A Theory of Human-Like Few-Shot Learning

Zhiying Jiang[1], Rui Wang[2], Dongbo Bu[2], Ming Li[1*]

[1]David Cheriton School of Computer Science, University of Waterloo,

200 University Ave W, Waterloo, ON N2L 3G1, Canada

[2]Institute of Computing Technology, Chinese Academy of Science, Beijing, China

[*]To whom correspondence should be addressed; E-mail: mli@uwaterloo.ca

We aim to bridge the gap between our common-sense few-sample human learning and large-data machine learning. We derive a theory of human-like few-shot learning from von-Neuman-Landauer's principle. Modeling human learning is difficult as how people learn varies from one to another among eight billion people. Under commonly accepted definitions, we prove that all human or animal few-shot learning, and major models including Free Energy Principle and Bayesian Program Learning that model such learning, approximate our theory, under Church-Turing thesis. We find that deep generative model like variational autoencoder (VAE) can be used to approximate our theory and perform significantly better than baseline models including deep neural networks, for image recognition, low resource language processing, and character recognition.

Introduction

During the past decade, fast progress in deep learning (*1*) has enabled computer speech recognition, image processing, natural language processing, protein folding, game playing and many other applications. However these great progresses fell short when we try to understand our own learning mechanism: How to model human learning (*2*), (*3*), (*4*)?

Species in nature learn quickly to survive. When a dragonfly is hatched, within hours it firms up its wings then flies to catch mosquitoes; a newborn does not need tons of repeated examples or transfer learning to identify an apple. Most human or animal learning exhibits a mixture of inherited intelligence, few-shot learning without prior knowledge, as well as long term many-shot learning. It is interesting to note that these learning programs are encoded in

Work in progress.

our genomes but they are not all the same, even for individuals within the same species. The diversity of these learning algorithms is vividly expressed by Spearman's "g" factor (*2*).

Unlike data-laden, model-heavy, and energy-hungry deep learning approaches, most human learning appear to be simple and easy. Merely scaling up current deep learning approaches may not be sufficient for achieving human level intelligence. We miss certain major components when modeling human or animal learning.

Modeling human or animal few-shot learning is difficult. There are eight billion people on earth, each with a unique few-shot learning model (*5*). Even if we just want to model one person, a single person often uses different parameters, features, and perhaps different algorithms to deal with different learning tasks. Ideally we want a framework that can cover the diversity in human and animal few-shot learning. Facing such a seemingly formidable task, traditional thinking in machine learning will only lead us to various traps. To avoid such traps we need to go back to the very first principles of physics.

Specifically, we start from an agreed-upon law in thermodynamics, to formally derive our model for few-shot learning, and prove this is the optimal model within our framework in the sense that all other models including human ones may be viewed as approximations to our framework. We show a deep connection between our framework and the free energy principle (*3*) and the Bayesian Program Learning model (*4*). By the end of this process, a component of data compression during the inference phase of learning emerges as a key component of all few-shot learning models.

First, we formalize our intuitive and commonly accepted concept of human few-shot learning. For example, our definition below is consistent with what is used in (*4*), and in the same spirit of (*3*).

Definition 1. *Consider a universe Ω, partitioned into H disjoint concept classes: $\mathcal{C}_h$, $h = 1, 2, \ldots, H$. Few-shot (k-shot) learning is described as follows:*

1. *n elements in or outside Ω are given as unlabelled samples $y_1, \ldots, y_n$;*
2. *There are k labelled examples for each class $\mathcal{C}_h$, for small k;*
3. *The learning program, using a computable metric $\mathcal{M}$, few-shot learns $\mathcal{C}_h, h = 1, 2, ...H$, if it uses the n unlabelled samples and k labelled samples and minimizes the objective function:*

$$\sum_{h=1}^{H} \sum_{i=1}^{|\mathcal{C}_h|} \mathcal{M}(x_i, core_h) \mid y_1, \ldots, y_n, x_i \in \mathcal{C}_h,$$

where $core_h = \psi(k \text{ samples of } \mathcal{C}_h)$ representing a transformed representation of the k labeled samples from $\mathcal{C}_h$.

This definition covers most of our common sense few-shot learning scenarios and other studies. In particular, this is used in one-shot learning by (*4*). As each independent individual, we do not all use a same metric, or even similar metric, to few-shot learning. For example, MN Hebart et al (*6*) identified 49 highly reproducible dimensions to 1854 objects to measure their similarity. Different people can be equipped to better observe some of these dimensional features.

We explain the intuition behind Definition 1 via a simple example. A human toddler may have already seen many unlabelled samples of fruits which, for example, contains two classes: apples and pears. Then given a new labelled sample from each class, the toddler learns how to differentiate between these two fruits. The number of labeled data required for one to classify may vary as people have different learning algorithms.

