

機械学習技術による創作物の知的財産評価のための情報理論測度について

金子 格†
Itaru Kaneko

1. はじめに

機械学習(以下ML)の利用拡大により「創作物」すなわち従来人間が創作活動によって作り出してきた絵画、音楽の生成への応用の可能性が広がってきた。

機械学習による創作物の例を図1に示す。この例はDeepDream[1]により生成された画像であり creative commons に収録されている[2]。以下ではこのような創作物を知的財産として評価する場合にその帰属をどう考えるかを議論する。

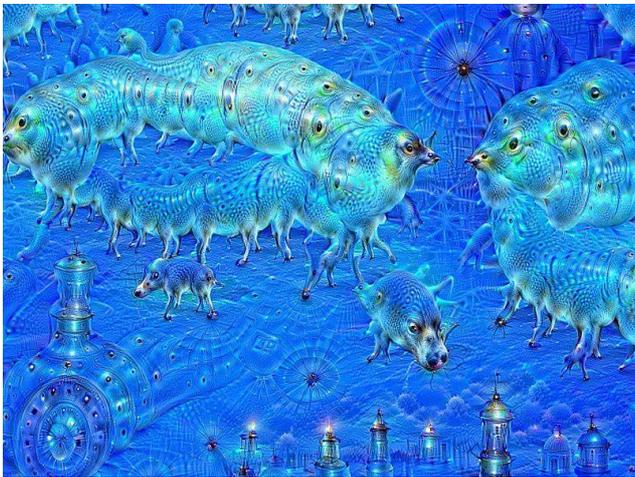


図1 機械学習による作画例(Aurelia-auria-3-0009.jpg, https://commons.wikimedia.org/wiki/Category:Deep_Dream_images, edited on 15 May 2018, at 13:20)

これらの「創作物」への機械学習の利用は広いスペクトルにわたっているが、機械学習の「学習データ」への依存度においても様々な段階がある。すなわち「学習データ」の単純コピーに近いものから、学習データの「統計的性質」を抽出して利用するものまでといったかなりの幅がある。以下では、これらの知的財産の評価においてなんらかの測度としていかなる基準がありえるかを論ずる。

2. 機械学習による創作のモデル

2.1. 機械学習による創作の参照モデル

筆者はすでにAIによる創作の参照モデルとして図2に示すモデルを提唱している[3][4]。この参照モデルをここでも機械学習による創作のモデルとして利用することにする。

F:知財化候補データはシステムの出力である「創作物」である。人間の作者は A:指令と、B:作者入力情報を入力し、D:補助機構を操作して創作を行う。C:参照情報とは、機械

学習の場合とレーニングデータセットである。内部の E:蓄積情報とは機械学習の場合学習によってつくられたネットワークに対応する。一般的な AI による創作の場合は D:補助機構はあらゆるプログラムでありえる。機械学習の場合は D はあらかじめ準備された機械学習を含むアルゴリズムである。たとえば DeepDream などがあげられる。

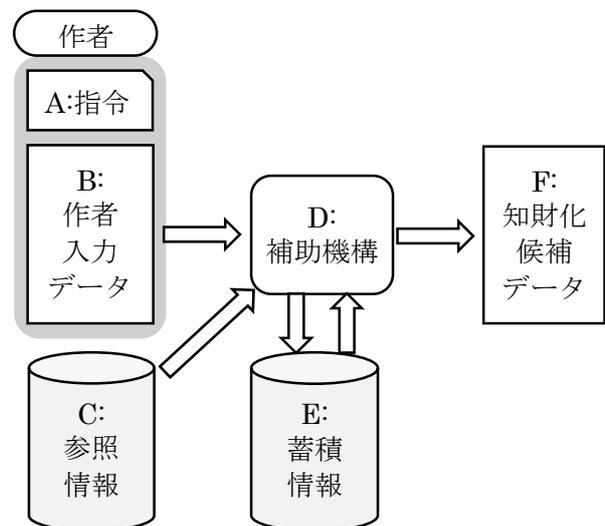


図2 AIによる創作の参照モデル

2. 情報量の影響

2.1. 均等確率の蓄積情報

DeepDream 型の生成において、操作者は画像としての作者入力データはほとんど与えず、機械学習結果と参照情報から出力を得る。そこで機械学習結果と入力画像から出力を得る状況について以下で考察する。

