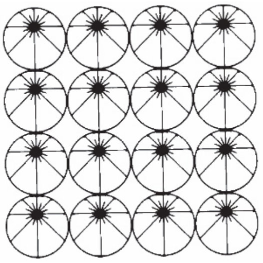
Recorded Lecture



At The Forefront of AI Technology − Current Frontiers

Graduate School of Information Sciences,
Tohoku University/RIKEN Center for Advanced Intelligence Project
Takayuki Okatani

This article is an edited version of an online lecture by Professor Okatani on December 16, 2021.

1 Introduction

I think the day has come when many people are looking at how they can use AI in their research. On the other hand, some of you will be most interested in what the latest AI looks like. Although we hear the term Digital Transformation (DX) more frequently than AI, today I am going to focus on AI.

Researchers in this field like me are surprised every day by developments. In the 10-year history of AI research, there have been very big changes (paradigm shifts) about once every 2 or so years, which I would like to share with you.

First, a simple self-introduction. My AI research field in imaging is called computer vision. This field has made major advances since 2010 due to the arrival of 'deep learning'.

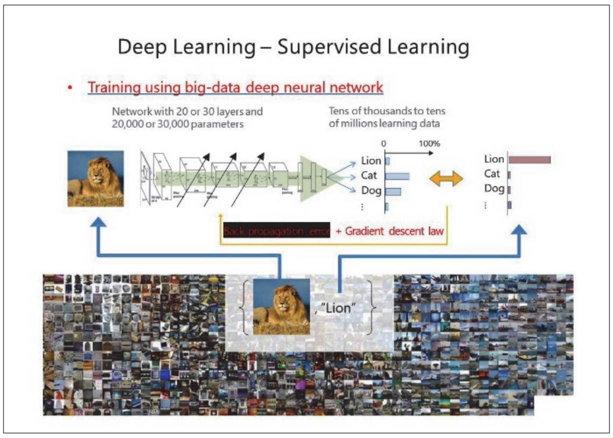
I have been researching computer vision since my student days but in the early days it was extremely rare to use machine learning for computer vision studies. Machine learning entered the field from about 2000 and the appearance of deep learning from 2010 seems to have accelerated developments.

The research has two 'wheels'–educating students and creating a new technology. I work collaboratively with industry as well as do consulting.

As a result, I do not just do scientific research, because I think scientific research and hands-on experience cannot be separated in this field.

2 AI – Deep Learning Success

People use the term 'AI' to include many things but most of the really interesting things for researchers are possible because of what we call "deep learning".



Deep Learning – Supervised Learning

- Training using big-data deep neural network

Network with 20 or 30 layers and 20,000 or 30,000 parameters

Tens of thousands to tens of millions learning data

Back propagation error + Gradient descent law

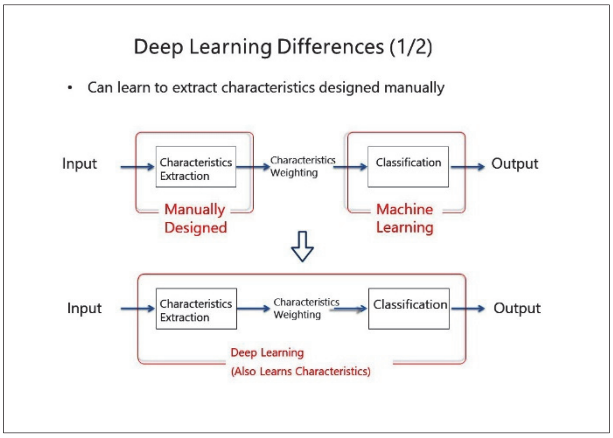
Slide 1

First, I will explain what deep learning is and how it is used.

Although there are various learning methods, basically we use a method called "supervised learning". I'll omit the details here, but in principle we arrange things called units into layers and then tie these units together by linking the layers. Information is transferred sequentially through the layers, so when something is input to a layer, something finally comes out. This type of layered network is called a "deep neural network".

As an example, when wanting to use object recognition to identify something in a photograph, first you need to create the appropriate neural network in which to put the image. It is necessary to predetermine first what type of object you want to ascertain by designing the network to output whether the input image more closely resembles a lion or a domestic cat. The aim is to obtain the response "resembles a lion" if we input a Merlion (Singapore mascot). In fact, since the early stages are unlikely to return this response, the wiring between levels is tuned weighting parameters to favor the correct response.

When modelled mathematically, linking 20 or 30 layers has a weighting order of several tens of millions. Today, there are much heavier weighting orders. The learning process tunes each parameter little-by-little to obtain the ideal output corresponding to the mathematical model input with massive degrees of freedom. Consequently, the requirements are the images of identifiable objects as well as correct information about the images, which means preparing at least several hundred images and ideally more than a thousand images for each thing to be distinguished. Deep learning means training to obtain the correct output at one-by-one input.



Deep Learning Differences (1/2)

- Can learn to extract characteristics designed manually

Input → Characteristics Extraction → Characteristics Weighting → Classification → Output

Manually Designed

Machine Learning

Input → Characteristics Extraction → Characteristics Weighting → Classification → Output

Deep Learning
(Also Learns Characteristics)

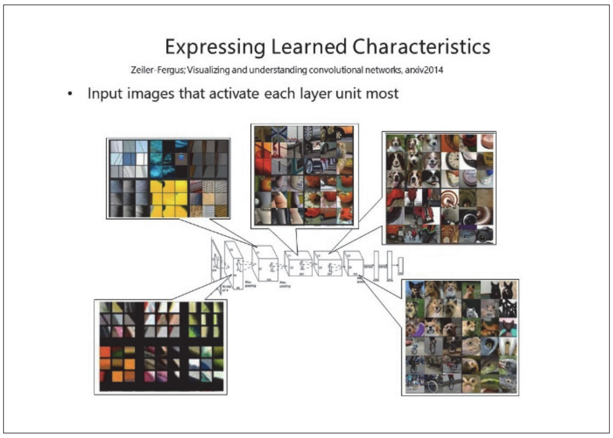
Slide 2

Next, I'll talk about the differences from earlier 'computer vision'.

Prior to deep learning, the attribute was obtained from the image. This was heavily compressed to minimize the amount of data. Doing this makes it possible to identify the thing you are looking for from the assorted characteristics. This has been the basis of machine learning for about the last 20 or 30 years.

The difficulty with this method is not knowing what the best characteristic is to take from the image. To identify a lion, when looking at an image of a lion, it is hard to answer the question, "What thing makes a lion a lion?" And it is even harder to code a program that does this, so the results didn't go well.

With the arrival of the deep-learning era, the neural network handles all problems up to the final output after inputting the image(s). Repeating the learning explained earlier solves the previously difficult issue of "Which are the best characteristics to extract?" because the neural network handles all learning.



Expressing Learned Characteristics

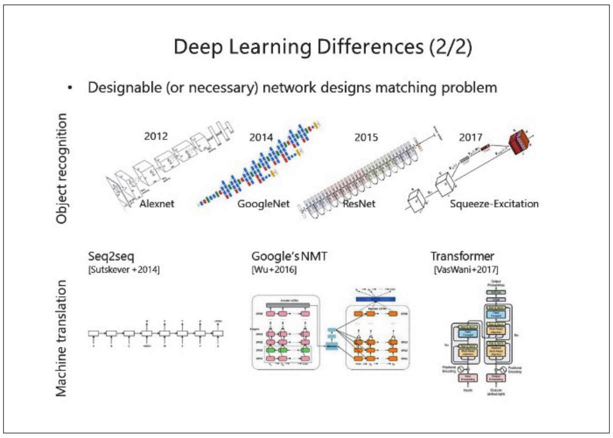
Zeiler-Fergus; Visualizing and understanding convolutional networks, arxiv2014

- Input images that activate each layer unit most

Slide 3

Here (slide 3) is a learning neural network with 1 million images showing each layer when the object is distinguished. It shows the response of some unit in some layer to the properties of the input image. Each unit corresponds to a nerve cell and is a 3 × 3 image array. The lower-layer unit reacts to diagonal line(s). The other upper-layer unit always reacts to yellow. Going slightly higher, the reaction is to orange circles and objects looking like train timetables. Going even higher, some units always react to dogs' faces while other units react to spiral shapes. This process comes about automatically through learning.

This might be called expression of hierarchical images and this type of structure may be found in a monkey or a cat or maybe sometimes even a person. Looking at various images while an electric probe is inserted into the brain, the results of investigating the reaction of nerve cells at the probe location when viewing some image clearly show that the brain's visual cortex handles how to output hierarchical information like that just shown here. However, we will not understand how to actually engineer this for a long time. In addition, when learning using a neural network described previously, we have discovered that this type of structure can be configured automatically.



Deep Learning Differences (2/2)

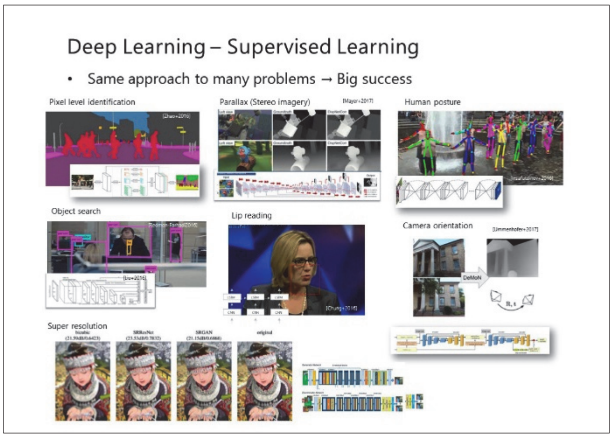
- Designable (or necessary) network designs matching problem

Slide 4

Listening to me, it may seem that everything can be left to the network, but this is not actually true. For the last 7 or 8 years I have been saying that it is very important for the network structure to match the problem, and so far, various structures have been developed for respective problems.

For example, there seems to be a very good structure for object recognition using images. In addition, despite being very simple, machine translation can now be used to output French from English input. In these circumstances, the design of the network structure is very important.

Although it is easy to talk about structural design, since the actual design must have very high degrees of freedom, we are now at the trial and error stage in gaining knowledge and experience. Progress is mainly by trial and error.



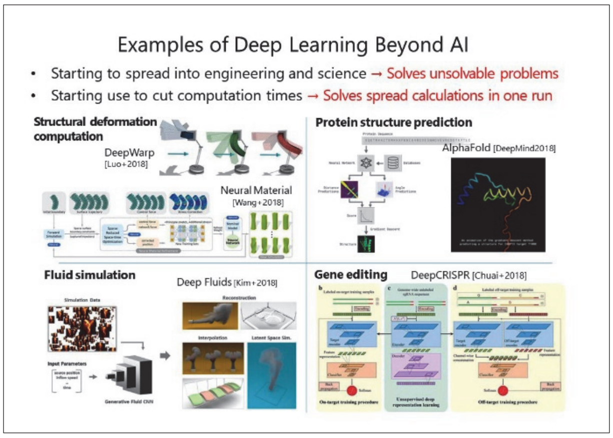
Deep Learning – Supervised Learning

- Same approach to many problems → Big success

Slide 5

Such a method is supervised learning. It learns what is the wanted output from whatever is input. Even when one image is input, there can be various outputs. To handle various problems using 'computer vision', objects are recognized at the pixel level, making it possible to paint different parts of the image.

Moreover, by inputting an image in which there is one object, instead of responding that it is something, deep learning can both search for that one object in images with various objects as well as use depth perception (parallax). Moreover, it can learn human posture in pictures. This has come to be a recent common application, but it is a technology from 5 or 6 years ago. It is a structure for outputting the coordinates of a person's joints in an image. The challenges of formulating a problem and designing a network structure in this way can be solved by inputting an image and learning the correct answer to the joint positions.



Examples of Deep Learning Beyond AI

- Starting to spread into engineering and science → Solves unsolvable problems
- Starting use to cut computation times → Solves spread calculations in one run

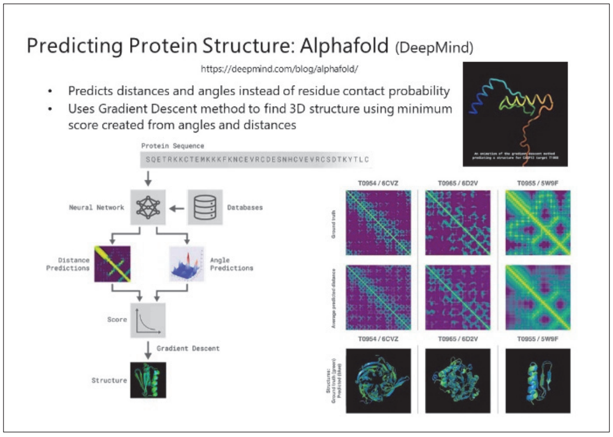
Slide 6

In our computer-vision research field, deep learning is being used for every problem. Almost nobody was using deep learning 10 years ago, but 99% of work is using it 10 years later.

