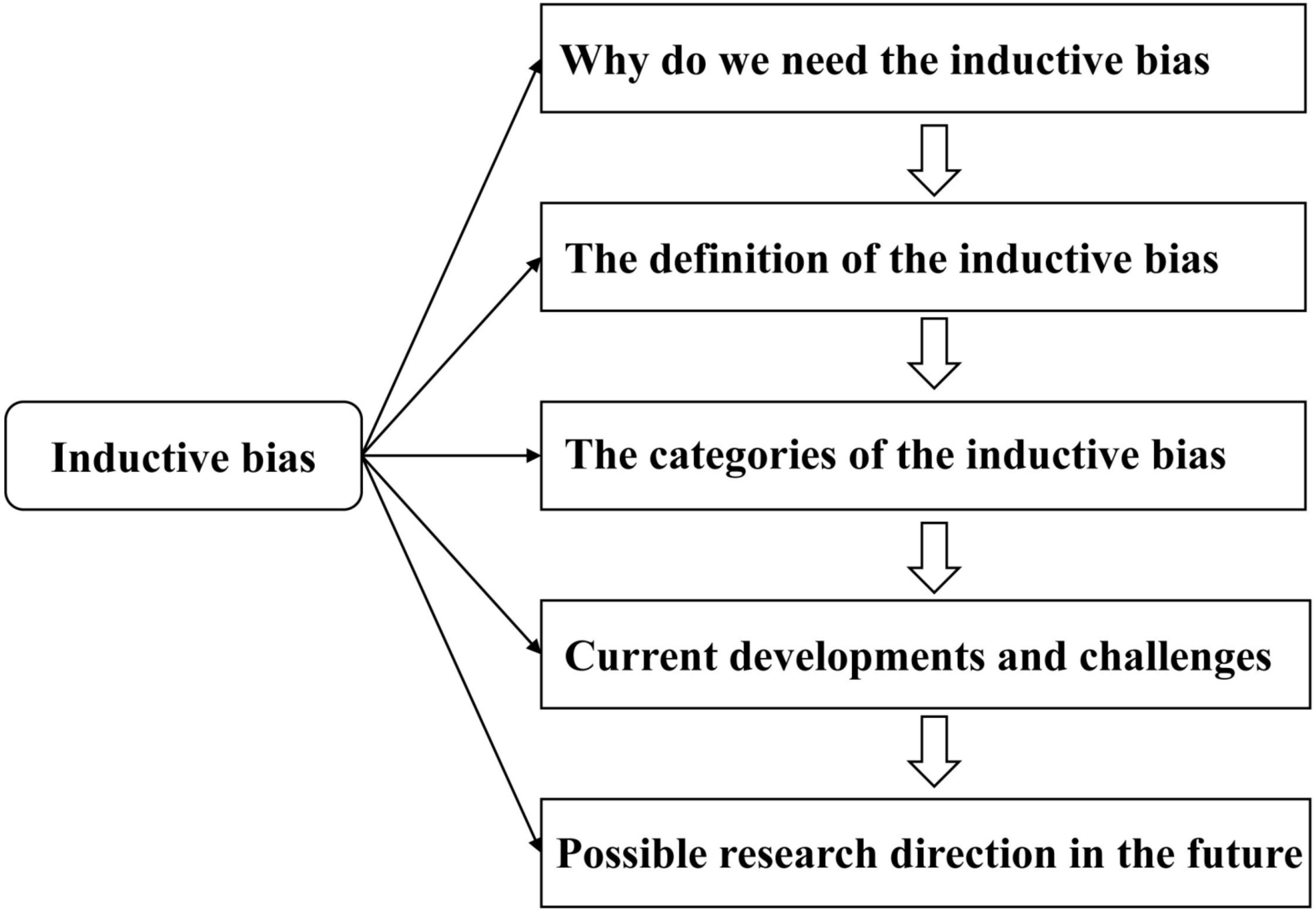
Inductive Bias

Weipeng Cao, PhD
Shenzhen University

2019.9.4

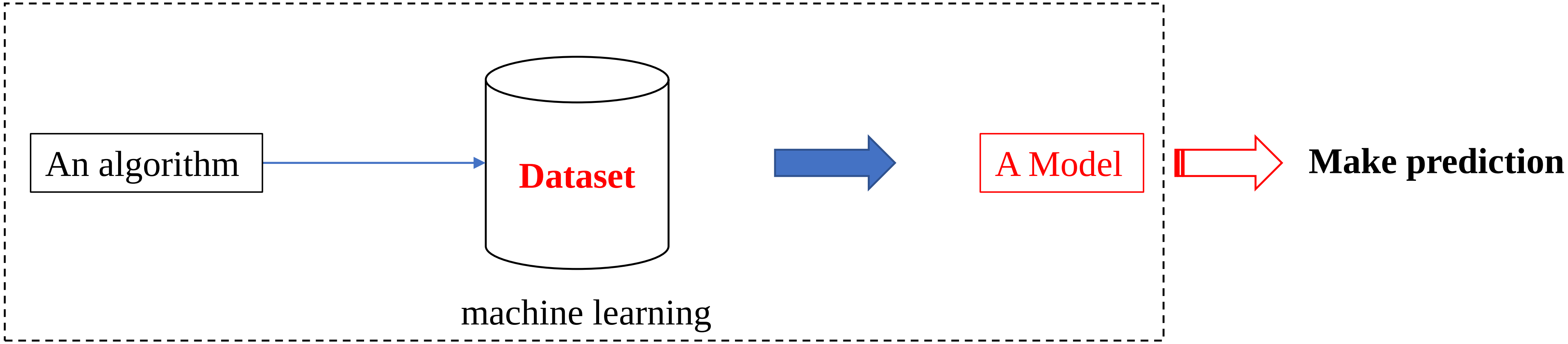
Content


Inductive bias
Why do we need the inductive bias
The definition of the inductive bias
The categories of the inductive bias
Current developments and challenges
Possible research direction in the future

Why does machine learning require inductive bias?

Why does machine learning require inductive bias?

The learning process of machine learning



In general, if the prediction accuracy of the model on the **unseen** samples is high, we call it a model with good **generalization ability**.

Why does machine learning require inductive bias?

How to get a model with good generalization ability?

unbiased strategy → rote learning ⇒ enumerators

biased strategy → introduce additional assumptions for model learning ⇒ inductive bias

The samples we can acquire in reality are often limited and accompanied by noise, errors, and inconsistencies. In this case, it is unreasonable or even harmful to keep the learner in consistency with the training samples.

Biased strategy is more feasible in practical scenarios and inductive bias is necessary for the effective learning of the machine learning algorithm

The definition and categories of inductive bias

The definition and categories of inductive bias

Mitchell T. M. gave the following description in 1980 [48]:

"bias refers to any basis for choosing one generalization over another, other than strict consistency with the observed training instances..."

Later, Utgoff P. E. gave a more specific definition [72][73]:

"bias is the set of all factors that collectively influence hypothesis selection. These factors include the definition of the space of hypotheses and definition of the algorithm that searches the space of concept descriptions..."

[48] Mitchell, T.M.: The need for biases in learning generalizations. Department of Computer Science, Laboratory for Computer Science Research (1980)
[72] Utgoff, P.E.: Shift of bias for inductive concept learning. Machine learning: An artificial intelligence approach 2, 107-148 (1986)