図3に参照情報と蓄積情報から出力を得る状況を示す。この場合情報量を尺度とすることができないかを考察する。

例として文字列情報の生成を考え、均等確率の E1:蓄積情報と C1:参照情報から F1:知的化候補データを生成するとする。蓄積情報 E1 を表 1, 指示:A1 を表 2 に示す。蓄積情報:E2 には、様々な長さの文字列が集積され、C1:参照情報はその蓄積情報をピックアップする符号である。適切な符号により任意の長さの任意の文字列が選択できるから任意の英文を生成することができる。この場合 F:知財化候補データの価値に E1 は寄与しておらず C1 が知財化候補データに価値を与えていると考えられる。

画像の場合同様な例として 8x8 の画素の単位で大量の画像の画素間の相互相関を学習させると E1 には DCT 基底が

† 東京工芸大学

生成されるはずである。DCT 基底を用いてももちろんあらゆる画像が生成可能である。DCT 基底は多くの画像の特徴を抽出してはいるが、DCT 基底を使ってあらゆる画像は表現可能であるし、DCT 基底自体は単純な濃淡模様でしかないから、DCT 基底に創作物の価値を求めることはできないだろう。

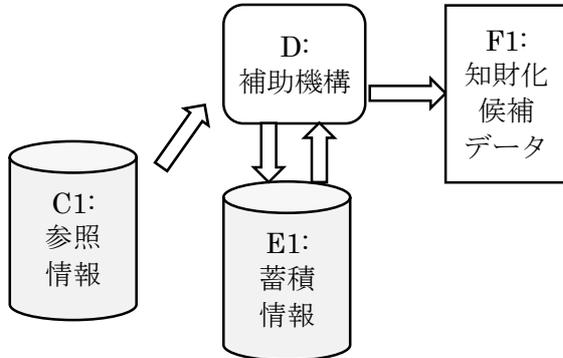


図 3 均等確率分布による蓄積情報 E1:により F1:知財化候補データを生成する場合

表 1 E1:蓄積情報. E1 では各 s と l について以下の文字列が蓄積情報として蓄積されている。文字列の分布はすべての文字列パターンで均等である。

	l=1	l=2	l=3	l=4
H1[bit]	4.75	9.5	14.25	19
S				
1	A	AA	AAA	AAAA
2	B	AB	AAB	AAAB
...
26	Z	AZ	AAZ	AAAZ
27		BA	ABA	AABA
28		BB	ABB	AABB
...	

表 2 参照データ C1, C1 では以下のような指示により文字列が出力される

Command	(s, l)	出力	H1[bit]
1	1, 1	I	4.75
2	2, 12	AM	9.5
3	2, 374	OK	9.5

2.1 不均等確率の蓄積情報

次に図 4 に蓄積情報 E2 が表 3 に示す不均等な確率分布によって構成されている場合を示す。この場合、E2 には英文が蓄積されているがその出現順序は英文中で利用される頻度の順に並べられている。したがって有名なフレーズほど上位になり、上位のフレーズは短い記号で指定することができるとする。

表 4 に示す参照データ C2 は均等確率の場合と同様に、蓄積情報の選択を示す指示である。ここでは 2 つの文章を選

んでいる。そしてこの 2 つの文章を接続すれば Simon & Garfunkel, "The Sound Of Silence" の冒頭の歌詞になっている。

この場合 E2 の上位には単純な文が多く収容されているが、やがて大量の、しかし英文の中で最も有名なフレーズが連続してあらわれることになる。したがってまったくでたらめな指令を行っても、出力には有名なセンテンスが多く出力されることになる。