In addition to AI applications, it is gradually entering every field of engineering and science and is already recognized as a general engineering tool beyond the limits of AI.

There are two types of use; first is solving previously difficult-to-solve problems, second is understanding physical phenomena that have been simulated by computer using replays as a substitute for simulation. Instead of simulation, the key is to establish the anticipated problem or causal relationship and let the network learn it. This can be used to simulate deformation of different objects and fluid dynamics.

One of the biggest impacts in resolving difficult problems has been in predicting protein folding using the AI system called AlphaFold.



Predicting Protein Structure: Alphafold (DeepMind)

https://deepmind.com/blog/alphafold/

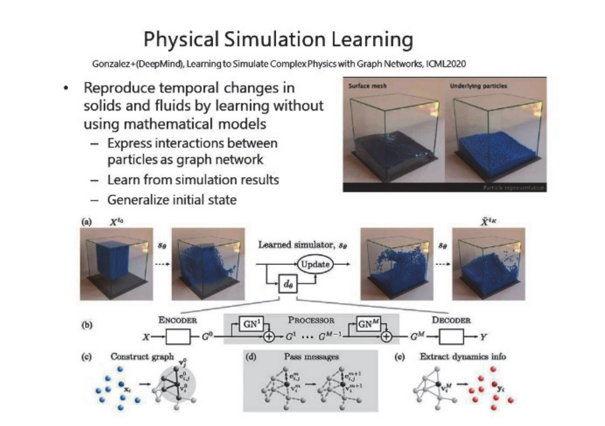
- Predicts distances and angles instead of residue contact probability
- Uses Gradient Descent method to find 3D structure using minimum score created from angles and distances

Slide 7

AlphaFold is published by the well-known DeepMind Technologies and was updated last year. Proteins are formed from sequences of amino acids starting with the complete amino-acid sequence as text data. The final 3D structure of the folded amino acid sequence determines the characteristic functions of the protein. More than 30 years have been spent considering ways to predict the final folded 3D structure from the first amino-acid sequence design.

Computer simulations based on physical models were run to conduct various tests and using deep learning just several years ago generated predictions with extremely high accuracy. Trial and error tests were run to see what was output for some input to the neural network by using the network to predict the final form resulting from changing the contact between bases, using further and closer distances, and different angles. The predicted 3D structures turned out well after having passed through this step.

With such high prediction accuracy, deep learning is solving problems that have remained unsolved for 30 years and is now progressing to more difficult problems.



Physical Simulation Learning

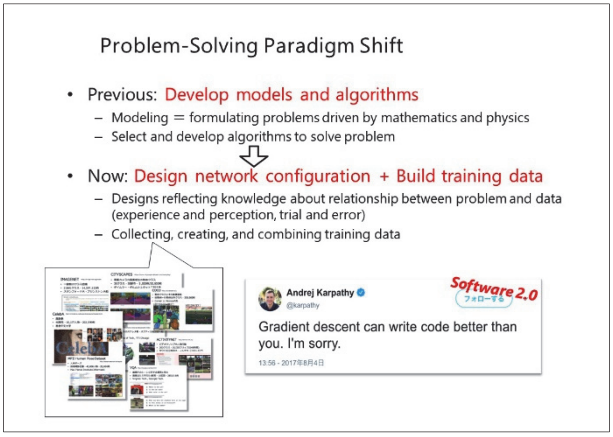
Gonzalez+(DeepMind), Learning to Simulate Complex Physics with Graph Networks, ICML2020

- Reproduce temporal changes in solids and fluids by learning without using mathematical models
  - Express interactions between particles as graph network
  - Learn from simulation results
  - Generalize initial state

Slide 8

Although I have talked about physical simulation as one use, here is another typical model. It is also by DeepMind Technologies and is used mainly to simulate fluids. This is the true calculated value based on the physical model. This is the result using a neural network to predict slightly ahead in the future only by learning, without considering the entire physical model. In concrete terms, it expresses the target medium as a combination of many particles each of which is moving freely. When one particle strikes a neighbor, it moves under some constraints, one of which is obviously gravity.

Conventionally, the system is mathematically and physically modelled based on neighboring relationships and physical laws and then the sequence direction is calculated based on this model; in other words, a simulation is performed. However, the deep-learning method uses only the interactions between adjacent particles to construct a graph of the interaction range showing the current state and speed, etc., of each particle to predict interactions with surrounding particles several ms in time ahead of the current condition. Using this prediction as learning data and running one simulation passes the movement of each particle like this to obtain the next data and predict the future movement by repeating the simulation over again.



Problem-Solving Paradigm Shift

- Previous: Develop models and algorithms
  - Modeling = formulating problems driven by mathematics and physics
  - Select and develop algorithms to solve problem
- Now: Design network configuration + Build training data
  - Designs reflecting knowledge about relationship between problem and data (experience and perception, trial and error)
  - Collecting, creating, and combining training data

Andrej Karpathy
@karpathy
Gradient descent can write code better than you. I'm sorry.
13:56 - 2017年8月4日

Slide 9

Therefore, there have been great changes in problem-solving methods outside AI. Our field in particular, has also changed completely. Conventionally, models were important and were created making full use of both mathematics and physics. In the case of computer vision, the key target of the model was how to handle the image in the first place. An image of an object is obtained by capturing light reflected from the object surface at an image sensor after passage through the camera lens. A physical model is created of this process, which is then used to solve the reverse problem, but there was a problem regarding what appeared with this development so the method did not go well. Since there was very little AI until about 10 years ago, despite this approach being the mainstream trend, progress did not go very well.

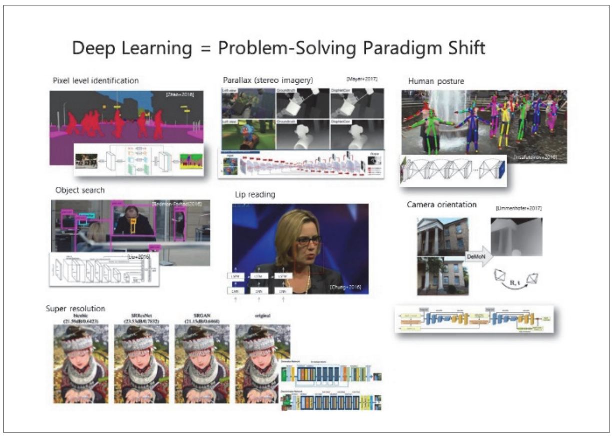
In contrast, deep learning is now being used as the problem-solving method. I touched on this already a little when saying what I did, but the first point is designing the neural network structure. The questions are "What problem are you trying to solve?" and "What are the inputs and outputs?" These questions must be decided first. First, you must consider the form of the inputs and outputs to be solved by deep learning. This is already a difficult task but must be included in considering how to create the internal network structure, and then large amounts of correct input and output pairs must be collected. If the network design is for object recognition, these will be collections of large numbers of images and objects appearing in images.

Therefore, although mathematics and physics have been extremely important, since these models are not used in the deep-learning world, although it may be a slight exaggeration, mathematics and physics are no longer essential. The hard part is thinking about the network structure. However, since there is no well-established methodology or theory for this structural design, the only way is to use various past experiences combined with trial and error. In many cases, mathematics and physics have nothing to do with the method for collecting training data using input and correct output. So now let's talk about this methodology.

Some people call this new paradigm Software 2.0. Gradient descent is a neural-network learning algorithm that can write better code than a programmer. Rather than actually writing a program, it can be likened to learning actions accurately and writing them. Conventionally, problems were solved by coding actions by considering a difficult model-based algorithm, but things are different now when problems are solved by various programs learning the network structure.

3 Deep Learning So Far

This approach has been successful for about 3 or 4 years. However, there have been various obstacles in solving some problems.

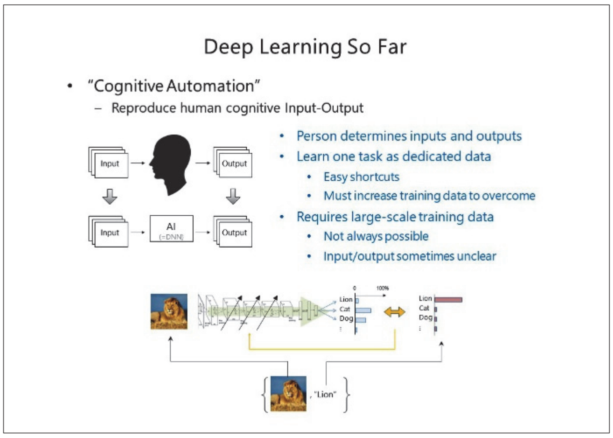


Deep Learning = Problem-Solving Paradigm Shift

Pixel level identification
[Zhao+2016]
Parallax (stereo imagery)
[Mayer+2017]
Human posture
[Insafutdinov+2016]
Object search
[Redmon-Farhadi2016]
[Liu+2016]
Lip reading
[Chung+2016]
Camera orientation
[Ummenhofer+2017]
DeMoN
R, t
Super resolution
bicubic
(21.59dB/0.6423)
SRResNet
(23.53dB/0.7832)
SRGAN
(21.15dB/0.6868)
original

Slide 10

Although this slide is the same as the previous slide on deep learning that can do various things, the scenery in our computer-vision world has changed almost completely compared to 20 or even 10 years ago.



Deep Learning So Far

- "Cognitive Automation"
  - Reproduce human cognitive Input-Output
- Person determines inputs and outputs
- Learn one task as dedicated data
  - Easy shortcuts
  - Must increase training data to overcome
- Requires large-scale training data
  - Not always possible
  - Input/output sometimes unclear

Input
Output
Input
AI
(=DNN)
Output
0
100%
Lion
Cat
Dog
Lion
Cat
Dog
"Lion"

Slide 11

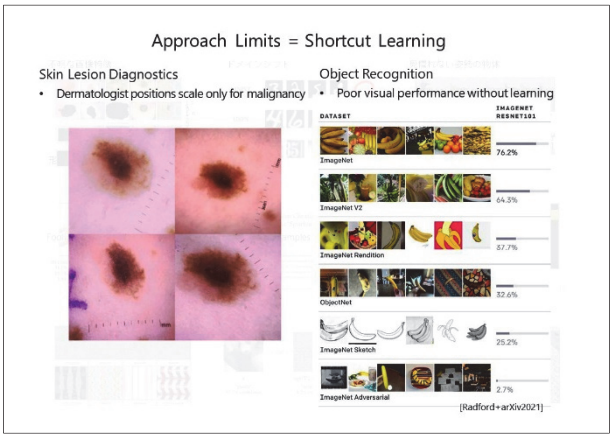
Thinking carefully, what I do in AI is not intelligence at all despite calling it AI. Therefore, I call it "cognitive automation", meaning automating the human cognitive process.

When talking about this kind of thing, human beings can do the various tasks shown in slide 10. First, consider what is input to people and what is output? For image recognition, the input is an image while the output might be a type of object or the position such as the screen coordinates of the body in the image. By specifying these inputs and outputs, perhaps it is possible to create a neural network that faithfully reproduces the human output from an input. What to input and output is decided by researchers using deep learning, but nobody has so far managed to reproduce these human functions. Consequently, this is why I think it is better to describe this as cognitive automation rather than intelligence.

However, if some human functions can be done by machines, there are not only huge benefits but the idea to isolate the bad aspects. But there will still be some problems even if bad aspects can be isolated. Anyway, these problems cannot be solved without having large-scale training data although sometimes training data is not always available.

Even worse, often the inputs and outputs are unclear, and solving problems without a clear definition of what to input and what is output needs many people.

There are other remaining issues too, such as specifying the inputs and outputs for the problem to solve, designing an exclusive network, and learning exclusive data, and there are clear problems with this general method.