[73] Utgoff, P.E.: Machine learning of inductive bias, vol. 15. Springer Science & Business Media (2012)

The definition and categories of inductive bias

Given a training dataset, we need some additional constraints or criteria to help us better fit the training samples, so that the trained model can make better predictions on the unseen samples (i.e., generalize beyond the training data).

The additional constraints or criteria here are called inductive bias.

The definition and categories of inductive bias

Generally, the inductive bias of a learning algorithm A is a minimal set of assertions B such that for any concept c and corresponding training data D_c, it holds that for any $x_i \in X$,

$$B \wedge D_c \wedge x_i \vdash A(x_i, D_c), \tag{1}$$

where D_c is the set of training examples, and x_i is the i-th instance of the input set X. Equ. 1 expresses the intuition that with the minimal hypothesis B, the learning algorithm A can give a sufficient classification. However, this logical definition of inductive bias is difficult to straightforward use in real learning problems, so we will consider its formal definition formally in the following part.

First we consider single-task supervised learning models. We assume an input space X, an output space Y, which are both measurable spaces, a probability measure P on $X \times Y$, a loss function $f : Y \times Y \to \mathbb{R}$, and a hypothesis space $\mathcal{H}$, which is a set of hypotheses or functions $h : X \to Y$. Here we assume that all the functions involved here are measurable. Our goal is to select a hypothesis $h \in \mathcal{H}$ to minimize the expected loss:

$$er_P(h) = \int_{X \times Y} l(h(x), y)\, \mathrm{d}P(x, y). \tag{2}$$

The definition and categories of inductive bias

Usually the probability measure P is not known by the learner, so the solution is to generate a sample space $\{(x_i, y_i)\}_{i=1}^{m}$, and we minimize the estimation

$$\hat{er}_P(h) = \frac{1}{m} \sum_{i=1}^{m} l(h(x_i), y_i). \tag{3}$$

It has been proved by Vapnik [4], and Blumer et al. [11] that $\hat{er}_P(h)$ will converge uniformly to $er_P(h)$ as $m \to \infty$ with probability arbitrarily close to 1.