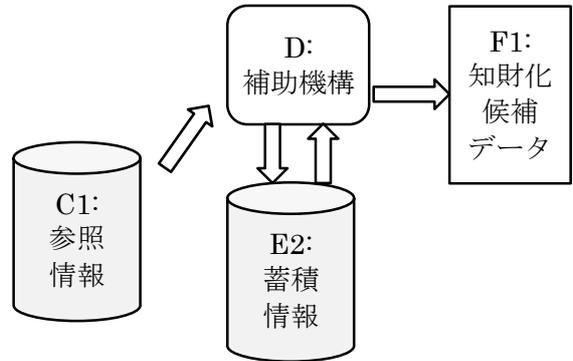


図 4 不均等確率分布による蓄積情報 E2:により F2:知財化候補データを生成する場合

表 3 蓄積情報 E2. E2 では出現頻度を考慮した英文の表の一例。頻度が多い文に短い可変長符号を表した場合の bit 数を H1 に示している。

S	H1[bit]	Sentence
1	1	HELLO
2	1.5	YES
3	2	NO
4	2.3	I LOVE YOU
5		WHO ARE YOU
...		
2145656	21	HELLO DARKNESS MY OLD FRIEND
...		
369310405	28	IVE COME TO TALK WITH YOU AGAIN
...		

表 4 参照データ C2. Command として示す指示により出力文字列が出力される。出力 bit 数として空白を含む文字数 × 4.75bit を H2 として示す。

Command	出力	H2[bit]
2145656	HELLO DARKNESS MY OLD FRIEND	137.8
369310405	IVE COME TO TALK WITH YOU AGAIN	152.1
...	...	

ここで着目すべきは、C1, C2 いずれの場合も指示次第で任意の文字列を生成し得ることであると思われる。E1 の蓄積情報は確率的に均等に分布しているが指示 C1 の与え方次第でいかなる単語、文字列も出力することができる。C2 の場合確かに命令の数値的な大きさが小さいほど、有名で

よく知られたセンテンスが出力されるが、だからといってあまり使われないセンテンスもやがて表に出現するのでそのような文字列を出力する可能性がないわけではない。

E2 のみが寄与する場合とはどんな場合だろうか。DeepDream を使った創作例の中で C:参照データとして「ホワイトノイズ」を利用した場合がすでに示されている。そのようにホワイトノイズを参照画像として使った画像の生成例を図 5 に示す。

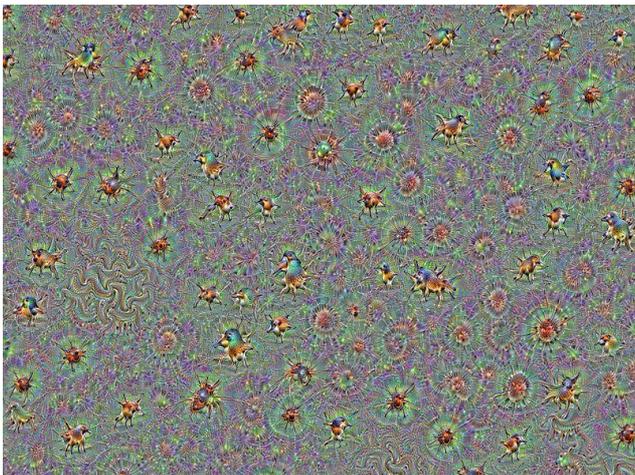


図 5 機械学習による作画例

(Deepdream-white-noise-0003.jpg,
https://commons.wikimedia.org/wiki/Category:Deep_Dream_images#/media/File:Deep-dream-white-noise-0003.jpg
 Created: 14 January 2016)

この場合、参照画像は単なるノイズであるから参照画像には創作物の知的財産への寄与はないと考えるべきだろう。したがってこの場合の創作物の価値はほぼ 100% 蓄積情報から得られたものと考えられる。

ここまでの考察をまとめると以下ようになる。

- (1) 機械学習による創作物には蓄積情報と参照情報が寄与する
- (2) 情報量が極度に小さい場合(たとえばゼロである場合)は寄与が少ないと考えられる

しかし、それでは情報量に比例して知的財産への寄与が決まるとして、単純に評価できるだろうか。以下では蓄積情報と参照情報の情報量によって創作物への寄与を評価する上で問題点について議論する。

3. AlgIM による評価

3.1 アルゴリズム的情報量(AlgIM)