Approach Limits = Shortcut Learning

Skin Lesion Diagnostics

- Dermatologist positions scale only for malignancy

Object Recognition

- Poor visual performance without learning

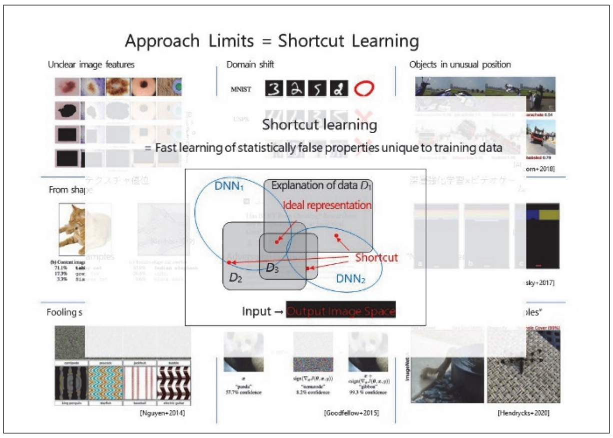
DATASET	IMAGENET RESNET101
ImageNet	76.2%
ImageNet V2	64.3%
ImageNet Rendition	37.7%
ObjectNet	32.6%
ImageNet Sketch	25.2%
ImageNet Adversarial	2.7%

[Radford+arXiv2021]

Slide 12

It turns out that there is a so-called ‘shortcut’ phenomenon that I will explain. This example is a slightly funny model that is like a story about skin lesions. The model outputs a diagnostic result about whether a lesion is benign like a common mole or malignant like a melanoma. Various training data were collected and used to correct respective benign and malignancy diagnostic results on a learning network. However, the dermatologist placed a scale to measure size only near training images of malignancies. When the learning network was trained using these data, it mistakenly learned to assess whether the lesion was benign or malignant based on whether a scale was present, which was meaningless. As a result, malignant lesions without a scale were not evaluated as malignant which was a false-negative result. This example demonstrates shortcut learning. This example is easy to understand but there are much harder-to-understand examples. When only training pictures of malignancies are taken in a slightly different way, the neural network soon learns that this difference is an indicator of malignancy.

Here (Slide12) is another example using photographs (top row) to recognize a banana as a banana with very high accuracy using network learning. However, bananas have various appearances, such as conceptual art and these types of design, so the network could not immediately recognize some bananas without learning images.


Approach Limits = Shortcut Learning

Unclear image features

Domain shift

Objects in unusual position

Shortcut learning

= Fast learning of statistically false properties unique to training data

[Nguyen+2014]

[Goodfellow+2015]

[Hendrycks+2020]

Slide 13

There are plenty of such examples. From the handwritten figures here (slide 13) in the top line, we can confirm the extremely high accuracy for the similar data range. However, although the number is somehow blurry, it is different at a dimmer place for slightly different data and this type of data does not transfer so well. When a neural network can definitely recognize the top numbers, people might think “The neural net must be understanding concepts such as the form of the numbers, otherwise numbers are not recognizable”. However, this is an example of where what a neural network ‘sees’ is slightly different from what we expect.

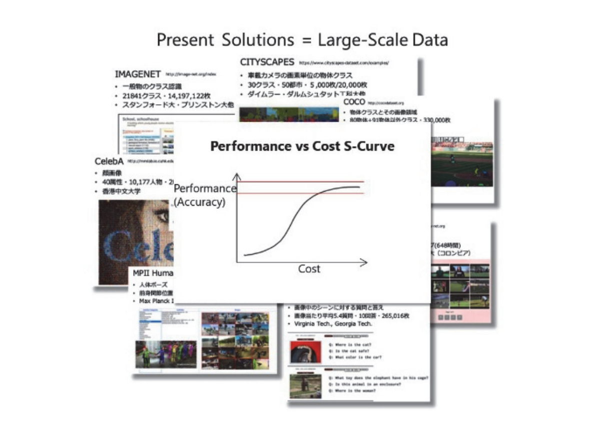
Similar examples can be found in many places. The principle is similar even for objects in unusual positions that seem completely different (slide 13).

Adversarial examples are also well known. When training properly with this noise-free panda image, a neural network can make serious errors, such as this misrecognition of a panda as a gibbon.

Like the top domain-shift example, what the network ‘sees’ is not like what we are thinking. This is similar to the example of shortcut learning where the presence of a scale in pictures of dermatology lesions caused misidentification (slide 12). In other words, learning mistakes can occur if there are false distinguishing points hidden in the training data because neural networks have extremely high ability from this perspective.

Moreover, there are limits to the size of available training data. If this (slide 13) represents the input-to-output space, each point expresses various representations from input to output. When deciding the neural-network configuration, various representations can be expressed by adding a large weighting parameter. For example, if there is an ideal representation for a problem to solve here at hand, data is provided so the network can acquire this representation. As I have said, although the data size is limited, there seems–somehow or other–to be a representation explaining the data in the represented range. Although just the thing is identified in these examples (slide 12, 13), people want to believe that 'a number is not recognizable if an object is never recognized when not seen in the same way as a human being'. However, neural nets are not like this, and we need to understand how human beings and neural nets distinguish things.

For this reason, like this shortcut example, AI finds a thing that we do not really want. In actual use, recognition errors occur immediately after inputting prepared training-data images with slightly different statistical properties.



Present Solutions = Large-Scale Data

IMAGENET http://image-net.org/index
• 一般物のクラス認識
• 21841クラス・14,197,122枚
• スタンフォード大・プリンストン大他
CITYSCAPES https://www.cityscapes-dataset.com/examples/
• 車載カメラの画素単位の物体クラス
• 30クラス・50都市・5,000枚/20,000枚
• ダイムラー・ダルムシュタット工科大他
COCO http://cocodataset.org
• 物体クラスとその画像領域
• 80物体+91物体以外クラス・330,000枚
CelebA
• 顔画像
• 40属性・10,177人物
• 香港中文大学
MPII Huma
• 人体ポーズ
• Max Planck I
• 画像中のシーンに対する質問と答え
• 画像当たり平均5.4質問・10回答・265,016枚
• Virginia Tech., Georgia Tech.
(648時間)
(コロンビア)
Performance vs Cost S-Curve
Performance (Accuracy)
Cost

Slide 14

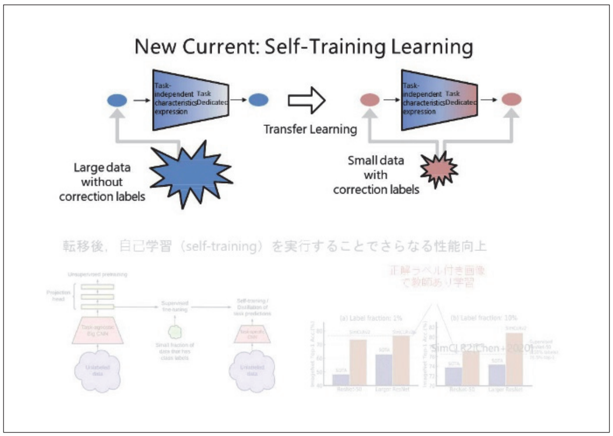
The only certain method to solve this problem is to use large-scale data, and research is progressing in this direction.

Data collection has a basic cost and slide 14 shows how performance rises as costs increase with more data. However, the rising performance curve reaches a plateau at some point. Rarely does performance keep rising as data is collected. In other words, although performance keeps rising by collecting data up to a plateau, very high costs are incurred to reach this region.

Therefore, although I am talking about 3 or 4 years ago, the many AI startup businesses in the world doing things in this domain have been able to build their networks this far. However, it will be difficult for these service applications to achieve the required real-word performance due to these rising costs. I think there are already many commercial applications as a result of overcoming these problems, but there could be solutions in areas like medicine and automotives requiring high reliability levels to protect life where high cost is not a consideration. However, it might be quite difficult for some businesses to use deep-learning AI.

4 Current and Future Deep Learning

Let's talk about what is happening now.



New Current: Self-Training Learning

Task-independent characteristics expression
Task Dedicated
Transfer Learning
Large data without correction labels
Small data with correction labels
転移後，自己学習（self-training）を実行することでさらなる性能向上
正解ラベル付き画像で教師あり学習
Unsupervised pretraining
Projection head
Task-agnostic Big CNN
Unlabeled data
Supervised fine-tuning
Small fraction of data that has class labels
Self-training / Distillation of task predictions
Task-specific CNN
(a) Label fraction: 1%
(b) Label fraction: 10%
ImageNet Top-1 Acc.(%)
SOTA
SimCLRv2
ResNet-50
Larger ResNet
SimCLR2[Chen+2020]

Slide 15

There have been big changes in about the last 2 years based on critical understanding of this type of problem. This could be described as a new current and the method called self-supervised learning is now an extremely big topic.

There is a mountain of data on the web such as image and text data. The problem is the task of finding data providing a solution to the problem. Believing that "I'll get the correct answer if I input these images." costs money to get the correct answer.

However, this self-supervised learning is not necessary to give the correct answer. As explained later, creating answers automatically is a problem. From the usage side, an image can be learned only by bringing in many images. So-called transfer learning is the major form for this neural network and learning uses input and output for the problem task one really wants to solve. 'Fine tuning' means preparatory learning using these large data and is an approach to using fewer data on just several upper layers of the neural network. However, some minimum amount of correct data is still necessary.

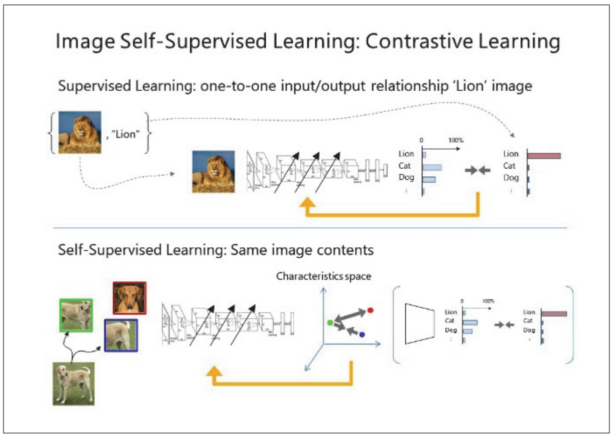


Image Self-Supervised Learning: Contrastive Learning

Supervised Learning: one-to-one input/output relationship 'Lion' image

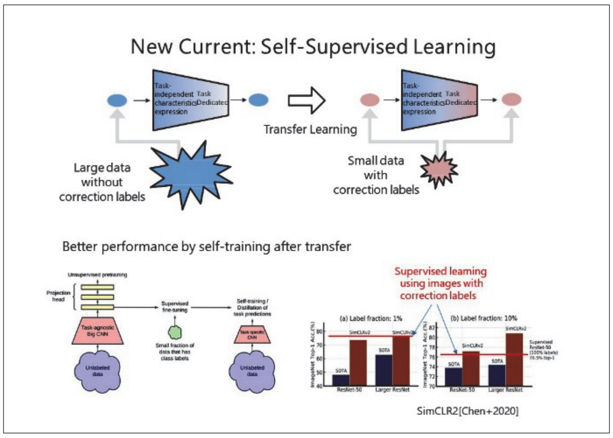
Self-Supervised Learning: Same image contents

Slide 16

In more concrete terms, as explained, supervised learning with images conventionally gives the answer 'lion' when inputting this lion image. Therefore, such an image is required for correct training of the network.

In contrast, although self-supervised learning collects many images, it does not need correction like "What is appearing there?" When saying what to do, image parts taken from slightly different parts are used. The various image cuttings are loaded into the network for handling as intermediate parts called 'characteristics vectors' by the network here (slide16). In this characteristics space (slide 16), the network is being trained under pressures such as 'The closest similarity for things extracted from the same image is near' and 'The largest dissimilarity for things extracted from the same image is far'.

The aim is to map features extracted from the image but from slightly different locations into the same characteristics space. Features are mapped to separate locations for different images. This is self-supervised learning of an image. After such learning, providing a few images of the object to be recognized and some images with correction labels supports learning of the more upstream part from this characteristics space. This method started to produce very good results from about 2 years ago.



New Current: Self-Supervised Learning

Task-independent characteristics expression
Task Dedicated
Transfer Learning
Task-independent characteristics expression
Task Dedicated
Large data without correction labels
Small data with correction labels

Better performance by self-training after transfer

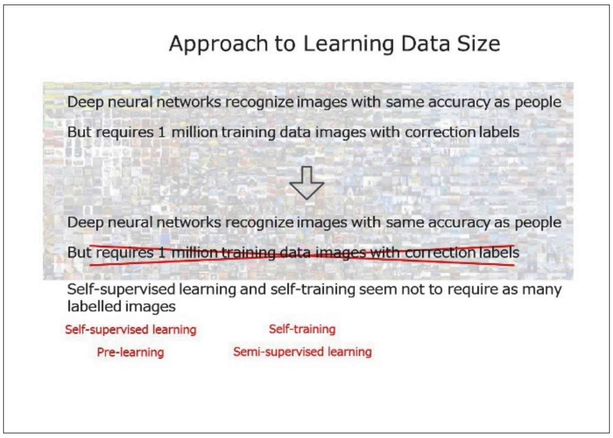
Unsupervised pretraining
Projection head
Task-agnostic Big CNN
Unlabeled data
Supervised fine-tuning
Small fraction of data that has class labels
Self-training / Distillation of task predictions
Task-specific CNN
Unlabeled data
Supervised leaming using images with correction labels
(a) Label fraction: 1%
(b) Label fraction: 10%
ImageNet Top-1 Acc.(%)
80
70
60
50
40
SimCLRv2
SOTA
ResNet-50
Larger ResNet
82
80
78
76
74
72
70
Supervised ResNet-50 (100% labels) 76.5% top-1

SimCLR2[Chen+2020]

Slide 17

In addition to self-supervised learning, there is also a method called self-training. For example, after image recognition with a small number of images and correction data, here (left side of slide17) we put the large-scale data without correction labels used once previously once more into the same networks. By doing so, although the network outputs something, that output is used as some form of self-training correction.