Now we define the bias learning problem. Different from the settings above, we need to learn the inductive bias for a set of problems, so we use a pair $(\mathcal{P}, Q)$ to represent the environment, where $\mathcal{P}$ is the set of all probability measures on $X \times Y$, and Q is a distribution on P, which gives the distribution of the problems that the learner will see. Also, we do not choose a single hypothesis for all these problems, but choose the best hypothesis space $\mathcal{H}$ in a given hypothesis family $\mathbb{H}$. Naturally, the goal of a bias learner is still to minimize the expected loss:

$$er_Q(\mathcal{H}) = \int_{\mathcal{P}} \inf_{h \in \mathcal{H}} er_P(h)\, \mathrm{d}Q(P) = \int_{\mathcal{P}} \inf_{h \in \mathcal{H}} \int_{X \times Y} l(h(x), y)\, \mathrm{d}P(x, y)\, \mathrm{d}Q(P). \tag{4}$$

[4] Baxter, J.: A model of inductive bias learning. Journal of artificial intelligence research 12, 149{198 (2000)

[11] Blumer, A., Ehrenfeucht, A., Haussler, D., Warmuth, M.K.: Learnability and the vapnik-chervonenkis dimension. Journal of the ACM (JACM) 36(4), 929{965 (1989)

The definition and categories of inductive bias

Similarly, we sample to obtain an estimation of this expected loss. Not only do we need to sample on $X \times Y$, we also need to sample from $\mathcal{P}$. We sample n times from $\mathcal{P}$ according to Q to get $\{P_1, \ldots, P_n\}$, and for each P_i, we sample m times to get $\{(x_{i1}, y_{i1}), \ldots, (x_{im}, y_{im})\}$. We define the matrix $\mathbf{z} = ((x_{ij}, y_{i,j}))_{n\times m}$, and z_i is the i-th row of $\mathbf{z}$. The estimation of the expected loss $er_Q(\mathcal{H})$ is

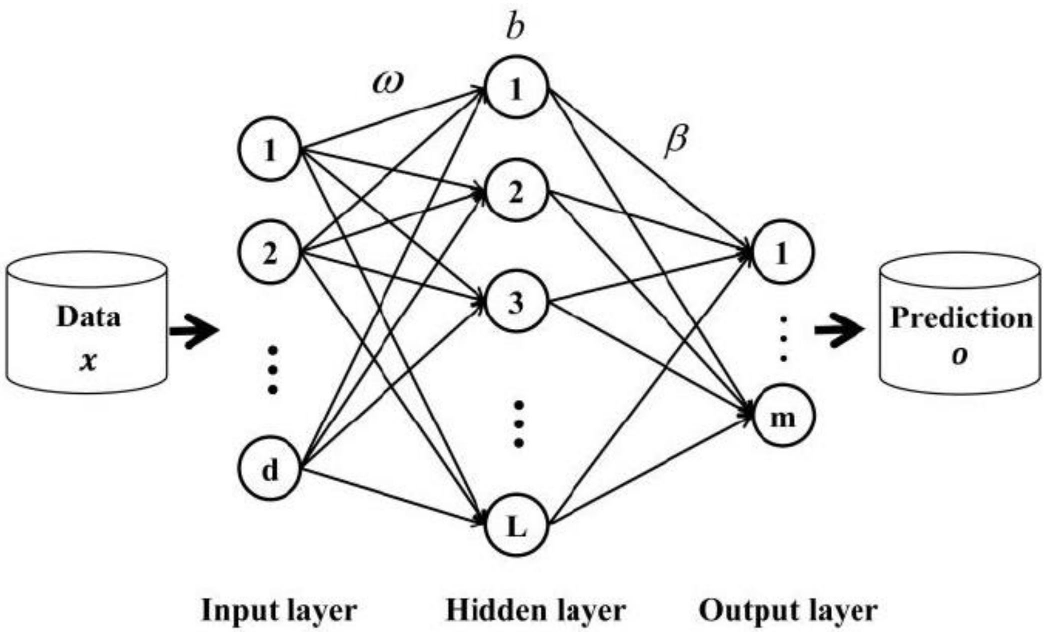
$$\hat{er}_{\mathbf{z}}(\mathcal{H}) = \frac{1}{n}\sum_{i=1}^{n} \inf_{h\in\mathcal{H}} \hat{er}_{z_i}(h). \tag{5}$$

It has been proved in [4] that if the sampling satisfies some good properties, then the estimation will converge uniformly to the loss expectation with probability arbitrarily close to 1. Also, this result can be expanded straightforward to multiple tasks.

[4] Baxter, J.: A model of inductive bias learning. Journal of artificial intelligence research 12, 149{198 (2000)