議論の前にアルゴリズム的情報量(以下 AlgIM)の概念を導入する。アルゴリズム的情報量の提案はチューリングにさかのぼることができる。たとえば円周率 π の小数点以下の桁の連続は、一見ランダムなデータで無限の情報量を持つように見えるが、一定のアルゴリズムにより「計算可能」である点でランダムな数字の列とは異なる。チューリングはこのような情報をあらわす「計算可能数」という概念を提唱した。

G. J. Chaitin は、これをさらに発展させアルゴリズム的情報量という概念を提唱した。 π の桁の並びはほぼランダムであるが、 π の生成アルゴリズム自体はコンパクトである。したがって π を記録再生するには無限のメモリは必要ではなく、 π のアルゴリズムと計算領域があればよい。データの情報量を、そのデータを生成し得るアルゴリズムのサイズとして定義すれば、 π のような計算可能数の情報量は有限でありアルゴリズムに基づいた情報量の測度を構築できる。

3.2 AlgIM による知的財産評価の利点

π の桁はほぼランダムであることが経験的に知られている。したがって π を 27 進法で表現し、各桁にアルファベットと空白を割り当てれば、Simon & Garfunkel, “The Sound Of Silence” の歌詞や、シェイクスピアの小説も、いずれかの桁において出現すると考えられる。

そこで π を蓄積情報とし、蓄積情報の位置を指定することで出力を行う場合を考える。先の議論において蓄積情報から得られる追加の情報量を問題にしたが、蓄積情報から得られる正味の情報のみが重要であるという立場であれば、この場合知的財産はあくまで蓄積情報から派生したとみなせる。つまりこの場合、創作物への寄与は蓄積情報が大部分であり、参照情報の寄与は少ないということになる。

しかし蓄積情報の文字列が π というひとつの数値の一部であることを考えると、そのような解釈は不自然とも考えられる。もしこのような場合に蓄積情報が創作の主な貢献要因と考えると、すべての文学作品は π から生まれると感ずることになってしまう。

このような困難を避けるために AlgIM による情報量の評価を取り入れるべきだと考えられる。

蓄積された π の AlgIM はごくわずかである。 π のすべての桁が、表面的に多くの情報量を有しているとしても、AlgIM は少ない。したがってアルゴリズム的歌词の情報のうち π に起因するものは相対的に少なく、大部分の情報は「桁数指定」という指示によって与えられたと解釈できる。 π の中に無限のコンテンツが含まれているとしても、AlgIM としては π の計算アルゴリズムの長さに対応する情報量でしかないから、 π に帰属する知的財産権は少ないと評価することが可能になる。

π と同様に、表面的に多くの情報量を生成する手段は数多く存在する。そのように多量の「表面的」情報を生成することで多くの知的財産権を与えると有用な情報へのインセンティブにならない。より有用な情報を創出する作者や作業者にインセンティブを与える制度設計上、アルゴリズムの情報量を導入することが解決策になると考えられる。

3.3 AlgIM による知的財産評価の欠点

一方で、AlgIM による評価の欠点として、アルゴリズムの創作物としての価値がそのコードの長さだけで論ずることができるだろうか、という問題がある。

AlgIM はアルゴリズムとして最小の情報表現を意味するだけでその創作物としての価値の大きさとは関係がない。例としてまったく出鱈目な命令列からなる手順を考える。手順自体が多数の出鱈目なコードからなり、ただし停止したり出力が途絶えたりすることがないように、制約されて作られた長大なプログラムであるとする。このようなアルゴ

リズムと同じ出力を得るには全く同じアルゴリズムを実行する他にない。またこのアルゴリズムを実行すればアルゴリズム自身よりも多くの、見かけの情報量を持つ出力が生成される。このようにして「巨大な」AlgIMを有するデータを作ることができるが、このようなデータは乱数と同じではほぼ有用性がないだろう。

情報量は同じ素性の情報による創作への寄与量の評価に用いる(情報量が2倍であれば寄与も2倍に評価)することはできるが、異なる生成源から得る情報の価値の比較に用いるのは合理的ではない。