In other words, self-learning is the second stage of self-supervised learning. Although not the same, using self-predictions as a form of correction data supports self-improvement, so performance is clearly improved by learning with correction labels right from the start.



Approach to Learning Data Size

Deep neural networks recognize images with same accuracy as people

But requires 1 million training data images with correction labels

Deep neural networks recognize images with same accuracy as people

But ~~requires 1 million training data images with correction labels~~

Self-supervised learning and self-training seem not to require as many labelled images

Self-supervised learning

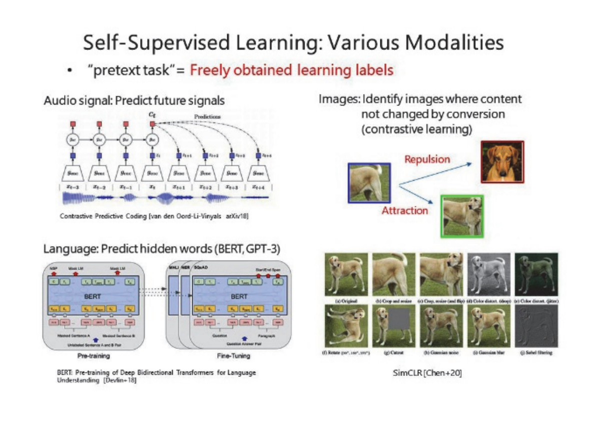
Self-training

Pre-learning

Semi-supervised learning

Slide 18

Therefore, although conventional deep learning can recognize objects with the same performance as people, it is quite difficulty because it requires 1 million images with a correction label attached to each image. In contrast self-supervised learning does not use 1 million images with attached labels. Although 1 million or more unlabeled images are necessary, any number can be obtained from the internet using a web crawler. On the other hand, correction labels are not so necessary, and the same or better accuracy degree can be obtained if just 1% of the 1 million images, or 10,000 images, have correction labels attached.



Self-Supervised Learning: Various Modalities

- "pretext task"= Freely obtained learning labels

Audio signal: Predict future signals

Contrastive Predictive Coding [van den Oord-Li-Vinyals arXiv18]

Images: Identify images where content not changed by conversion (contrastive learning)

Language: Predict hidden words (BERT, GPT-3)

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [Devlin+18]

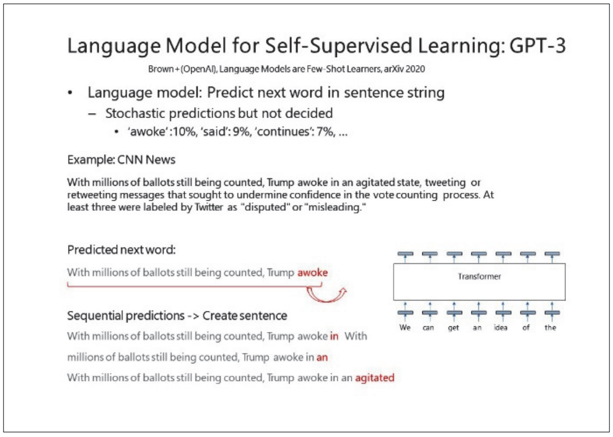
SimCLR [Chen+20]

Slide 19

The same approach is being adopted for other AI problems and not just for images. As I said, after imaging, the point is language accuracy.

Similar methods are used for both audio and language. Regarding what to do, some signals are input to the neural network at some point in time while others are not; in other words, we are thinking about future expected problems. Although self-supervised learning is finally supervised learning, the supervisor and correction labels are created automatically freely. These are called the pretext tasks which the neural network uses for learning with the idea of predicting the future audio-signal input midway. This can be done freely if there is one audio signal.

Language is similar to audio; a proper sentence is input word-by-word to the network. The problem of hiding one word at random in a sentence and predicting the hidden word can be solved by training. Part of the sentence in slide 20 is hidden but the correction data can easily fill the sentence gaps.



Language Model for Self-Supervised Learning: GPT-3

Brown+(OpenAI), Language Models are Few-Shot Learners, arXiv 2020

- Language model: Predict next word in sentence string
  - Stochastic predictions but not decided
    - 'awoke':10%, 'said': 9%, 'continues': 7%, ...

Example: CNN News

With millions of ballots still being counted, Trump awoke in an agitated state, tweeting or retweeting messages that sought to undermine confidence in the vote counting process. At least three were labeled by Twitter as "disputed" or "misleading."

Predicted next word:

With millions of ballots still being counted, Trump awoke

Sequential predictions -> Create sentence

With millions of ballots still being counted, Trump awoke in With

millions of ballots still being counted, Trump awoke in an

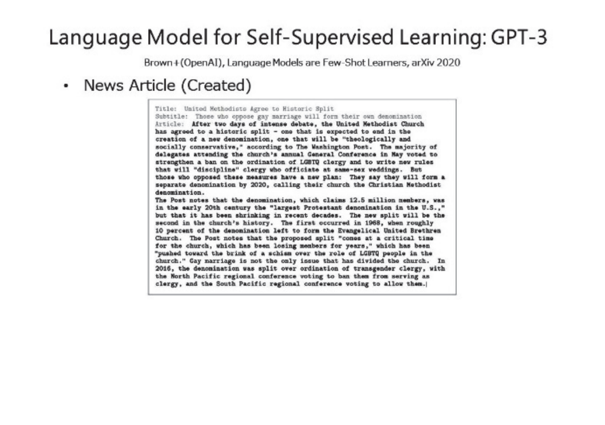
With millions of ballots still being counted, Trump awoke in an agitated

Slide 20

Even if we say this is the strongest impact of self-supervised learning, the biggest change to AI itself is called the "language model".

The most well-known is OpenAI's GPT-3 model. Despite being self-supervised learning, the language model was created in the first place to do self-supervised learning and is in fact an old thing. When explaining what it does, like the previously explained audio data, it uses pretext task network training to predict the next word in a sentence string.

As explained previously, there is a mountain of sentences on the web, such as this CNN article on something ex-president Trump did. When inputting the part of this article up to the word 'Trump' into the network, the model predicted the next word to be 'awoke'. Of course, the network is using stochastic prediction, so this is not the only possible correct answer. Although 'awoke' was the correct answer in this example, the next word in the sentence can be predicted again by inputting that answer to the network as shown in slide 20, and after reaching some midpoint, the full sentence is predicted accurately word-by-word.



Language Model for Self-Supervised Learning: GPT-3

Brown+(OpenAI), Language Models are Few-Shot Learners, arXiv 2020

- News Article (Created)

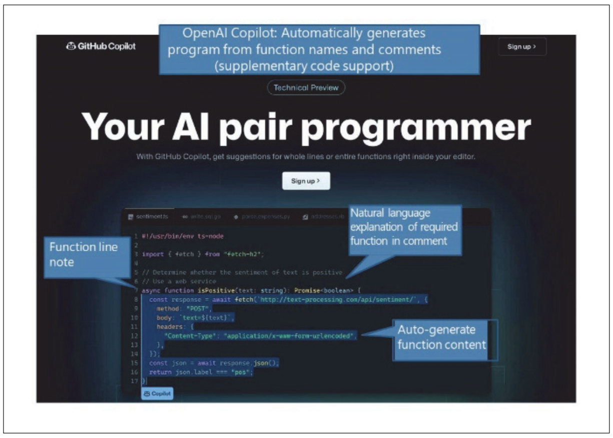
Title: United Methodists Agree to Historic Split
Subtitle: Those who oppose gay marriage will form their own denomination
Article: After two days of intense debate, the United Methodist Church has agreed to a historic split - one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church's annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination.
The Post notes that the denomination, which claims 12.5 million members, was in the early 20th century the "largest Protestant denomination in the U.S.," but that it has been shrinking in recent decades. The new split will be the second in the church's history. The first occurred in 1968, when roughly 10 percent of the denomination left to form the Evangelical United Brethren Church. The Post notes that the proposed split "comes at a critical time for the church, which has been losing members for years," which has been "pushed toward the brink of a schism over the role of LGBTQ people in the church." Gay marriage is not the only issue that has divided the church. In 2016, the denomination was split over ordination of transgender clergy, with the North Pacific regional conference voting to ban them from serving as clergy, and the South Pacific regional conference voting to allow them.

Slide 21

GPT-3 is a massive network with about 4 trillion data items managed by 133 billion parameters and the cost of the learning computers is on the order or about ¥500 million. Such a thing can support huge learning when allowed.

Slide 21 shows another example where a news article was generated by inputting this single grayed-out text group to the network. The news Title and Subtitle run from here, and after inputting the word 'Article', the learning network predicted 'After' as the next word and then fed the word 'After' back to the network to predict the next word 'two' and so on to predict 'days' and the rest of this entire article.

Since I did not get the point of the article in English, I used a neural network to perform a machine translation into Japanese, but the text is not easy to understand although it feels like a high-quality article. Although I can't say whether the content is all lies or is true, I think some parts of previous articles have been pasted in. However, it is an example of AI generating quite a plausible article.



OpenAI Copilot: Automatically generates program from function names and comments (supplementary code support)

Your AI pair programmer

With GitHub Copilot, get suggestions for whole lines or entire functions right inside your editor.

Function line note

Natural language explanation of required function in comment

Auto-generate function content

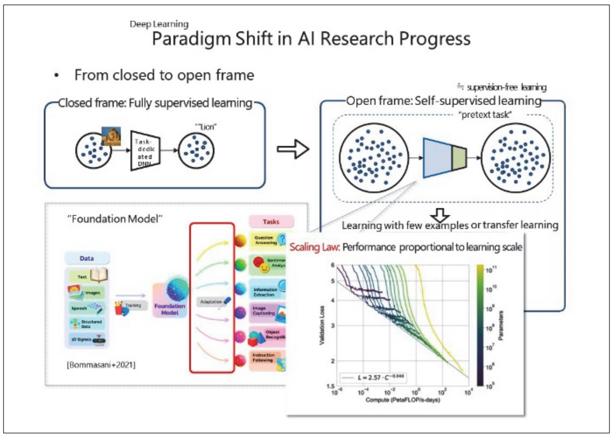
```
#!/usr/bin/env ts-node

import { fetch } from "fetch-h2";

// Determine whether the sentiment of text is positive
// Use a web service
async function isPositive(text: string): Promise<boolean> {
  const response = await fetch(`http://text-processing.com/api/sentiment/`, {
    method: "POST",
    body: `text=${text}`,
    headers: {
      "Content-Type": "application/x-www-form-urlencoded",
    },
  });
  const json = await response.json();
  return json.label === "pos";
}
```

Slide 22

Doing such things is quite nice! Slide 22 shows an easy-to-understand example of a tool announced in 2021 for auto-generating programs like Copilot and Codex. Inputting part of a program generates supplementary code and is functionally similar to the previously described one-by-one word prediction networks. It automatically generates natural-language comments one-by-one about the actions of previously coded functions. Letting previous 'language models' learn various programs supports predictions where 'the content will be like this if such a thing is wanted'.



Deep Learning

Paradigm Shift in AI Research Progress

- From closed to open frame

Closed frame: Fully supervised learning

Open frame: Self-supervised learning

≒ supervision-free learning

"pretext task"

Learning with few examples or transfer learning

"Foundation Model"

[Bommasani+2021]

Scaling Law: Performance proportional to learning scale

Slide 23

Slide 23 is a summary. With self-supervised learning as the keyword, there is now a paradigm shift in AI learning methods and the deep-learning world.

Although called "fully-supervised learning" some years ago, the approach used input and output designed according to the problem to be solved along with training data and a dedicated network.