Differences between inductive bias and other biases

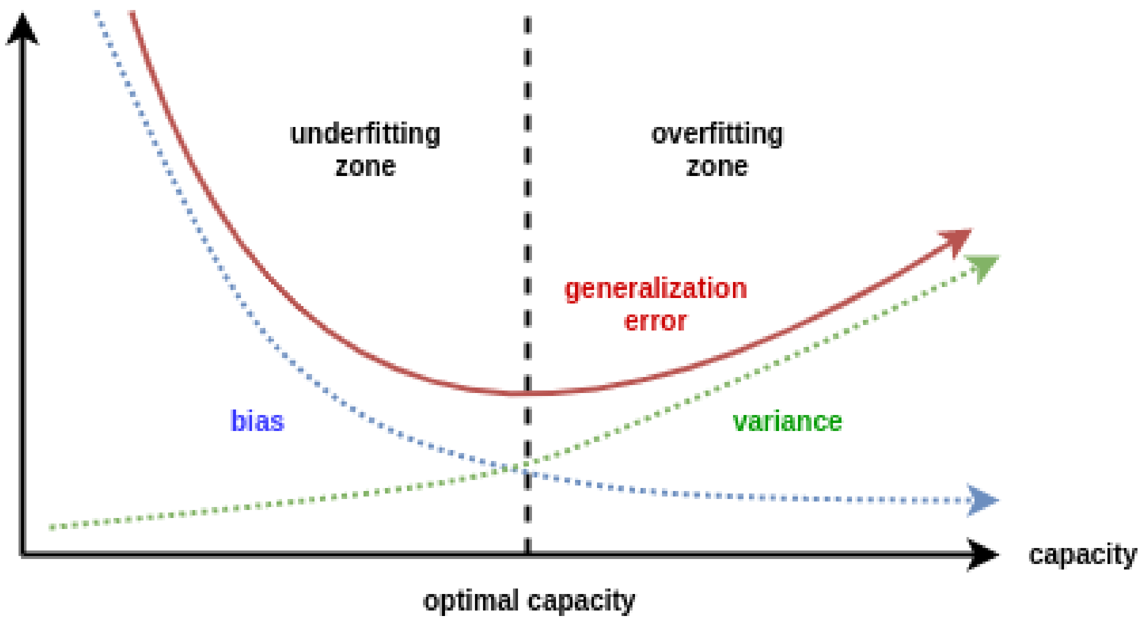
The bias in neural networks



$G(wx + b)$

The biases of neural networks are used to improve the translation ability of the model.

The bias in the Bias-Variance tradeoff



The bias in the Bias-Variance tradeoff is one of the important indicators for model evaluation.

The choice of inductive bias determines whether the machine learning algorithm can effectively learn on a problem.

The categories of inductive bias

Inductive biases

restriction bias

play a role of adding some restrictions to the search space and limiting the hypothesis space

various regularization methods, such as L1 regularization and L2 regularization. It is noted that the pooling strategy [3], convolution operation [7][26], dropout strategy [62][63], and Shakeout [64] in deep learning also belong to the restriction bias.

preference bias

set different priority for the hypothesis space

various strategies for selecting models, such as Occam's razor principle (preference for simplicity), minimum cross-validation error, maximum margin (Support Vector Machine, SVM), and nearest neighbors (K-nearest neighbors algorithm, KNN) [2].

[2] desJardins M., Gordon D.F. (1995). Evaluation and selection of biases in machine learning. Machine Learning Journal, 5:1-17, 1995.
[3] Cohen N., Shashua A. Inductive bias of deep convolutional networks through pooling geometry. ICLR, 2017.
[7] Fukushima K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological cybernetics, 36(4), 193-202.
[26] LeCun Y., Boser B., Denker J. S., Henderson D., Howard R. E., Hubbard W., Jackel L. D. (1989). Backpropagation applied to handwritten zip code recognition. Neural computation, 1(4), 541-551.
[62] Helmbold D. P., Long P. M. (2015). On the inductive bias of dropout. The Journal of Machine Learning Research, 16(1), 3403-3454.
[63] Helmbold D. P., Long P. M. (2017). Surprising properties of dropout in deep networks. The Journal of Machine Learning Research, 18(1), 7284-7311.
[64] Kang G., Li J., Tao D. (2017). Shakeout: A new approach to regularized deep neural network training. IEEE transactions on pattern analysis and machine intelligence, 40(5), 1245-1258.

The properties of inductive bias

inductive bias

Strength. The strength of inductive bias describes its limitation on the size of the hypothesis space that the learner can search. Strong inductive bias gives the learner a relatively small search space, while weak inductive bias provides a broader search space for the learner.

How to measure it? **VC dimension theory (Vapnik–Chervonenkis dimension)**

Correctness. Only the correct inductive bias can ensure that the learner successfully learns the target concept. Conversely, under the incorrect induction bias, the learner cannot learn the correct target concept no matter how many training samples are used.