3.4 情報単価の算定

生成源の違いによる価値の違いを反映する方法についてはまだまだ議論の余地があるがその一つの方法として因子分析が考えられる。 $p_1 \dots p_N$ を生成されたN個の創作物の最終的な評価、AをM個の情報源からN個の創作物を生成する場合の各情報源から利用した情報量とすると、M個の情報源の単価を、 $r_1 \dots r_M$ とすれば両者の関係は式1のようにならわせるだろう。ここで $e_1 \dots e_N$ は近似誤差である。誤差最小となる $r_1 \dots r_M$ を求めることで各情報源の創作への寄与率を推定することができる。この問題は一種の推定問題であるので、様々な手法が知られているからそれらを利用して各情報源の成果への寄与を推定できるし、その結果寄与の大きい情報源の充実により多くのインセンティブを与えることは、結果として得られる創作物の発達にも寄与すると期待できる。

$$\begin{pmatrix} p_1 \\ \vdots \\ p_N \end{pmatrix} = A \begin{pmatrix} r_1 \\ \vdots \\ r_M \end{pmatrix} + \begin{pmatrix} e_1 \\ \vdots \\ e_N \end{pmatrix} \quad [\text{式 1}]$$

4. まとめ

本稿では機械学習を利用した創作において創作物への寄与を情報理論的測度の観点から考察した。

第一に、創作の参照モデルを用い形式的な議論の準備を行った。

第二に、創作モデルで交換される情報量を分析し、情報量が著作物の依拠性に対応する場合と、情報量のみでは創作物への寄与度の評価に問題が生じる場合を示した。

第三に、アルゴリズムの情報量(AlgIM)を導入し、第二の例の問題の一部が解消されるが、AlgIMでも異なる素性の素材の価値を公平に評価するには問題があることを示した。

最後に「付加価値」の評価を提案し、付加価値の定量的な評価が著作物に寄与する様々な素材の評価を与える可能性を論じた。

機械学習による創作は今後爆発的に普及する可能性があると考えられる。その際、創作に寄与した素材となる創作物と最終的な創作物の間には複製という関係がないので、素材となる創作物の供給のインセンティブをいかに高めるかが、知的財産権の制度設計において重要であると筆者は考える。機械学習を用いた創作はきわめて多くの構成要素からなり、仮にその派生物の知的財産権が高く評価されたとしてもその帰属は一つの要素によるのではなく、機械学習のトレーニングに用いた創作物や機械学習に与えるパラ

メータ、最終的な創作時に参照する参照データなど、多くの素材に依存する。それらの寄与度の評価をどうするかという点は現在議論が続いている点である。

しかし同様の議論は、印刷、録音録画が実現した際に起こった議論と同じ側面を持つ。すなわち、録音が実現する前はすべての演奏に演奏者が必要だったか、録音により単純な演奏の付加価値が下がる。その結果演奏者は、複製の元となる高度な品質を持つ演奏者。録音では対応できない、個別の条件に合わせたアドリブ演奏ができる演奏者。作曲家などのように分化していったのではなかろうか。

機械学習による創作の産業育成を考える場合、早期にその産業をささえるすべての利害関係者に正しいインセンティブが与えられることが重要であると考えられる。

本報告は一提案にすぎないが、ここで論じたような、価値ある創作物の供給が涵養される制度設計の議論が活発に行われ、機械学習を組み込んだ人間-機械学習共生システムによる創作活動が発展することを期待したい。

参考文献

- [1] Alexander Mordvintsev, DeepDream - a code example for visualizing Neural Networks, Google research blog, Posted July 1, 2015, <https://web.archive.org/web/20150708233542/http://googleresearch.blogspot.co.uk/2015/07/deepdream-code-example-for-visualizing.html> (2015)
- [2] WIKIMEDIA COMMONS, Category: Deep Dream Images, https://commons.wikimedia.org/wiki/Category:Deep_Dream_images edited on 15 May 2018, at 13:20.
- [3] 金子格, AI, AIの産業応用の拡大における知的財産の扱いに関する考察, 情報処理学会研究報告, 電子化知的財産・社会基盤(EIP), 2015-EIP-69(8), 1-4 (2015-09-03)
- [4] 金子格, 人工知能による著作物の創作性尺度に関するアルゴリズム情報理論から見た考察, 2017-EIP-76