Current deep learning based approaches for few-shot learning generally depend on 1) many auxiliary labeled training samples or task-specific data augmentation for transfer learning or meta learning (*7*); or 2) very large scale self-supervised pre-training (*8*). These approaches thus fall short to model few-shot learning in nature by humans and animals as they can hardly account for the diversity in learning algorithms and they either neglect the unsupervised scenario that humans are mostly exposed to or use the scale of unlabeled data and training parameters that are far beyond creatures need.

Many attempts have been made to understand human learning through cognitive, biological, and behavior sciences. Some studies have established basic principles a human learning model should obey. One theory is the two-factor theory of intelligence by Charles Spearman in 1904 (*2*). where the “g” factor is an indicator of the overall cognitive ability, and the “s” factor stands for the aptitude that a person possesses in specific areas. As “g” factor is genetically-related (*9*), it indicates the necessity of a learning theory that can account for the diversity in creatures’ learning ability. Another theory is the Free Energy Principle by Karl Friston (*3*) that human (and all biological systems) learning tends to minimize the free energy between internal understanding in the sense of Bayesian (under internal perceived distribution p) and that of the environmental event (under distribution q), measured by KL-divergence (*10*). In a similar spirit, Lake, Salakhutdinov and Tenenbaum (*4*) proposed a Bayesian program learning (BPL) model, learning a probabilistic model for each concept and achieve human-level performance. Two articles by Schmidhuber (*11*) and by Chater and Vitanyi (*12*) linked simplicity to human cognition and appreciation of arts.

Instead of exploring a biological basis for few-shot learning, we think it is possible to mathematically derive an optimal framework that can unify the above theories. We further demonstrate by experiments that our new model indeed works significantly better than other classical deep learning neural networks for few-shot learning. As a byproduct of our new model, a new concept class of "interestingness" is learned; this class implies where our appreciation of art, music, science and games comes from. Extending this observation, some aspects of consciousness may be modelled as a set of few-shot learned concepts. Consequently, the ability of labeling input data becomes a key step to acquiring some aspects of consciousness.

A theory of few-shot learning

We mathematically derive an optimal few-shot learning model for Definition 1 that is effective and is able to cover enormous diversities existed in different species. The task may appear to be formidable because of conflicting and seemingly very general goals: each individual is allowed to have a different learning model, yet our model has just one program to model everybody; we do not yet exactly know the complete underlying biological mechanisms, yet we need to implement the right functionality; there are infinite number of models, yet we need to choose one that is optimal; we are not really interested in "proposing models" out of blue, yet we wish our model to be a mathematical consequence of some basic laws of physics; the model needs to be theoretically sound, yet practically useful.

For simplicity and readability, we begin with one-shot learning, $k = 1$ in Definition 1. Thus, $core_h$ in Definition 1 is just the single labeled sample x_h. For larger k, $core_h$ can be some form of average of the k samples. As Definition 1 defined, some unlabelled objects are assumed and it's also possible to extend the definition by adding distribution, learnt from either unlabelled or labelled data, to Ω. Using metric $\mathcal{M}$ that is responsible for k-shot learning of an individual, the learning system seeks to minimize the energy function

$$\sum_{h=1}^{H}\sum_{i=1}^{|\mathcal{C}_h|}\mathcal{M}(x_i, x_h|y_1, \ldots, y_n),$$

or, assuming $\mathcal{H}(y_1, \ldots, y_n)$ is a pre-trained model of $y_1, \ldots, y_n$, or other labeled samples, capturing the distribution.

$$\sum_{h=1}^{H}\sum_{i=1}^{|\mathcal{C}_h|}\mathcal{M}(x_i, x_h|\mathcal{H}(y_1, \ldots, y_n)),$$

Now the question is, what sort of $\mathcal{M}$ should we use? Indeed, this varies from person to person. Can we unify all such measures, algorithms and inferences? Let's go back to the fundamentals.

Principle 1 (von-Neuman-Landauer Principle)**.** *Irreversibly processing 1 bit of information costs 1kT; reversible computation is free.*

Then for two objects x, y, the minimum energy needed to convert between x and y in our brain is:

$$\mathcal{E}_U(x, y) = \min\{|p| : U(x, p) = y, U(y, p) = x\},$$

where U is a universal Turing machine or our brain, assuming Church-Turing thesis. Since we can prove a theorem showing all Universal Turing machines are equivalent modulo a constant and efficiency, we will drop the index U (see (*13*)). To interpret, $\mathcal{E}(x, y)$ is the length of the shortest program that reversibly converts between x and y. These bits used in the shortest program p when they are erased will cost $|p|$kT of energy, according to the John von Neuman and Rolf Landuaer's law. This leads us to a fundamental theorem (*14*):