How to measure it? **PAC learning theory (Probably Approximately Correct)**

Inductive bias in the field of neural networks

Inductive bias in the field of neural networks

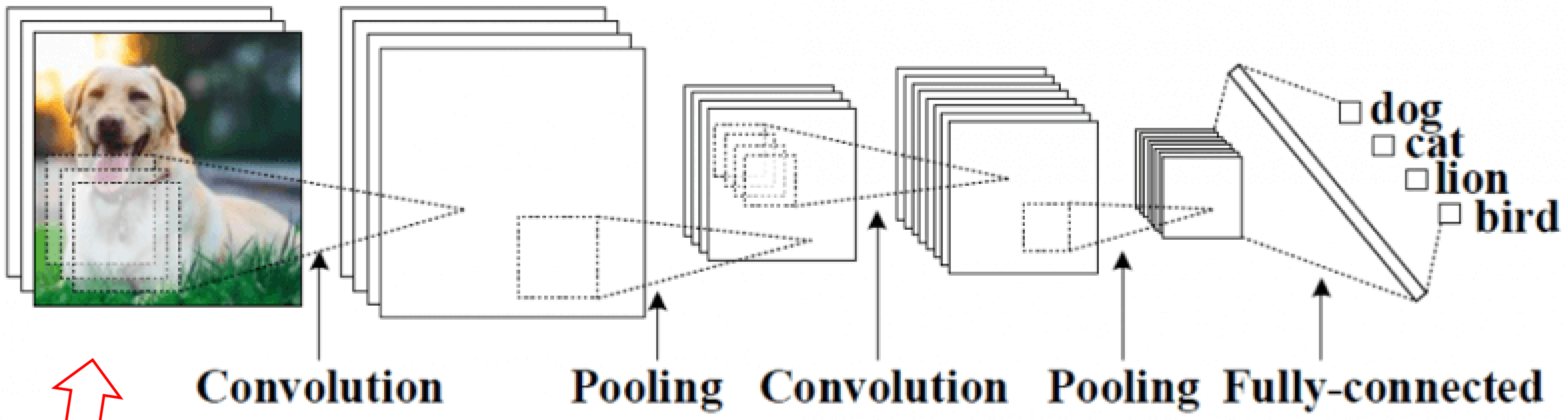


Image data has the characteristics of locality and translation invariance.

Different pooling schemes would induce different preferences of the model [3] .

Convolution operation and local weight sharing strategy can make the neural network algorithms extract and transform these local features effectively.

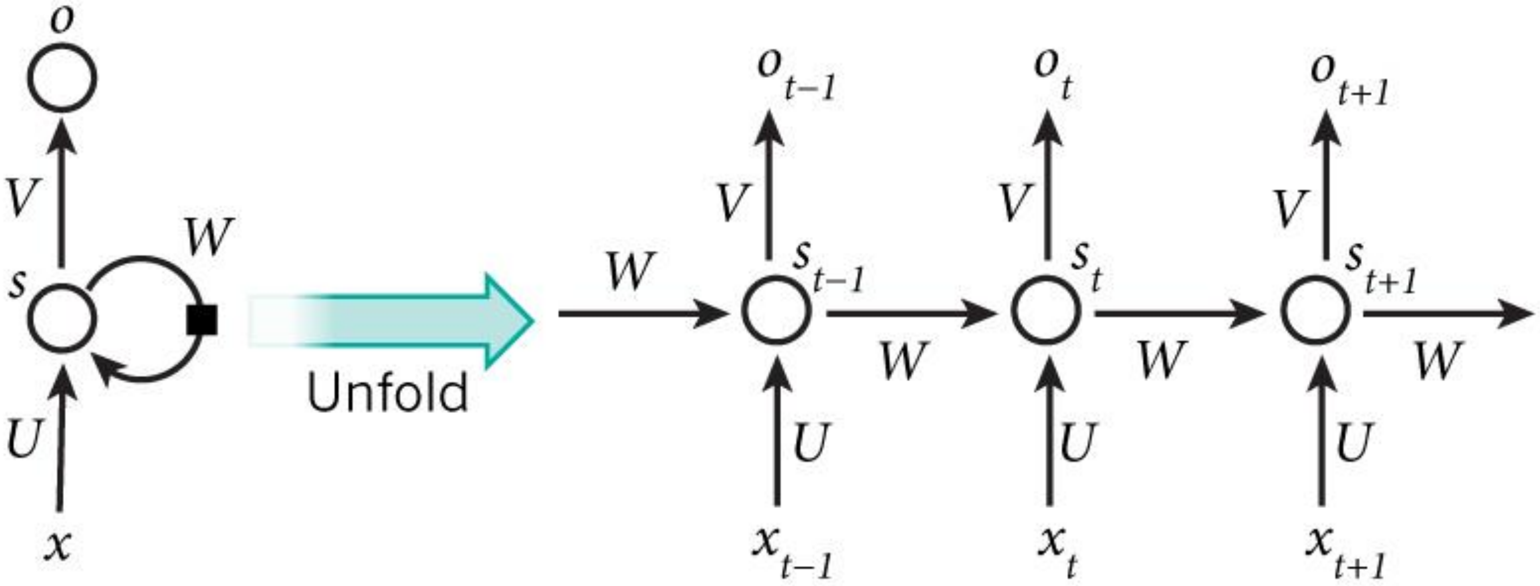
restriction bias

preference bias

[3] Cohen N., Shashua A. Inductive bias of deep convolutional networks through pooling geometry. ICLR, 2017.