However, this has now changed, first, to using self-supervised learning with massive data. Since correction labels were unnecessary, a lot could still be learned by collecting large amounts of equivalent data. After doing this first, it was possible to move to the problem to be solved. Although called transfer learning or 'few-shot learning', since the majority of networks is end by them, it is better to move to a method that tunes only the highest-layer parameters for the task to be solved, which reduces the amount of data, especially at this part (task-independent characteristics expression). After completing training, some recent networks supporting various later problems are described as using the Foundation Model. Conventionally, training was performed using this type of dedicated (task dedicated DNN), but I would recommend using the Foundation Model that supports a variety of problems.

The problem is that since this Foundation Model itself uses massive data and computation resources, it cannot be used by general people. Since it requires money, computing time, and costs, we university researchers cannot enter the battle in this field. Therefore, we are considering research using this partial approach (Adaptation) as a means to handling problems we want to solve.

Although the language model is a key part of this, the learning scale of this GPT-3 model itself has increased in the last 18 months. A slightly frightening article was published at about the same time about the Scaling Law which says that performance increases as more money is spent.

The accuracy of next-word prediction is a function of three factors: the data used for learning, the network size, and the time actually spent on learning, and performance rises as these factors increase in balance. Although performance continues to increase, there are comments that people do not understand the upper limit at the moment. I will show some of the latest articles later.

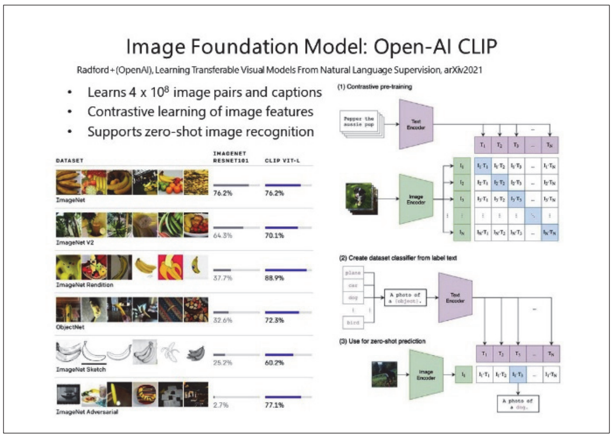


Image Foundation Model: Open-AI CLIP

Radford+(OpenAI), Learning Transferable Visual Models From Natural Language Supervision, arXiv2021

- Learns 4 x 10^8 image pairs and captions
- Contrastive learning of image features
- Supports zero-shot image recognition

DATASET	IMAGENET RESNET101	CLIP VIT-L
ImageNet	76.2%	76.2%
ImageNet V2	64.3%	70.1%
ImageNet Rendition	37.7%	88.9%
ObjectNet	32.6%	72.3%
ImageNet Sketch	25.2%	60.2%
ImageNet Adversarial	2.7%	77.1%

(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
Image Encoder
T1 T2 T3 ... TN
I1 I2 I3 ... IN
I1·T1 I1·T2 I1·T3 ... I1·TN
I2·T1 I2·T2 I2·T3 ... I2·TN
I3·T1 I3·T2 I3·T3 ... I3·TN
IN·T1 IN·T2 IN·T3 ... IN·TN
(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
(3) Use for zero-shot prediction
Image Encoder
I1
I1·T1 I1·T2 I1·T3 ... I1·TN
A photo of a dog.

Slide 24

Open-AI has also released a Foundation Model not just for language but also for images and especially when there is a relation to an image and language.

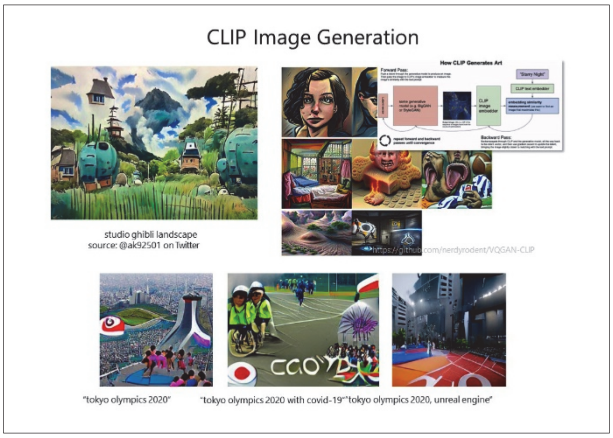
There has been talk about how the data is obtained but there are about 400 million pairs of image-to-image captions and the method is unknown but probably comes from mining SNS messages, etc. These data are used to train two networks: one for converting a caption to a characteristic's vector and the other for converting an image to a vector for a different characteristic.

These two networks perform training by applying two pressures: converting the image and explanation becoming a pair to as-close-as-possible vectors, and converting an unrelated image and explanation to far-apart vectors. Using the large-scale 400 million data pairs can do amazing things, such as recognition by zero-shot image recognition.

Until now, object recognition (slide 23) has been by one-to-one teaching by pairing the image input with the correct answers "This is a lion; This is a banana; This is a panda." However, this Contrastive pre-training method does not say which is a panda, lion or banana at all. Simply pairing captions only with contrastive images can automatically recognize a banana and a lion.

The image of the item to recognize is input here (slide 24). For example, various candidate prompts questions, such as pictures of an airplane, automobile, dog, and bird are prepared. Then the vector of the image to be recognized is compared to the vector of some close image. For example, if a dog appears, what is supposed to be the dog image becoms the closest vector to the dog image, meaning it can be judged as like a dog. Therefore, it is recognized properly without teaching "This is a lion; this is a banana; this is a dog."

Previously, it was impossible to recognize a thing without training the image, but with this method, it is even possible to recognize a sketch of a banana as well as hard-to-understand banana conceptual art as a banana.



CLIP Image Generation

studio ghibli landscape
source: @ak92501 on Twitter

https://github.com/nerdyrodent/VQGAN-CLIP

"tokyo olympics 2020" "tokyo olympics 2020 with covid-19" "tokyo olympics 2020, unreal engine"

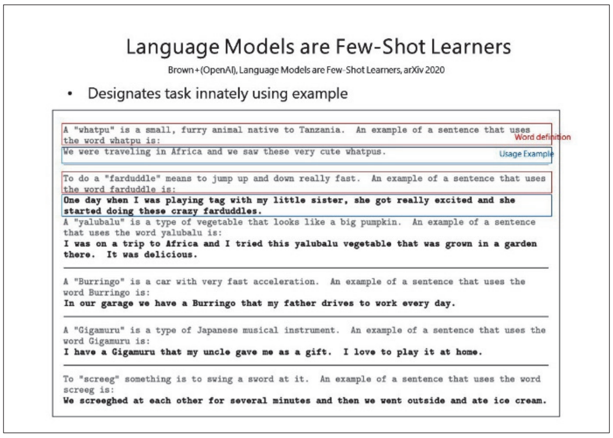
How CLIP Generates Art
Forward Pass:
some generative model (e.g. BigGAN or StyleGAN)
CLIP image embedder
"Starry Night"
CLIP text embedder
embedding similarity measurement
repeat forward and backward passes until convergence
Backward Pass:

Slide 25

Slide 25 shows the image Foundation Model called CLIP. CLIP creates interesting, combined images and I have been playing with it for about 1 year. In brief, CLIP learns the corresponding relationship of things reflected in images very well. For example, this (upper right of slide 25) is an example of an auto-generated image with a close correlation to text saying, 'like Studio Ghibli'.

In concrete terms, when some image is created by inputting random images to the network, the correspondence of the explanatory test with the created image, or in other words the vector made when the image is created, is associated with the closest image by inverse operation. Somehow, CLIP seems to understand the Studio Ghibli feeling and can generate various images using various texts using this 'feeling'!

This is something I have been playing with for 6 months and created these images for the 2020 Tokyo Olympics; I think they have just the right feeling.



Language Models are Few-Shot Learners

Brown+(OpenAI), Language Models are Few-Shot Learners, arXiv 2020

- Designates task innately using example

A "whatpu" is a small, furry animal native to Tanzania. An example of a sentence that uses the word whatpu is:
We were traveling in Africa and we saw these very cute whatpus.

To do a "farduddle" means to jump up and down really fast. An example of a sentence that uses the word farduddle is:
One day when I was playing tag with my little sister, she got really excited and she started doing these crazy farduddles.

A "yalubalu" is a type of vegetable that looks like a big pumpkin. An example of a sentence that uses the word yalubalu is:
I was on a trip to Africa and I tried this yalubalu vegetable that was grown in a garden there. It was delicious.

A "Burringo" is a car with very fast acceleration. An example of a sentence that uses the word Burringo is:
In our garage we have a Burringo that my father drives to work every day.

A "Gigamuru" is a type of Japanese musical instrument. An example of a sentence that uses the word Gigamuru is:
I have a Gigamuru that my uncle gave me as a gift. I love to play it at home.

To "screeg" something is to swing a sword at it. An example of a sentence that uses the word screeg is:
We screeghed at each other for several minutes and then we went outside and ate ice cream.

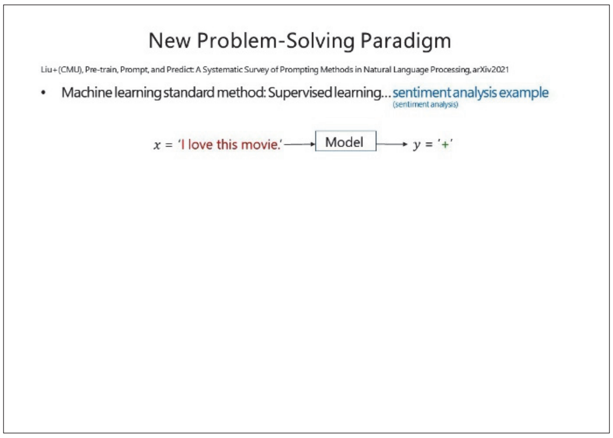
Slide 26

I have gone off at a tangent and will get a bit more on track now.

This language model is now a big turning point in AI. It can create fake news as well as program code supplements. I have only shown a few examples of the greatest things where it can solve some problems without learning everything.

For example, it can even create novel words using word-pairs and define the word in a sentence with the right feeling (slide 26). In this case, only a word-pair is shown, but for example, after being shown several word-pairs this model can create and define a novel word with a usage example. In brief, the paired novel word and definition can be input as text into the language model to predict the next word. Doing so outputs expected examples. Here (slide 26), the non-existent word 'farduddles' is defined as 'jump up and down really fast' in the usage example.

This is not learning everything about a problem to generate usage definitions as sentences with the right feeling from the created novel words but is instead using the few-shots method to understand and predict the next word.



New Problem-Solving Paradigm

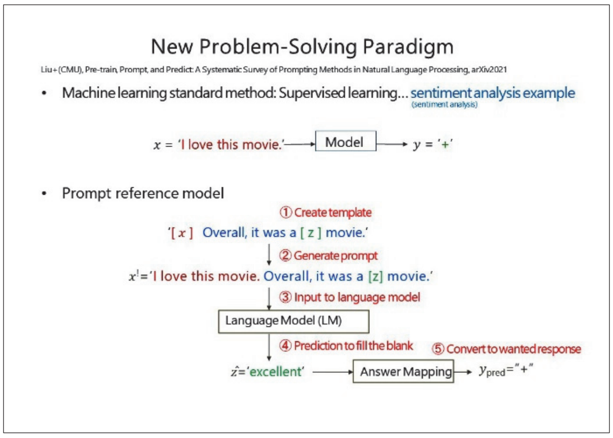
Liu+(CMU), Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing, arXiv2021

- Machine learning standard method: Supervised learning... sentiment analysis example (sentiment analysis)

Slide 27

Perhaps some people don't understand what I mean but various problems can be solved using a language model that has been pre-taught corresponding sentences.

For example, sentiment analysis is a typical natural-language processing problem. As an example, when reviewing images, the simplest prediction is whether the viewer's comment is positive or negative. In more detailed terms, movies have five ranking, ranging from "I love this movie" to "I hate this movie." Although this comment is short, inputting longer impressions can predict whether the viewers evaluation is good or bad, so this type of problem can be solved by machine learning.