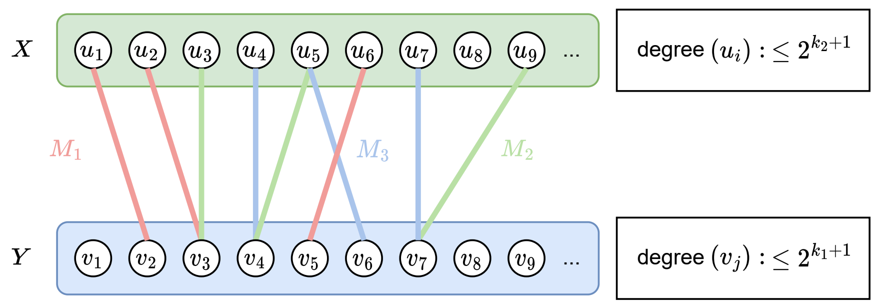

Figure 1: Bipartite Graph

Theorem 1. $\mathcal{E}(x, y) = \max\{K(x|y), K(y|x)\} + O(1)$.

$K(x|y)$ is the Kolmogorov complexity of x given y, or informally, the length of the shortest program that outputs x given input y (details are shown in (*13*)). As this theorem was proved thirty years ago and it is vital in our theory, to help our readers, we will provide an intuitive but less formal proof here.

Proof. By the definition of $\mathcal{E}(x, y)$, it follows $\mathcal{E}(x, y) \geq K(x|y)$ and $\mathcal{E}(x, y) \geq K(y|x)$, thus we have $\mathcal{E}(x, y) \geq \max\{K(x|y), K(y|x)\}$.

To prove the other direction $\mathcal{E}(x, y) \leq \max\{K(x|y), K(y|x)\}$, we need to construct a program $\mathtt{p}$ such that $\mathtt{p}$ outputs y on input x and $\mathtt{p}$ outputs x on input y, and length of $\mathtt{p}$ is bounded by $\max\{K(x|y), K(y|x)\} + O(1)$.

Let $k_1 = K(x|y)$, and $k_2 = K(y|x)$. Without loss of generality, assume $k_1 \leq k_2$. We first define a bipartite graph $\{X, Y, E\}$, where $X, Y = \{0, 1\}^*$, as shown in Figure 1 and E is a finite set of edges defined between X and Y as follows:

$$E = \{\{u, v\}, u \in X, v \in Y, K(u|v) \leq k_1, K(v|u) \leq k_2\}$$

Note that a particular edge (x, y) is in E. If we find edge (x, y), then given x, $\mathtt{p}$ can output y, and vice versa. So the idea of the proof is to partition E properly so that we can identify (x, y) easily. Two edges are disjoint if they do not share nodes on either end. A matching in graph theory is a set of disjoint edges in E.

Claim. E can be partitioned into at most 2^{k_2+2} matchings.

Proof of Claim. Consider edge $(u, v) \in E$. The degree of a node $u \in X$ is bounded by 2^{k_2+1} because there are at most 2^{k_2+1} different strings v such that $K(v|u) \leq k_2$, accumulating possible strings from $i = 1$ to $i = k_2$ gives us $\sum_{i=1}^{i=k_2} = 2^{k_2+1} - 2$. Hence u belongs to at most 2^{k_2+1} matchings. Similarly, node $v \in Y$ belongs to at most 2^{k_1+1} matchings. We just need to put edge (u, v) in an unused matching. (End of Proof of Claim)

Let M_i be the matching that contains edge (x, y) We now construct our program $\mathtt{p}$. $\mathtt{p}$ operates as follows:

- Generate M_i following the proof of Claim, i.e. enumerating the matchings. This uses information k_1, k_2, and i. $K(i) \leq k_2 + O(1)$

- Given x, p uses M_i to output y, and given y, p uses M_i to output x.

$\square$

A conditional version of Theorem 1, using information in Definition 1, can be obtained $\mathcal{E}(x, y|y_1, \ldots, y_n) = \max\{K(x|y, y_1, \ldots, y_n), K(y|x, y_1, \ldots, y_n)\}$, conditioning on unlabelled samples $y_1, \ldots, y_n$. According to (*14*), this distance is universal, in the sense that $\mathcal{E}(x, y)$ is the minimum among any other computable distances:

Theorem 2. *For any computable metric D, there is a constant c, such that for all x, y, $\mathcal{E}(x, y) \leq D(x, y) + c$.*

This theorem implies: if D metric finds some similarity between x and y, so will $\mathcal{E}$. Thus, the above theorem implies, up to some constant $O(H)$

$$\sum_{h=1}^{H}\sum_{i=1}^{|\mathcal{C}_h|}\mathcal{E}(x_i \in \mathcal{C}_h, core_h|y_1, \ldots, y_n) \leq \sum_{h=1}^{H}\sum_{i=1}^{|\mathcal{C}_h|}\mathcal{M}(x_i \in \mathcal{C}_h, core_h|y_1, \ldots, y_n).$$