Inductive bias in the field of neural networks

Recurrent Neural Networks



The inductive biases of RNNs include sequence and time invariance [13].

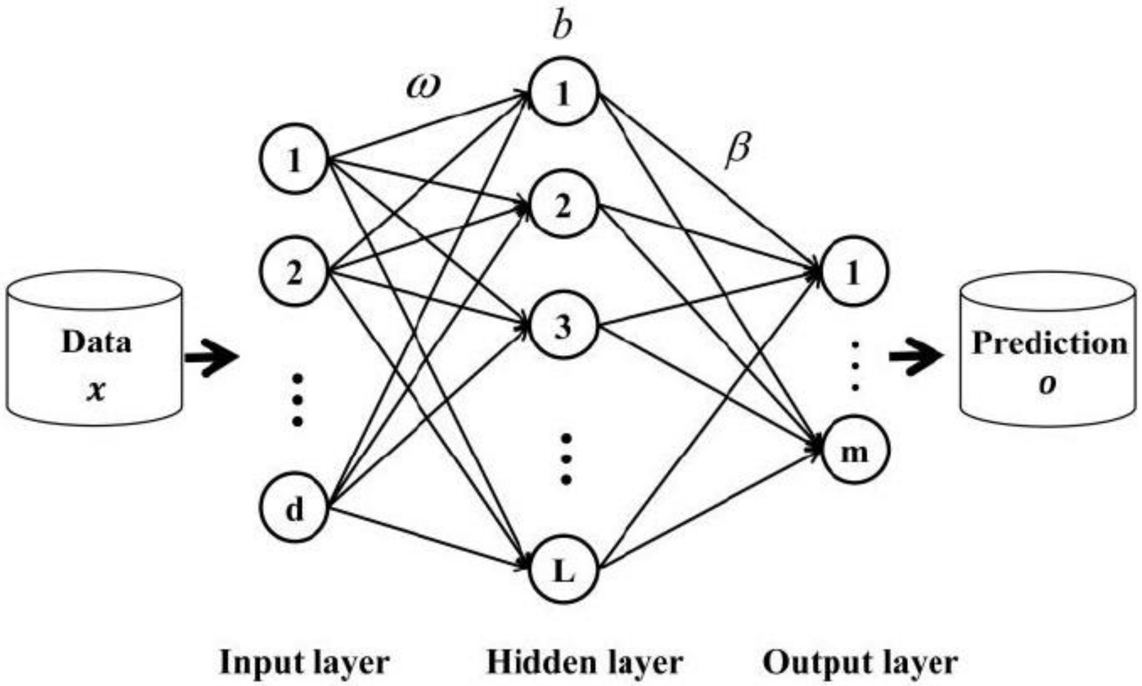
Each token in the sequence can be modeled as a result of previous tokens and tokens far away from it can be forgotten.

restriction bias

[13] Battaglia P. W., Hamrick J. B., Bapst V., Sanchez-Gonzalez A., Zambaldi V., Malinowski M., Tacchetti A., Raposo D., Santoro A., Faulkner R., Gulcehre C., Song F., Ballard A., Gilmer J., Dahl G., Vaswani A., Allen K., Nash C., Langston V., Dyer C., Heess N., Wierstra D., Kohli P., Botvinick M., Vinyals O., Li Y. J., Gulcehre C. (2018). Relational inductive biases, deep learning, and graph networks. arXiv preprint arXiv:1806.01261.

Inductive bias in the field of neural networks

Extreme Learning Machine



The inductive biases of ELM mainly include two assumptions as follows [32].

- the random feature mapping matrix is linearly related to the output matrix of the output layer.
- the optimal regression line has the smallest sum of squares of errors on the training dataset.

Wong P. K. et al. [32] pointed that the inductive biases followed by ELM are consistent with the linear regression algorithm and they analyzed the rationality of the reasoning process of ELM from the perspective of inductive bias.

[32] Wong P. K., Gao X. H., Wong K. I., Vong C. M. (2016). An analytical study on reasoning of extreme learning machine for classification from its inductive bias. Cognitive Computation, 8(4), 746-756.

Inductive bias in the field of neural networks

Snyders S. et al. [69] proposed a heuristic approach to make the network search from the point where the gradient has the maximal value in the weight space to achieve the goal of accelerating the convergence rate of the model. Later, they applied this method to recurrent neural networks [58].

Helmbold et al. [62][63] found that the inductive bias of neural networks induced by Dropout is different from standard regularizers such as L2 regularization. For example, Dropout enables penalty to grow exponentially with the increase of the hidden layer of neural networks, while L2 regularization can only cause linear growth. This feature allows Dropout to better balance the complexity and performance of the model.

Zhao S. J. et al. [33] empirically analyzed the inductive bias of GAN (Generative Adversarial Networks) and VAE (Variational Autoencoders) in the learning process by using a well-designed training dataset, and found that there are some similarities in dealing with the features of images between these algorithms and human cognition, such as emphasis on the salient features and ways to combine different features.