New Problem-Solving Paradigm

Liu+(CMU), Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing, arXiv2021

- Machine learning standard method: Supervised learning... sentiment analysis example (sentiment analysis)
- Prompt reference model

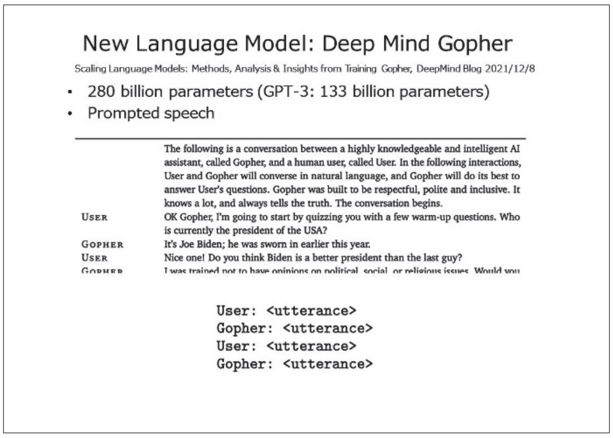
x = 'I love this movie.'
Model
y = '+'
① Create template
'[x] Overall, it was a [z] movie.'
② Generate prompt
x' = 'I love this movie. Overall, it was a [z] movie.'
③ Input to language model
Language Model (LM)
④ Prediction to fill the blank
⑤ Convert to wanted response
ẑ = 'excellent'
Answer Mapping
y_pred = "+"

Slide 28

Conventionally, neural networks have been trained by providing large amounts of 'positive ranking if this sentence', and 'negative if this sentence' data pairs. However, using the new language model paradigm, the same result can be obtained by devising a 'prompt' type input.

How is this done? In slide 28, x is called a slot (location) for accepting the impression and comment. Followed by 'Overall it was a 'something' movie', which, for example, generates the prompt, ‘I love this movie. Overall, it was a ‘something’ movie.’ The ‘something’ part is filled in by the language model. Since GPT-3 predicts the next word and the sentence has a good feeling up to ‘it was a’, the language model predicts the word ‘excellent’. But why choose ‘excellent’? Well, it’s because the language evaluated ‘I love this movie.’ as a positive sentiment. The positive evaluation is the result of learning several trillion texts so ‘… it was a ‘something’ movie’ was evaluated as predicting insertion of a positive adjective. As a result of this prediction, the model itself understands that a positive meaning has priority and decides to use the positive word ‘excellent’.

I am being repetitive in saying this over, but the language model does not need conventional full input-to-output learning. It just needs to learn an existing sentence to predict the next word. Consequently, this learning-free use can solve new problems.



New Language Model: Deep Mind Gopher

Scaling Language Models: Methods, Analysis & Insights from Training Gopher, DeepMind Blog 2021/12/8

- 280 billion parameters (GPT-3: 133 billion parameters)
- Prompted speech

	The following is a conversation between a highly knowledgeable and intelligent AI assistant, called Gopher, and a human user, called User. In the following interactions, User and Gopher will converse in natural language, and Gopher will do its best to answer User’s questions. Gopher was built to be respectful, polite and inclusive. It knows a lot, and always tells the truth. The conversation begins.
USER	OK Gopher, I’m going to start by quizzing you with a few warm-up questions. Who is currently the president of the USA?
GOPHER	It’s Joe Biden; he was sworn in earlier this year.
USER	Nice one! Do you think Biden is a better president than the last guy?
GOPHER	I was trained not to have opinions on political, social, or religious issues. Would you

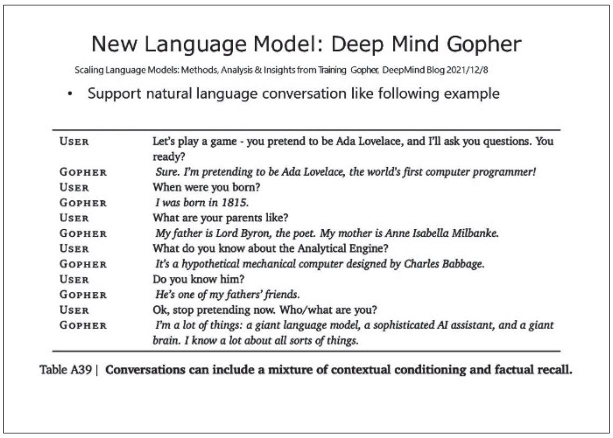
```
User: <utterance>
Gopher: <utterance>
User: <utterance>
Gopher: <utterance>
```

Slide 29

I touched on this earlier, so we know performance increases as the learning scale increases (slide 14). About 2 weeks ago, DeepMind launched its latest language model called Gopher. This massive network uses 280 billion parameters, and many very interesting applications appear in this paper (slide 29) but the greatest impression is the so-called ‘Conversation’ problem. AI-based conversation generally uses the ‘prompt’ (something input?), and Gopher uses input sentences.

The following is a conversation between Gopher, a highly knowledgeable and intelligent AI assistant, and a human user, called User.

The next part of this description explains the Gopher development. This is just the following explanation, these word of the explanation are entered into a language model. The user inputs something to the Gopher language model and Gopher replies. The user responds to the reply and the sequence is repeated over. Gopher’s replies use next-word prediction to spin-out the conversation more and more.



New Language Model: Deep Mind Gopher

Scaling Language Models: Methods, Analysis & Insights from Training Gopher, DeepMind Blog 2021/12/8

- Support natural language conversation like following example

USER	Let’s play a game - you pretend to be Ada Lovelace, and I’ll ask you questions. You ready?
GOPHER	*Sure. I’m pretending to be Ada Lovelace, the world’s first computer programmer!*
USER	When were you born?
GOPHER	*I was born in 1815.*
USER	What are your parents like?
GOPHER	*My father is Lord Byron, the poet. My mother is Anne Isabella Milbanke.*
USER	What do you know about the Analytical Engine?
GOPHER	*It’s a hypothetical mechanical computer designed by Charles Babbage.*
USER	Do you know him?
GOPHER	*He’s one of my fathers’ friends.*
USER	Ok, stop pretending now. Who/what are you?
GOPHER	*I’m a lot of things: a giant language model, a sophisticated AI assistant, and a giant brain. I know a lot about all sorts of things.*

Table A39 | **Conversations can include a mixture of contextual conditioning and factual recall.**

Slide 30

So, let’s look at this example of a conversation in slide 30.

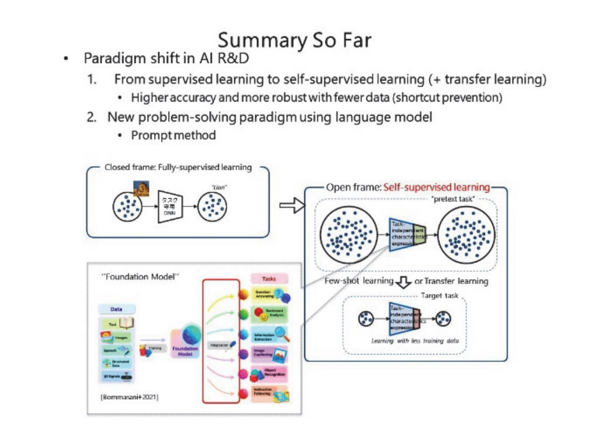
First, the user (called User) says “Let's play a game” and then suggests that the Gopher AI pretends to be Ada Lovelace to whom User will ask question. Finally, User starts the dialog by asking Gopher the question “Are you ready?” Gopher replies “Sure.” (meaning OK) and then repeats back to User the first statement that Gopher is pretending to be the world’s first computer programmer, Ada Lovelace. After User asks, “When were you born?”, Gopher replies, “I was born in 1815.” When next asked by User, “What are your parents like?”, Gopher replies, “My father… . My mother is… .”

Ada Lovelace is famous for her work with Charles Babbage’s concept for a mechanical computer called the Analytical Engine, so when User asks ‘Ada’,

“What do you know about the Analytical Engine?” Gopher replies that “It’s a hypothetical mechanical computer designed by Charles Babbage.” In reply to User’s question. “Do you know him?” Ada (Gopher) replies “He’s one of my father’s friends.” Gopher has learned these trivia from text, so the answers are correct. The next amazing thing is Gopher’s reply to User’s question “Ok, stop pretending now. Who/what are you?” as “I am a lot of things: a giant language model, a sophisticated AI assistant, and a giant brain. I know all sorts of things.” This reply ends the dialog.

At least we can finally have a conversation but talking to a computer with such precision feels like the world of science fiction.

I'll repeat it again that a language model is not made to perform such dialog because it does not learn. It has just been trained with common sentences that would probably contain such conversations too. As a result, the model seems to acquire this type of experience-based dialog format, so we could say it can do such a thing.



Summary So Far

- Paradigm shift in AI R&D
  1. From supervised learning to self-supervised learning (+ transfer learning)
     - Higher accuracy and more robust with fewer data (shortcut prevention)
  2. New problem-solving paradigm using language model
     - Prompt method

Slide 31

So, to summarize up to here.

It feels like there has been a paradigm shift in AI R&D. There has been a tectonic shift from the previous supervised learning to self-supervised learning and data scales have already been reduced by using correct-answer labels.

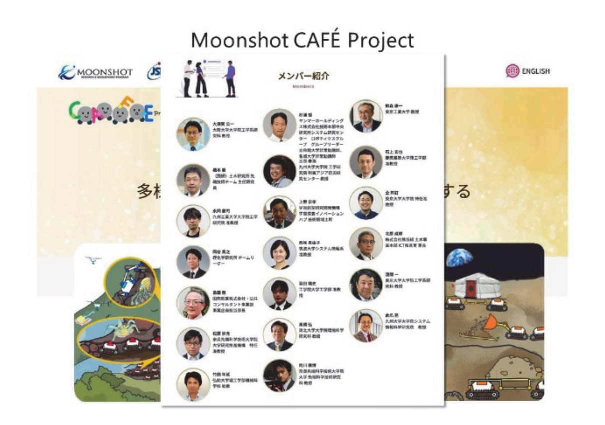
While this is true for images too, more amazingly, language models can be used to solve various problems without the vocabulary learning task by devising prompt input to the language model. As a result, it has become possible to hold something like the dialog shown previously.

Until several years ago, AI researchers commented that artificial general intelligence almost made fools of us because it was thought impossible. In the first part, I mentioned cognitive automation where, until a few years ago, it was believed AI could only automate cognitive processes by reproducing human-decided inputs and outputs. However, looking at what massive language models can do now, the potential of artificial general intelligence seems limitless.

5 Introduction to Current Research

This section introduces some examples of our research.

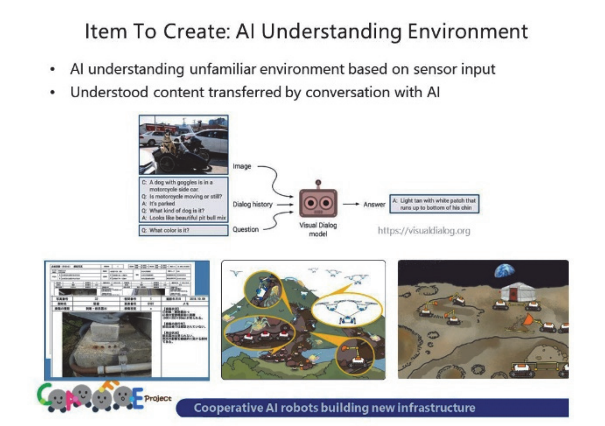
We are doing a lot of things. One concrete problem is visual inspection of completed industrial products. On the other hand, looking at the future, we are researching how to solve harder problems by understanding massive language models as I have just mentioned.



Moonshot CAFÉ Project

Slide 32

For example, we are a part of the nationally financed large project using revolutionary cooperative AI robots to build infrastructure for various environments, and I am a team member (slide 32).



Item To Create: AI Understanding Environment

- AI understanding unfamiliar environment based on sensor input
- Understood content transferred by conversation with AI

C: A dog with goggles is in a motorcycle side car.
Q: Is motorcycle moving or still?
A: It's parked
Q: What kind of dog is it?
A: Looks like beautiful pit bull mix

Q: What color is it?

A: Light tan with white patch that runs up to bottom of his chin

Cooperative AI robots building new infrastructure

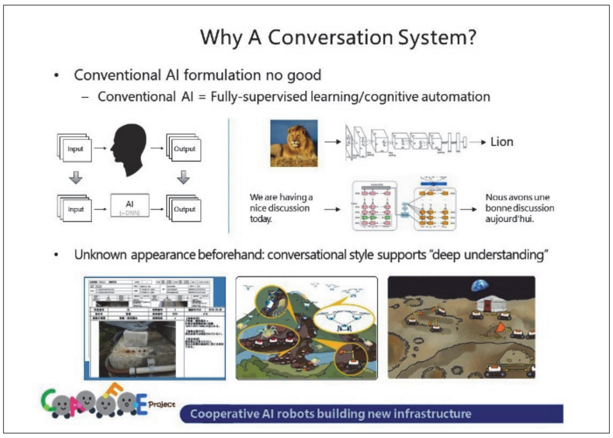
Slide 33

My role in this project is to investigate whether we can build an AI that understands the environment.