When unlabeled samples $y_1, \ldots, y_n$ plus other irrelevant historical labeled samples are modeled by some model $\mathcal{H}$ such as a generative model (e.g., VAE), then the above inequality can be rewritten as:

$$\sum_{h=1}^{H}\sum_{i=1}^{|\mathcal{C}_h|}\mathcal{E}(x_i \in \mathcal{C}_h, core_h|\mathcal{H}) \leq \sum_{h=1}^{H}\sum_{i=1}^{|\mathcal{C}_h|}\mathcal{M}(x_i \in \mathcal{C}_h, core_h|\mathcal{H}). \tag{1}$$

Thus, $\mathcal{E}$ gives optimal metric for few-shot learning algorithm. Other algorithms satisfied Definition 1 are the approximation to this optimal solution. [1]

In addition, we show that our theory's deep connection to two well-established principles of learning in neuroscience and psychology. Friston's Free Energy Principle (FEP) (*3*), derived from Bayesian brain hypothesis (*15*), states that brain seeks to minimize surprises. Specifically, it assumes the brain has its internal state (a.k.a. generative model) that implicitly models the environment according to the sensory data. Hidden (latent) variables need to be defined for the internal state, which are drawn from prior beliefs. Ideally, these prior knowledge is also modelled, which is made possible by hierarchical generative models. The free energy principle (FEP) is often interpreted as Bayesian optimization, using the Evidence Lower Bound (ELBO) as ELBO $= \log p(x; \theta) - D(q(z) \| p(z|x; \theta)$ optimization function. Here the evidence $\log p(x; \theta)$ is the encoding length of x under probability p, and the Kullback-Leibler divergence term is the p-expected encoding length difference. This is half of Theorem 1 and FEP is asymmetric if we view it as a distance. However, the symmetry is important to few-shot learning. For example, a scarlet king snake may look like a coral snake, but the latter certainly has more deadly features the former lacks, one way compression $K(\mathrm{ScarletKingSnake}|\mathrm{CoralSnake})$ is not sufficient to

[1]Note that $\mathcal{E}$ is a metric: it is symmetric, and satisfies triangle inequality

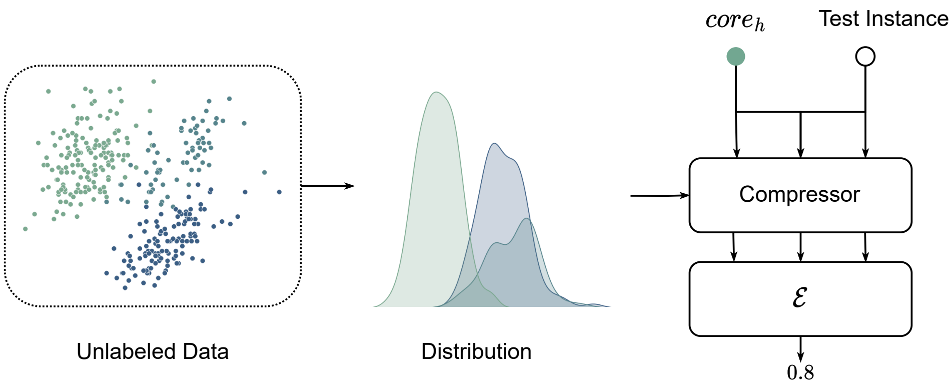

Figure 2: Illustration of our framework, dashed line indicates optional component when learning.

distinguish the two. Despite of the fact *H. influnza* with genome size 1.8 million and *E. coli* with genome size 5 million they are sister species but *E. coli* would be much closer to a species with zero genome G_0 or just a covid-19 genome with this asymmetric measure ($K(G_0|E.coli)$ than with *H. influnza* ($K(H.\ influnza|E.\ coli)$). A symmetric interpretation of Friston's FEP can be derived by requiring minimum conversion energy as we show in Theorem 1.

Different individuals may use different compression algorithms to do data abstraction and inference. It can be viewed that these algorithms all approximate $\mathcal{E}(x, y)$. Some are more efficient than others in different situations. The individuals with better compression algorithms have bigger "g" factor. Diversified compression algorithms also guarantee better survival chances of a community when facing a pandemic. As compression neural networks are genetically encoded, the "g" factor is thus inheritable. This can be seen via Figure 2, compression algorithms vary from one to another. The distribution of the data to be learnt is either implicitly or explicitly captured by creatures. Those who can better utilize unlabeled data to capture distribution may have a more efficient compression algorithm.