Li X. H. et al. [41] found that making the parameters of the fine-tuned model keep coherent with those of the pre-trained model is a very import criterion for transfer learning. Letting this criterion as the inductive bias of transfer learning and designing penalty terms of the optimization objective based on it would help to improve the prediction accuracy of the model.

[69] Snyders S., Omlin C. W. (2000). What inductive bias gives good neural network training performance?. In IJCNN 2000. (Vol. 3, pp. 445-450). IEEE.
[58] Snyders S., Omlin C. W. (2001, June). Inductive bias in recurrent neural networks. In International Work-Conference on Artificial Neural Networks (pp. 339-346).
[62] Helmbold D. P., Long P. M. (2015). On the inductive bias of dropout. The Journal of Machine Learning Research, 16(1), 3403-3454.
[63] Helmbold D. P., Long P. M. (2017). Surprising properties of dropout in deep networks. The Journal of Machine Learning Research, 18(1), 7284-7311.
[33] Zhao S., Ren H., Yuan A., Song J., Goodman N., Ermon S. (2018). Bias and generalization in deep generative models: An empirical study. In NIPS (pp. 10792-10801).
[41] Li X. H., Grandvalet Y., Davoine, F. (2018). Explicit inductive bias for transfer learning with convolutional networks. arXiv preprint arXiv:1802.01483.

Inductive bias in the field of neural networks

Ewald V. et al. [56] gave an example of how domain knowledge from the aerospace engineering field can be added to the design of deep learning algorithms. They regarded external knowledge as the inductive bias of model learning.

Mansour T. [37] proved that deep neural networks prefer simple and generalizable hypotheses when learning. The author used this inductive bias to explain the reason why deep neural networks are robust to the substantial noise, that is, noisy data is often unstructured, while this inductive bias makes deep neural networks preferentially learn those simple structured data first.

Yang X. et al [45] proposed a Scene Graph Auto-Encoder (SGAE) algorithm, which can incorporate the inductive bias of human language into the prevailing encoder-decoder image captioning framework to make the captions generated more human-like.

[56] Ewald V., Goby X., Jansen H., Groves R. M., Benedictus R. (2018). Incorporating Inductive Bias into Deep Learning: A Perspective from Automated Visual Inspection in Aircraft Maintenance. In Proc. 10th Intl Symposium on NDT in Aerospace, Dresden (pp. 1-9).
[37] Mansour T. (2019). Deep neural networks are lazy: on the inductive bias of deep learning (Master dissertation, Massachusetts Institute of Technology).
[45] Yang X., Tang K., Zhang H., Cai J. (2019). Auto-encoding scene graphs for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 10685-10694).

Inductive bias in the field of neural networks

Remarks:

Many factors can influence the inductive bias of neural networks such as the topology (i.e., the network structure), the type of activation function, the training regime, regularization, and the injection of prior knowledge.

These factors are often implicitly encoded into the learning algorithm, making it extremely difficult to quantify the inductive bias of the neural network.

Challenges

Challenges

1) Automating the selection of inductive bias (before and/or during learning) is still a very valuable and challenging task [5][65].

Although many research results have been achieved in recent years, there is still no universal way to ensure that the inductive biases for a new problem are chosen correctly and the strength is controlled flexibly.

No inductive bias can adapt to all the problems and the inductive bias adapting to the same problem may be quite different at different times. Therefore, some researchers classify inductive biases as static bias and dynamic bias [6][10][31].

Static bias means that the inductive bias is determined at the beginning of training and no adjustment in the learning process of the model.

Dynamic bias means that the inductive bias may be changed (in time and space) according to the need of learning so that the model can adapt to the current environment faster.

[5] Bensusan H. N. (1999). Automatic bias learning: an inquiry into the inductive basis of induction. University of Sussex.
[6] Utgoff P. E. (1986). Shift of bias for inductive concept learning. Machine learning: An artificial intelligence approach, 2, 107-148.
[10] Gordon D. F., Desjardins M. (1995). Evaluation and selection of biases in machine learning. Machine Learning, 20(1-2), 5-22.
[31] Utgoff P. E. (1983). Adjusting Bias in Concept Learning. In IJCAI (pp. 447-449).
[65] Brodley C. E. (1995). Recursive automatic bias selection for classifier construction. Machine Learning, 20(1-2), 63-94.

Challenges

2） Another important challenge is to ensure the correctness of inductive bias.

Different inductive biases correspond to different hypothesis spaces. Only the hypothesis space is learnable can the algorithm work well. For a specific hypothesis space, its learnability refers to whether the hypothesis space satisfies the probability that its generalization error is less than the error parameter in the confidence space. By analyzing the range of generalization error bounds of the hypothesis space under different conditions, it can be judged whether the hypothesis space is learnable.

The learnability of the hypothesis space reflects the correctness of the inductive bias. It is very important to ensure the learnability of the hypothesis space before learning.