The key points are how to express the understood contents, as well as how to extract this data. We've looked at dialog examples using previously described language models but this (slide 33) uses real-world input including images. The input to the neural-net AI is images and dialog to understand things extracted from the images.

For example, think about the following application. The current things we are dealing with are understanding infrastructure diagnostics and soundness. Presently, people onsite perform various evaluations and data documentation but when inputting these images to the network and having a dialog, we need to consider how to evaluate what the situation is now.

This project is described as moonshot-type research because the overall aim is to use robots to automatically build a base on the moon by 2050 in the longer term, as well as to use robots for natural-disaster assistance and recovery in the shorter term. Since the robot workforce will be operating autonomously to some extent, it is critical to understand the work environment.



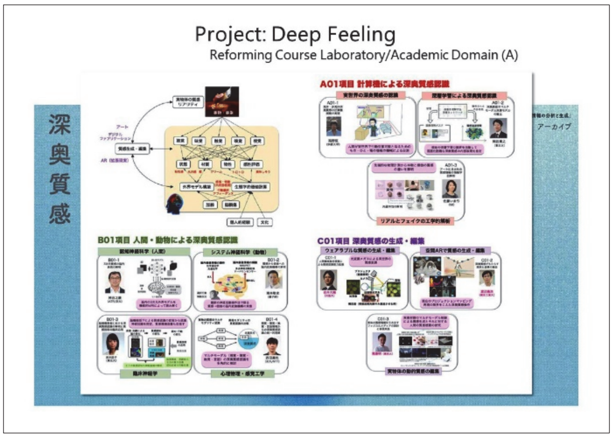
Why A Conversation System?

- Conventional AI formulation no good
  - Conventional AI = Fully-supervised learning/cognitive automation
- Unknown appearance beforehand: conversational style supports "deep understanding"

Input
Output
Input
AI
Output
Lion
We are having a nice discussion today.
Nous avons une bonne discussion aujourd'hui.
Project
Cooperative AI robots building new infrastructure

Slide 34

I keep repeating this over, but when including the paradigm shift that I just explained, getting an instantaneous input–output response using the previous AI style is no good. When conversing with another person or AI agent, the first-understood contents become deeper and then the requirements become clear, and an appropriate response can be made.



Project: Deep Feeling

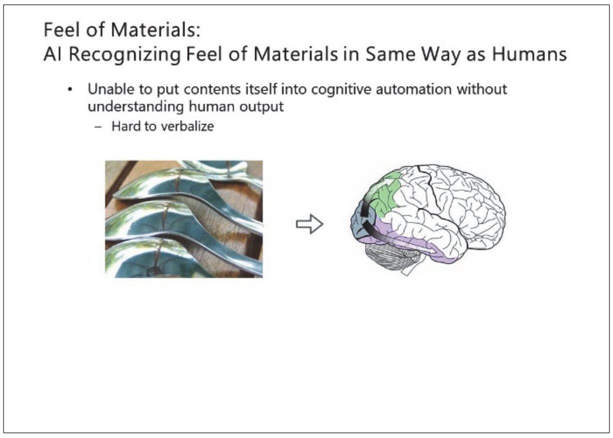
Reforming Course Laboratory/Academic Domain (A)

深奥質感
A01項目 計算機による深奥質感認識
B01項目 人間・動物による深奥質感認識
C01項目 深奥質感の生成・編集
リアルとフェイクの工学的解析
ウェアラブルな質感の生成・編集
臨床神経学
心理物理・感覚工学
実物体の動的質感の編集
アーカイブ

Slide 35

The project in slide 35 is different and slightly academic research into 'deep feeling.'

This is also a group research project; I am leading the work thinking about the feeling of things using AI. 'Feeling' is a very difficult thing to express; it is a word that expresses a wide range of things, so it is wide-ranging research. We are using an image-based approach to finding the next word to create the feeling of materials by using how a person or animal recognizes the feel of a material.



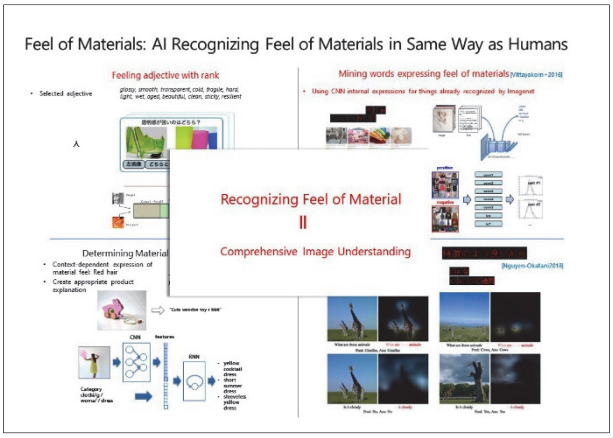
Feel of Materials:
AI Recognizing Feel of Materials in Same Way as Humans

- Unable to put contents itself into cognitive automation without understanding human output
  - Hard to verbalize

Slide 36

The deep feeling of materials is a difficult problem for AI now. Most things can be done by preparing training and specifying inputs and outputs. However, the first problem is that it is unclear what a person outputs and I think recognizing the feel of material is a good example.

Although looking at the image on the left side of slide 36 will get an image-recognition response like, "These are spoons.", this is not the whole response and there will be other understood ideas like "cold when touched", "hard", "reflecting something". I think these 'feelings' can be recognized in the same way. If the conventional approach does not reveal what a human being outputs, it does not fit into the deep-learning frame, which is difficult. Clearly, it is hard to verbalize a feeling. There are various problems in how to communicate one's feeling about a material to another person simply because it is difficult.



Feel of Materials: AI Recognizing Feel of Materials in Same Way as Humans

Feeling adjective with rank

- Selected adjective glossy, smooth, transparent, cold, fragile, hard, light, wet, aged, beautiful, clean, sticky, resilient

Mining words expressing feel of materials [Vittayakorn+2016]

- Using CNN internal expressions for things already recognized by Imagenet

Recognizing Feel of Material
||
Comprehensive Image Understanding

Determining Material

- Context-dependent expression of material feel: Red hair
- Create appropriate product explanation

[Nguyen-Okatani2018]

Slide 37

Various research has been done based on these properties. There isn't time to discuss the details, but it will clearly be impossible for AI to recognize the feeling of materials without comprehending the thing in the image.

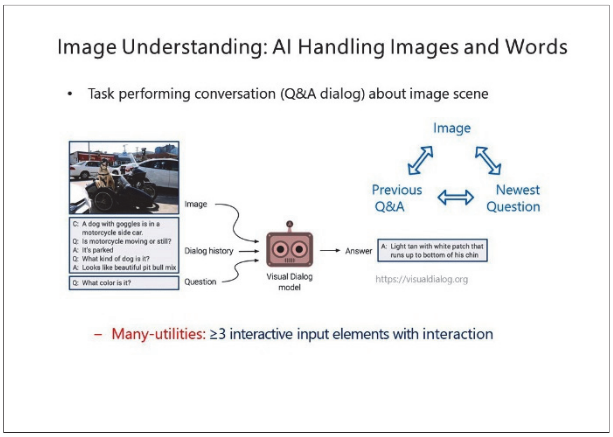


Image Understanding: AI Handling Images and Words

- Task performing conversation (Q&A dialog) about image scene

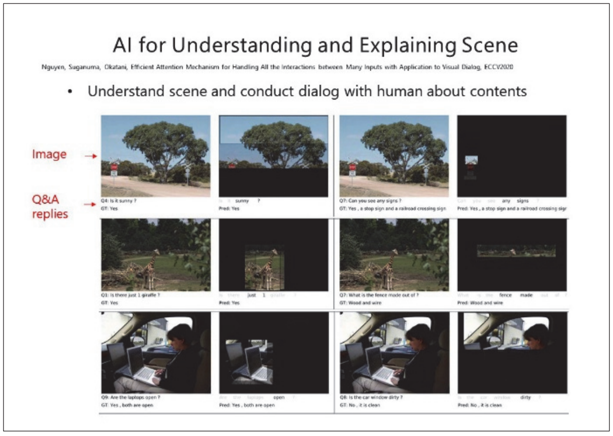
C: A dog with goggles is in a motorcycle side car.
Q: Is motorcycle moving or still?
A: It's parked
Q: What kind of dog is it?
A: Looks like beautiful pit bull mix
Q: What color is it?
Image
Dialog history
Question
Visual Dialog model
Answer
A: Light tan with white patch that runs up to bottom of his chin
Image
Previous Q&A
Newest Question

https://visualdialog.org

  - Many-utilities: ≥3 interactive input elements with interaction

Slide 38

As mentioned earlier, when holding a dialog about images with AI, the dialog is driven by what is taking place and the scene contents.



AI for Understanding and Explaining Scene

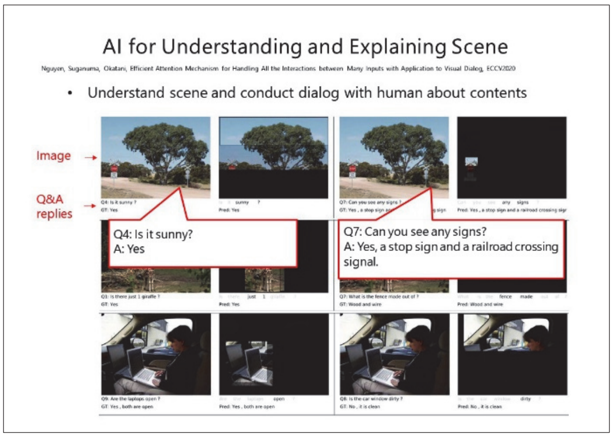
Nguyen, Suganuma, Okatani, Efficient Attention Mechanism for Handling All the Interactions between Many Inputs with Application to Visual Dialog, ECCV2020

- Understand scene and conduct dialog with human about contents

Image
Q&A replies
Q4: Is it sunny ?
GT: Yes
sunny ?
Pred: Yes
Q7: Can you see any signs ?
GT: Yes , a stop sign and a railroad crossing sign
any signs
Pred: Yes , a stop sign and a railroad crossing sign
Q1: Is there just 1 giraffe ?
GT: Yes
just 1
Pred: Yes
Q7: What is the fence made out of ?
GT: Wood and wire
fence made
Pred: Wood and wire
Q9: Are the laptops open ?
GT: Yes , both are open
open
Pred: Yes , both are open
Q8: Is the car window dirty ?
GT: No , it is clean
dirty
Pred: No , it is clean

Slide 39

Having this ability may open-up a research avenue.



AI for Understanding and Explaining Scene

Nguyen, Suganuma, Okatani, Efficient Attention Mechanism for Handling All the Interactions between Many Inputs with Application to Visual Dialog, ECCV2020

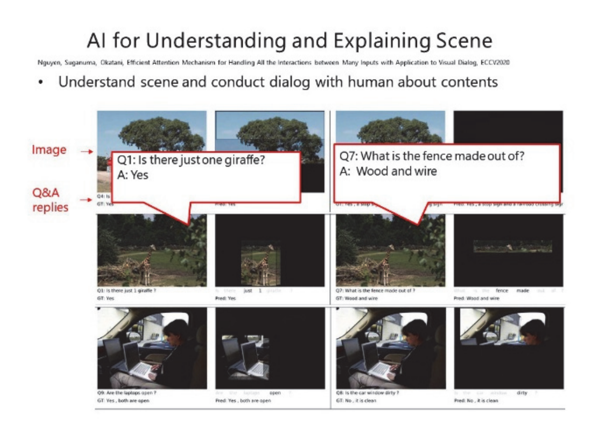
- Understand scene and conduct dialog with human about contents

Q4: Is it sunny?
A: Yes

Q7: Can you see any signs?
A: Yes, a stop sign and a railroad crossing signal.

Slide 40

We are already responding to these feelings. When looking at these images, the AI replies "Yes." to the question, "Is it sunny?" by looking at the sky, showing where the AI actually looked. The dialog continues several times and then the AI answers the question (obviously in English), "Can you see any signs?" with the reply, "Yes, a stop signal and a railroad crossing signal."