Experimental Results

Image Experiments

To approximate our universal few-shot learning model, we use a hierarchical VAE as our underlying model $\mathcal{H}$ in Inequality 1 to model the unlabelled samples $y_1, \ldots, y_n$. This hierarchical structure coincides with our visual cortex and brain structure (*16*). According to integrated information theory (*17*), an input y may come from all sensing terminals: vision, hearing, smell, taste, sensation. Often, creatures are exposed to an unsupervised environment where objects are unknown and unlabeled. Revisiting the negative ELBO, we can see it can be interpreted as changing perceptions to minimize discrepancy (minimize KL divergence) or changing observations to maximize evidence, in the context of FEP. When the creatures are exposed to a "tree" and they do not fully realize what it is, the sensory information of the objects are internalized with hidden states (inner belief) that can describes how it believes the generation process of a "tree". This process of generation, helps the creatures to identify the latent similarities among objects that belong to the same category, without the full awareness. This process of "unconsciously" training to generate helps the creatures to better categorize in future. When the identity of a "tree" is finally revealed, they can generalize quickly. This explains our rationale of using a VAE to process unlabelled samples. Consequently, the Kolmogorov complexity terms in Inequality 1 are naturally approximated by a VAE based compressor (*18*).

To test the hypothesis, we carry out the experiment on five datasets, MNIST, KMNIST, FashionMNIST, STL-10 and CIFAR-10. We first train a hierarchical VAE on unlabeled data to learn to generate $\hat{x}$ that's as close to x as possible. This corresponds to the time when creatures exposed a environment without knowing the object, implicitly learning the latent representation among objects. When the identity of objects are revealed, a VAE based universal compressor can be used to identify the new objects. Specifically, after training a hierarchical VAE unsupervisedly, we compare the $\mathcal{E}$ energy function between a labelled image and a test image, as in Definition 1. In our experiment, we use 5 labeled samples per class to test the accuracy of classification. The energy function $\mathcal{E}$ relies on a compressor to approximate. We thus use the bits-back argument to directly use our trained VAE for the compressor in (*18*). Our result shows that using only 5 samples, our method outperforms traditional supervised models like SVM, CNN, VGG and Vision Transformer (ViT) on all five datasets. These supervised methods are chosen to represent different model complexity with wide range of number of parameters. As we can see, when labeled data are scarce, supervised methods are not effective: complex models like VGG cannot perform better than SVM and this tendency is more obvious on ViT without pre-training. The improvement that our method brings is more obvious on more complex datasets like STL-10 and CIFAR-10. Similar result is also obtained in the recent work, across different shot settings (*19*).

We also compare with using latent representation directly with k-Nearest-Neighbor classifier, labelled as "Latent" in the table. The architecture and training procedure for "Latent" method is exactly the same to our method — we train on unlabelled data to generate the sample and then take the latent representation for classification. We can see using latent representation outperforms all supervised methods on four out of five datasets. But the accuracy is still way lower than our method, indicating our method can better utilize the generative models.

Text Experiments

Our theory is generally applicable, even without pre-training on unlabeled data. Here, we demonstrate significant advantages of our approach with a simple compressor *gzip* over lower resource languages.

Languages with Abundant Resources We first test our method on datasets with abundant resources. Specifically, we compare with three datasets — AG News, SogouNews and DBpedia. Similar to image classification, we compare with both supervised methods, including fasttext (*20*),

	MNIST	KMNIST	FashionMNIST	STL-10	CIFAR-10
SVM	69.4±2.2	40.3±3.6	67.1±2.1	21.3±2.8	21.1±1.9
CNN	72.4±3.5	41.2±1.9	67.4±1.9	24.8±1.5	23.4±2.9
VGG	69.4±5.7	36.4±4.7	62.8±4.1	20.6±2.0	22.2±1.6
ViT (disc)	58.8±4.6	35.8±4.1	61.5±2.2	24.2±2.5	22.3±1.8
Latent	73.6±3.1	48.1±3.3	69.5±3.5	31.5±3.7	22.2±1.6
Ours	77.6±0.4	55.4±4.3	74.1±3.2	39.6±3.1	35.3±2.9

Table 1: 5-shot image classification accuracy on five datasets.

	AG News	SogouNews	DBpedia
fasttext	27.3±2.1	54.5±5.3	47.5±4.1
Bi-LSTM+Attn	26.9±2.2	53.4±4.2	50.6±4.1
HAN	27.4±2.4	42.5±7.2	35.0± 1.2
W2V	38.8±18.6	14.4±0.5	32.5±11.3
BERT	80.3±2.6	22.1±4.1	96.4±4.1
Ours	58.7±4.8	64.9±6.1	62.2±2.2

Table 2: 5-shot text classification accuracy on three datasets.

BiLSTM (*21*) with attention mechanism (*22*) and Hierarchical Attention Network (HAN) (*23*), and non-parametric methods that use Word2Vec (W2V) (*24*) as representation. We also compare with pre-trained language models like BERT (*25*) We use five labeled data for each class (5-shot) for all the methods.