How to measure it? PAC learning theory (Probably Approximately Correct)

Challenges

3) Besides, it is still difficult to properly measure and set the strength of inductive bias.

There is an important intrinsic relationship between the strength of inductive bias and the effectiveness of learning [28].

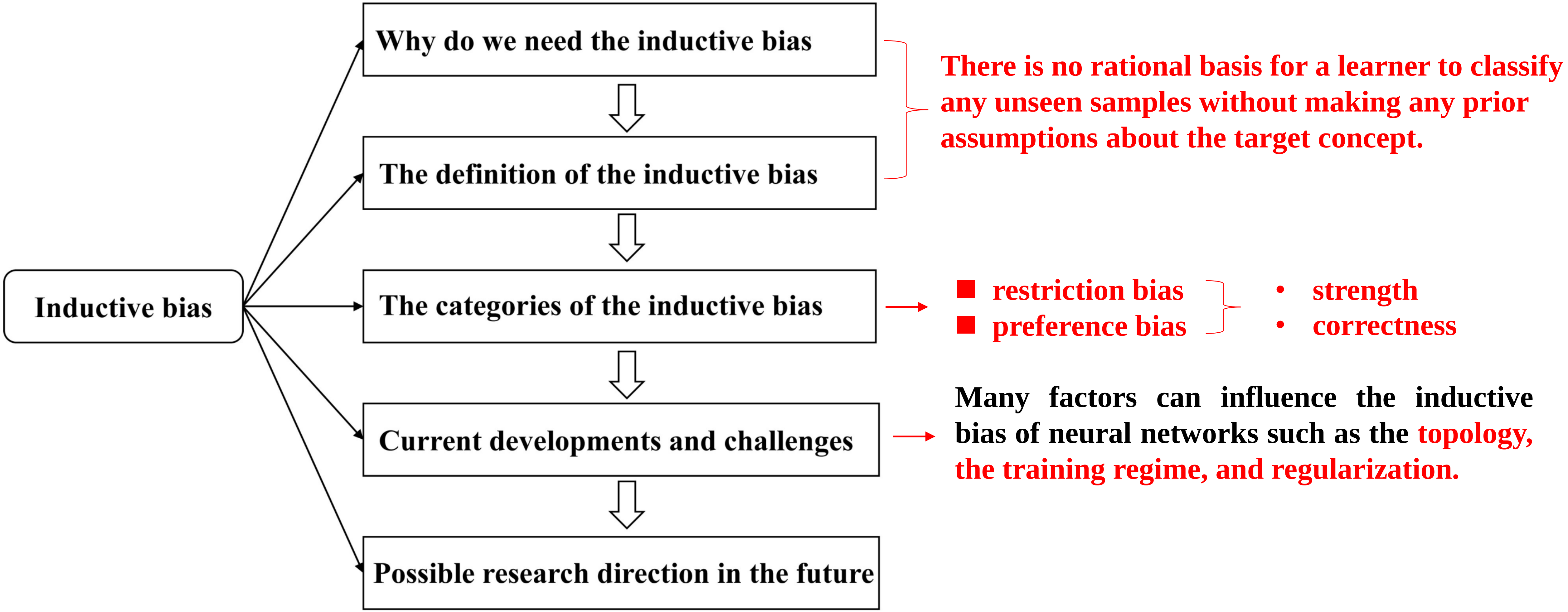
Assuming that the inductive bias chosen for a problem is correct, strong bias can accelerate the training process of the model. However, it is difficult to accurately measure the strength of the inductive bias.

It is very meaningful and challenging to build a quantitative analysis model among the strength of inductive bias, the characteristics of problems, and the performance of the model.

[28] Haussler, D. (1986, August). Quantifying the inductive bias in concept learning. In AAAI (pp. 485-489).

Conclusions and open questions

Conclusions and open questions



Conclusions and open questions

By the way, to get a machine learning model with good generalization ability, the prior knowledge of problems, appropriate inductive bias, and high-quality training samples are essential elements.

There is no free inductive bias. It is not true that the more inductive bias, the better the model. For example, DeepMind removes all Go-specific inductive biases and human historical data when training Alpha-Zero, but gets a better model than AlphaGo [50].

[50] Hessel M., van Hasselt H., Modayil J., Silver D. (2019). On inductive biases in deep reinforcement learning. arXiv preprint arXiv:1907.02908.

Conclusions and open questions

1) How to construct a relationship model between the characteristics of problems and inductive biases to help the selection of inductive bias for a new problem?

2) How to accurately measure the correctness and the strength of inductive bias in a simple way?

3) How to dynamically and timely choose appropriate inductive biases for lifelong learning models or online learning models?

4) Ensemble learning combines the inductive biases of different models and always achieves better performance than a single basic model. This phenomenon requires more theoretical proof and explanation from the perspective of inductive bias.

5) How to use inductive bias to help select the structure and hyper-parameters of neural networks?

6) How to inject human knowledge as an inductive bias to make the machine learning model more intelligent?

Thanks!