Surprisingly, even without any pre-training and with a simple compressor like *gzip*, our method outperforms all non-pretrained supervised methods and non-parametric methods in low data regime. This indicates that compressor serves as an efficient method to capture the regularity and our information distance is effective in comparing the similarity based on the essential information. When comparing with pre-trained models like BERT, we can see our method is significantly higher on SogouNews, a special dataset that includes Pinyin — a phonetic romanization of Chinese, which can be viewed as an Out-Of-Distributed (OOD) dataset as it uses the same alphabet as `english` corpus.

Low-Resource Languages Sufficiently pre-trained language models are exceptional few-shot learners (*8*). Our few-shot methodology come to rescue by providing a trade-off strategy when we face low resource data or distributions that are significantly different from any pre-trained data. Therefore, we compare our method with BERT on four different low-resource language datasets - `Kinyarwanda`, `Kirundi`, `Swahili` and `Filipino`. These datasets are curated to have the Latin alphabets, same as `english` corpus. BERT has performed extremely well as shown in Table 2 due to pre-training on billions of tokens. However, when facing low-resource

	Kinnews	Kirnews	Swahili	Filipino
BERT	24.0±6.0	38.6±10.0	39.6±9.6	40.9±5.8
mBERT	22.9±6.6	32.4±7.1	55.8±16.9	46.5±4.8
Ours	45.8±6.5	54.1±5.6	62.7±7.2	65.2±4.8

Table 3: 5-shot text classification accuracy on low-resource datasets

datasets, BERT perform significantly worse than our method only using *gzip* as we can see in Table 3, no matter using multilingual pre-trained version or the original one. Note that mBERT is pre-trained on 104 languages including `Swahili` and `Tagalog` (on which Filipino is based on). As we can see on `Swahili` and `Filipino`, mBERT performs better than BERT, but still significantly lower than our method.

Omniglot one-shot-classification dataset

In (*4*), a one-shot learning framework Bayesian program learning (BPL) was proposed. It learns a simple probabilistic model for each concept. Taking a negative logarithm converts a Bayesian formula to a description length paradigm, hence BPL can be viewed as one particular approximation to our theory. Here we provide another simple approximation of our theory for the Omniglot one-shot-classification dataset of (*4*).

Our system first decompose a given character into strokes, then compute $\mathcal{E}(a, b)$ between characters a and b, using all their possible stroke decomposition. We provide how to calculate $\mathcal{E}(a, b)$ here and details of decomposition program is given in Appendix A.

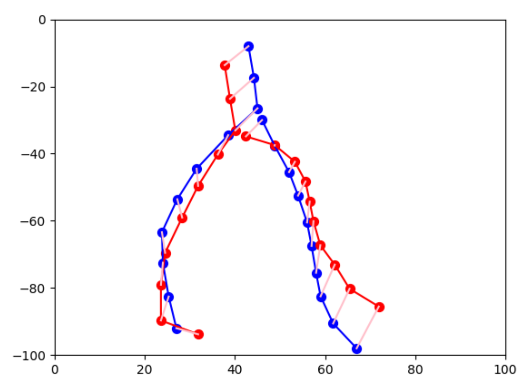

Figure 3: Distance between two Bezier curves

1. Fit a stroke by a Bezier curve;
2. Ensure the number of points on two curves are same. This algorithm utilize equally split method to select certain same number of points on each curve Figure 3;
3. Ensure the area of the convex hull and the barycenter of the compared characters are the same;
4. Use max Cartesian distance between parallel points on two Bezier curves to approximate the minimum encoding distance between two Bezier curves, as shown in Figure 3;

5. Choose the character with minimum distance.

This simple implementation achieves 92.25% accuracy 20-way-1-shot on this dataset. The point here is to demonstrate various approximations of our theory that work rather than comparing accuracy. At 96.75% (*4*) or at 92.25% might be two different individuals with different compression algorithms.

Unification

Our framework can unify other popular deep neural networks for few-shot learning.

Siamese Network: Siamese network uses twin subnetwork to rank the similarity between two inputs in order to learn useful features. $\mathcal{M}$ here is often a contrastive loss. This framework shows strong performance in one-shot image recognition (*26*).

Prototypical Network: Prototypical networks (*27*) propose to optimize the distance metric $\mathcal{M}$ directly by learning $core_h$ in representation space. $core_h$ are represented as the mean of embedded support samples.

Bi-Encoder: In the context of natural language processing, one of the dominant structure is the Bi-Encoder design with each encoder being a pre-trained language model. For example, in information retrieval, Dense Passage Retrieval (DPR), with two encoders encoding query and document respectively, has become the new state of the art. To capture semantic similarity, sentenceBERT (*28*) also adopts the bi-encoder design and becoming one of the most prevalent methods for semantic textual similarity. $\mathcal{M}$ in both cases can either be cosine similarity or Euclidean distance between representation learned through pre-trained models.

Information Distance based Methods: Hundreds of algorithms were published, before the deep learning era, on parameter-free data mining, clustering, anomaly detection, classification using information distance $\mathcal{E}$ (*29–34*), with a comprehensive list in (*13*). Recently (*19*) have discovered using information distance with deep neural networks and leverage the generalizability of few-shot image classification. This work shows that with the help of deep generative models, unlabelled data can be better utilized for few-shot learning under our framework.

Conclusion and a discussion on consciousness

We have defined human-like few-shot learning and derived an optimal form of such few-shot learning. Note there is an interesting difference between our theory and classical learning theory. In classical learning theory, it is well-known that if we compress training data to a smaller consistent description, whether it is a classical Bayesian network or a deep neural networks (*13, 35*), we would achieve learning. In this paper, we observe a tradeoff phenomenon that compression also happens in at the inference stage when there are not enough labelled data to train a small model. On the biological side, compression circuits using predictive coding in human cortex has been studied by (*36*). Experiments have also strongly supported our theory.

We expect to see more practical systems approximating our theory can be implemented to solve commonplace few-shot learning problems when large amounts of labelled data for deep learning is lacking. We now wish to explore two consequences of our few-shot learning model, to consciousness.

A binary classifier of interestingness

Our few-shot learning model has a by-product. We have proved compression is a universal goal that few-shot learning algorithms approximate. Thus this implies immediately a (subconscious) *binary classifier*: if something is compressed, then something interesting happens, and attention is given. It turns out that this "Interestingness" has been theoretically studied as logical depth first proposed by Charles Bennett (*13*). According to Bennett, a structure is deep if it is superficially random but subtly redundant. When few-shot learning happens, significant compression happens, and these deep objects gain attention. Such a binary classifier might explain our appreciation of arts, music, games, and science, since these all share a common feature of dealing with non-trivially compressible objects: whether it is a shorter description of the data that gives rise of Newton's laws (*13*), or a piece of art or music that itself is compressible or that reminds us of something we have experienced before, hence very compressible, we feel we understand it and hence appreciate it. Science is nothing but compressing data into simpler descriptions of nature.

Consciousness and the ability of labeling data

Do other species have consciousness? The difficulty to answer this question is consciousness is not testable. Thomas Nagel (*37*) made similar comment: We will never know if a bat is conscious because we are not bats.

Consider an alternative data-driven approach by asking what a species can do instead of how they feel. That is, if we treat some aspects of consciousness as a collection of learned concepts, then if we already have a compression network, the ability of acquiring the relevant concepts becomes a matter of labelling relevant data. We know learning and consciousness are both located at posterior cortex region (*38*). This is in agreement with some injured patients when they lost consciousness. This is also in agreement with "bistable perception" training results with monkeys (*39*).

Many different kinds of consciousness are being pragmatically studied (*40*). These include: 1) the ability of consciously perceive the environment; 2) the ability of evaluating some conscious emotions; 3) the ability of having a unified conscious experience; 4) the ability of integrating across time as a continuous stream, one moment flowing into the next; 5) the conscious awareness of oneself as distinct from the world outside. Many of these abilities may be seen as a few-shot learnable concepts, given properly labeled data.

Different animals have various levels of some of such consciousness by passing certain tests. For example, chimpanzees, dolphins, Asian elephants, and magpies can recognize themselves by passing some mirror-mark tests. The corvids display some emotions, and are able to plan ahead.

Octopus have powerful perceptual facilities obtaining and processing data independently with each foot. Experimentally, awareness emerges when information travels back and forth between brain areas (*41*) instead of a linear chain of command.

According to our theory, the brain really only needs to use a universal compressor to compress information, regardless of one processor in the head or a few processors in the feet (in case of Cephalopods). Thus we can conjecture that "consciousness" then is a matter of *ability of labelling the data* from sensory terminals. Food or enemy in the environment are easy to label. Emotional labeling requires some level of abstraction. Self-awareness of "me" and "others" thus is just another binary classifier trainable depending on if the species is able to do "displaced reference" mental labeling. Other than the human beings, only orangutans are known to have limited displaced reference ability (*42*).

Thus we have just reduced the non-testable question of whether an animal has consciousness in some aspects to if it is able to label the corresponding data properly.

Acknowledgement

We thank Dr. Hang Li for suggestions and bringing (*43*) to our attention and Dr. Amy Sun for bringing (*44*) to our attention. The work is supported in part by Canada's NSERC operating grant OGP0046506, Canada Research Chair Program, and the Leading Innovative and Entrepreneur teams program of Zhejiang, number 2019R02002, and NSFC grant 61832019.

References and Notes

1. Y. LeCun, Y. Bengio, G. Hinton, *Nature* **521**, 436 (2015).
2. C. Spearman (1961).
3. K. Friston, *Nature Review Neuroscience* **11**, 21 (2010).
4. B. M. Lake, R. Salakhutdinov, J. B. Tenenbaum, *Science* **350**, 1332 (2015).
5. E. Stern, *npj Science of Learning* **2**, 1 (2017).
6. M. Hebart, C. Zheng, F. Pereira, C. Baker, *Nature, Human Behaviour* pp. 1173–1185 (2020).
7. C. Finn, P. Abbeel, S. Levine, *International conference on machine learning* (PMLR, 2017), pp. 1126–1135.
8. T. Brown, *et al.*, *Advances in neural information processing systems* **33**, 1877 (2020).
9. T. J. Bouchard Jr, *Annals of Human Biology* **36**, 527 (2009).
10. M. N. Bernstein, *mbernste.github.io/posts/elbo/* .

11. J. Schmidhuber, *arXiv:0812.4360v2 [cs.AI]* (2009).

12. N. Chater, P. Vitányi, *Trends in Cognitive Sciences* **7**, 19 (2003).

13. M. Li, P. Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications* (Springer-Verlag, 1993, 1997, 2008, 2019).

14. C. Bennett, P. Gács, M. Li, P. Vitányi, W. Zurek, *IEEE Trans. Inform. Theory* **44**, 1407 (1998).

15. D. C. Knill, A. Pouget, *TRENDS in Neurosciences* **27**, 712 (2004).

16. K. Friston, *PLoS computational biology* **4**, e1000211 (2008).

17. C. Koch, G. Tononi, *Scientific American* **304** (2011).

18. J. Townsend, T. Bird, D. Barber, *International Conference on Learning Representations* (2018).

19. Z. Jiang, Y. Dai, J. Xin, M. Li, J. Lin, *Advances in Neural Information Processing Systems* (2022).

20. A. Joulin, E. Grave, P. B. T. Mikolov, *EACL 2017* p. 427 (2017).

21. M. Schuster, K. K. Paliwal, *IEEE transactions on Signal Processing* **45**, 2673 (1997).

22. Y. Wang, M. Huang, X. Zhu, L. Zhao, *Proceedings of the 2016 conference on empirical methods in natural language processing* (2016), pp. 606–615.

23. Z. Yang, *et al.*, *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies* (2016), pp. 1480–1489.

24. T. Mikolov, K. Chen, G. Corrado, J. Dean, *arXiv preprint arXiv:1301.3781* (2013).

25. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (2019), pp. 4171–4186.

26. G. Koch, R. Zemel, R. Salakhutdinov, *et al.*, *ICML deep learning workshop* (Lille, 2015), vol. 2, p. 0.

27. J. Snell, K. Swersky, R. Zemel, *Advances in neural information processing systems* **30** (2017).

28. N. Reimers, I. Gurevych, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (2019), pp. 3982–3992.

29. M. Li, *et al.*, *Bioinformatics* **17**, 149 (2001).

30. E. Keogh, S. Lonardi, C. A. Ratanamahatana, *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (2004), pp. 206–215.

31. C. H. Bennett, M. Li, B. Ma, *Scientific American* **288**, 76 (2003).

32. M. Nykter, *et al.*, *Physical review letters* **100**, 058702 (2008).

33. D. Benedetto, E. Caglioti, V. Loreto, *Physical Review Letters* **88**, 048702 (2002).

34. M. Nykter, *et al.*, *Proceedings of the National Academy of Sciences* **105**, 1897 (2008).

35. Y. Bengio, *et al.*, *Foundations and trends® in Machine Learning* **2**, 1 (2009).

36. R. P. Rao, D. H. Ballard, *Nature neuroscience* **2**, 79 (1999).

37. T. Negel, *Readings in philosophy of psychology* (1974).

38. C. Koch, *Scientific American.* (2018).

39. G. Miller, *Science* **309**, 79 (2005).

40. J. Birch, A. Schnell, N. Clayton, *Trends in cognitive sciences* (2020).

41. M. Boly, *et al.*, *Science* **332** (May, 2011).

42. H. Lyn, *et al.*, *Animal Cognition* **17** (2014).

43. Y. Ma, D. Tsao, H. Shum (2022).

44. F. Scherr, C. Stöckl, W. Maass, *BioRxiv* (2020).

A Algorithm for extracting strokes from a character

Repeat until all pixels of a character are marked, by depth-first search:

(1) Extract its skeleton so that the stroke width is 1 pixel point. Then convert the image to a graph and shrink adjacent cross points. (2) Randomly select an endpoint as starting point, endpoint at top left has a greater chance of being selected. Walk until a cross point or endpoint. If there is a circle then select a cross point of a top left point if there is no cross point. Record this stroke and mark it on the character. Allow small number of marked pixel points to make

the decomposition more natural. (3) When meeting a cross point, then enumerate two situations of pen-up and turning, randomly. Pen-up means end of a stroke, go to step (2) with the marked graph. Turning means continuation hence repeat step (2). If walking to an endpoint, then attempt to turn by going back to find a new unmarked pixels within some small number of pixels or directly end the stroke and repeat step (2) with marked graph.