AI for Understanding and Explaining Scene

Nguyen, Suganuma, Okatani, Efficient Attention Mechanism for Handling All the Interactions between Many Inputs with Application to Visual Dialog, ECCV2020

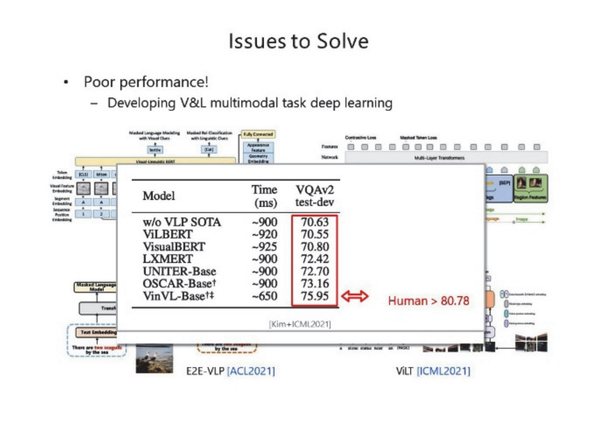
- Understand scene and conduct dialog with human about contents

Q1: Is there just one giraffe?
A: Yes

Q7: What is the fence made out of?
A: Wood and wire

Slide 41

In slide 39-41, the reply is "Yes." to the question, "Is there just one giraffe?" The question, "What is the fence made out of?" is answered by a 'feeling' type reply, "Wood and wire."



Issues to Solve

- Poor performance!
  - Developing V&L multimodal task deep learning

Model	Time (ms)	VQAv2 test-dev
w/o VLP SOTA	~900	70.63
ViLBERT	~920	70.55
VisualBERT	~925	70.80
LXMERT	~900	72.42
UNITER-Base	~900	72.70
OSCAR-Base†	~900	73.16
VinVL-Base†‡	~650	75.95

Human > 80.78

[Kim+ICML2021]

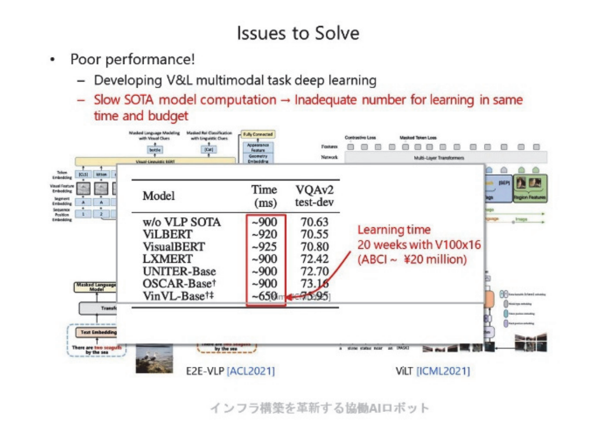
E2E-VLP [ACL2021]

ViLT [ICML2021]

Slide 42

These things are solved by the current technologies but have various problems. I haven't really touched on this in concrete terms during this talk, but performance is a problem anyway. Although the language model has made great advances in handling both images and natural language very well, responses are still a problem.

There is already a mountain of research, and as I said at the beginning, research has taken different approaches to the neural network internal structure and how to formulate the problem. However, for Q&A dialog tasks, the performance by the latest systems for correct responses to questions about images still cannot match the human response. It has gotten close recently (75.95 vs >80.78 for human) but there is still a need for improved performance.



Issues to Solve

- Poor performance!
  - Developing V&L multimodal task deep learning
  - Slow SOTA model computation → Inadequate number for learning in same time and budget

Model	Time (ms)	VQAv2 test-dev
w/o VLP SOTA	~900	70.63
ViLBERT	~920	70.55
VisualBERT	~925	70.80
LXMERT	~900	72.42
UNITER-Base	~900	72.70
OSCAR-Base†	~900	73.16
VinVL-Base†‡	~650	75.95

Learning time
20 weeks with V100x16
(ABCI ~ ¥20 million)

E2E-VLP [ACL2021]

ViLT [ICML2021]

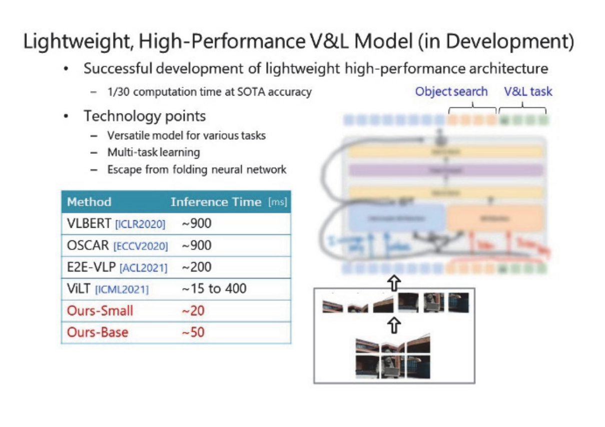
インフラ構築を革新する協働AIロボット

Slide 43

Another issue is how to improve the long 1-second question and response time for each dialog round without increasing the neural network size, although waiting 1 second may not actually be a problem in real conversation. However, since conversational learning repeats Q&A rounds millions or tens of millions of times, a longer Q&A round time can have a large impact on the system. Shortening the Q&A round time is expected to cost money!

Laboratory computers can be used for normal research but they are inadequate for this level, so cloud service time must be purchased. In Japan, the GPU server with the best cost-performance is the ABCI (AI Bridging Cloud Infrastructure/AI) at the National Institute of Advanced Industrial Science and Technology, but even ABCI costs about ¥20 million to simulate the learning reported in this article, which is an impossible cost burden.

Actually, I need about 10 times this figure because I run a lot of trial and error tests. Consequently, these studies have become impossible in our laboratory.



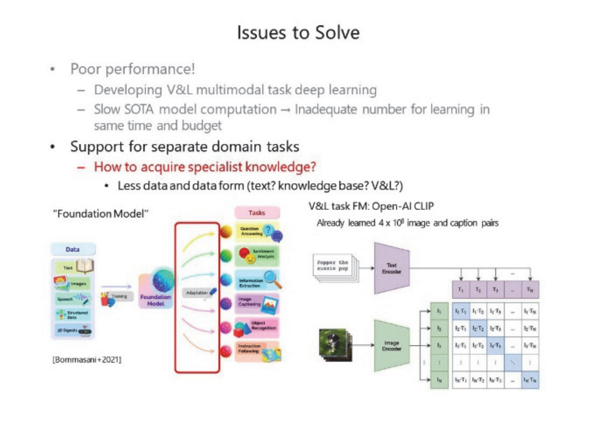
Lightweight, High-Performance V&L Model (in Development)

- Successful development of lightweight high-performance architecture
  - 1/30 computation time at SOTA accuracy
- Technology points
  - Versatile model for various tasks
  - Multi-task learning
  - Escape from folding neural network

Method	Inference Time [ms]
VLBERT [ICLR2020]	~900
OSCAR [ECCV2020]	~900
E2E-VLP [ACL2021]	~200
ViLT [ICML2021]	~15 to 400
Ours-Small	~20
Ours-Base	~50

Slide 44

As a solution, we are currently working on building a lightweight, high-performance model although some details are obscured here (slide 44). It is expected to cut computation times to 3% or even 2%. A cut to 3% would slash the ¥20 million time cost to a more sustainable few ¥million.



Issues to Solve

- Poor performance!
  - Developing V&L multimodal task deep learning
  - Slow SOTA model computation → Inadequate number for learning in same time and budget
- Support for separate domain tasks
  - How to acquire specialist knowledge?
    - Less data and data form (text? knowledge base? V&L?)

"Foundation Model"

[Bommasani+2021]

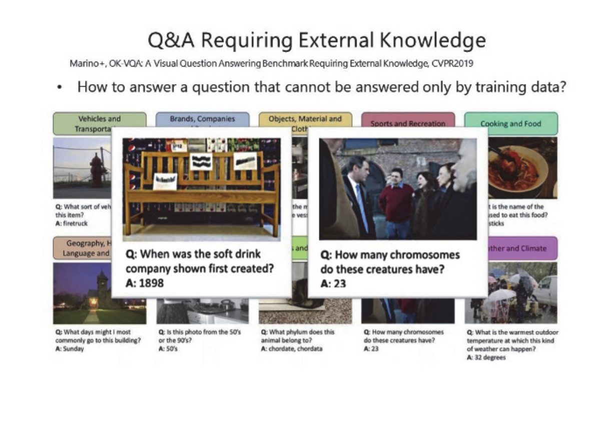
V&L task FM: Open-AI CLIP

Already learned 4 x 10^8 image and caption pairs

Slide 45

Last, I want to talk about another big problem for a series of projects–how to acquire specialist knowledge. In slide 33, I mentioned this bridge, but the first problem in inspecting an infrastructure bridge is an inspection specialist so how do we acquire the required specialist knowledge?

If, like a language model, specialist knowledge could be expressed as text, text learning might progress well. However, in the first place, we still don't completely understand how to express knowledge.



Q&A Requiring External Knowledge

Marino+, OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge, CVPR2019

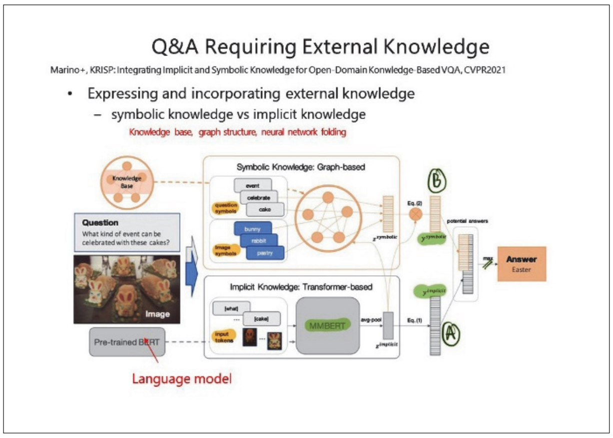
- How to answer a question that cannot be answered only by training data?

Q: When was the soft drink company shown first created?
A: 1898

Q: How many chromosomes do these creatures have?
A: 23

Vehicles and Transporta
Brands, Companies
Objects, Material and Cloth
Sports and Recreation
Cooking and Food
Geography, H Language and
and
ther and Climate
Q: What sort of veh this item?
A: firetruck
the m e vess
t is the name of the sed to eat this food? sticks
Q: What days might I most commonly go to this building?
A: Sunday
Q: Is this photo from the 50's or the 90's?
A: 50's
Q: What phylum does this animal belong to?
A: chordate, chordata
Q: How many chromosomes do these creatures have?
A: 23
Q: What is the warmest outdoor temperature at which this kind of weather can happen?
A: 32 degrees

Slide 46



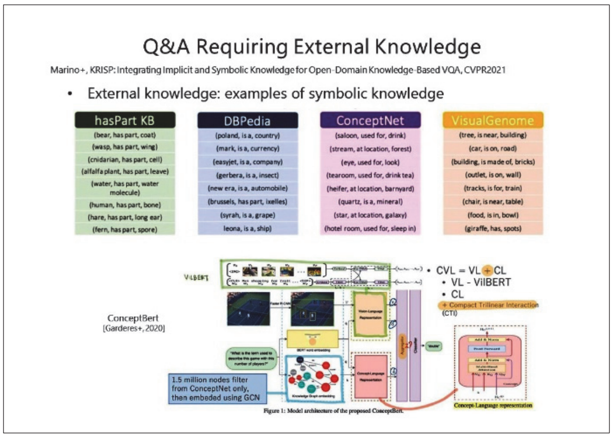
Q&A Requiring External Knowledge

Marino+, KRISP: Integrating Implicit and Symbolic Knowledge for Open-Domain Konwledge-Based VQA, CVPR2021

- Expressing and incorporating external knowledge
  - symbolic knowledge vs implicit knowledge

Knowledge base, graph structure, neural network folding

Slide 47



Q&A Requiring External Knowledge

Marino+, KRISP: Integrating Implicit and Symbolic Knowledge for Open-Domain Knowledge-Based VQA, CVPR2021

- External knowledge: examples of symbolic knowledge

hasPart KB	DBPedia	ConceptNet	VisualGenome
(bear, has part, coat)	(poland, is a, country)	(saloon, used for, drink)	(tree, is near, building)
(wasp, has part, wing)	(mark, is a, currency)	(stream, at location, forest)	(car, is on, road)
(cnidarian, has part, cell)	(easyjet, is a, company)	(eye, used for, look)	(building, is made of, bricks)
(alfalfa plant, has part, leave)	(gerbera, is a, insect)	(tearoom, used for, drink tea)	(outlet, is on, wall)
(water, has part, water molecule)	(new era, is a, automobile)	(heifer, at location, barnyard)	(tracks, is for, train)
(human, has part, bone)	(brussels, has part, ixelles)	(quartz, is a, mineral)	(chair, is near, table)
(hare, has part, long ear)	(syrah, is a, grape)	(star, at location, galaxy)	(food, is in, bowl)
(fern, has part, spore)	leona, is a, ship)	(hotel room, used for, sleep in)	(giraffe, has, spots)

- CVL = VL + CL
  - VL - VilBERT
  - CL

  + Compact Trilinear Interaction (CTI)

ConceptBert [Garderes+, 2020]

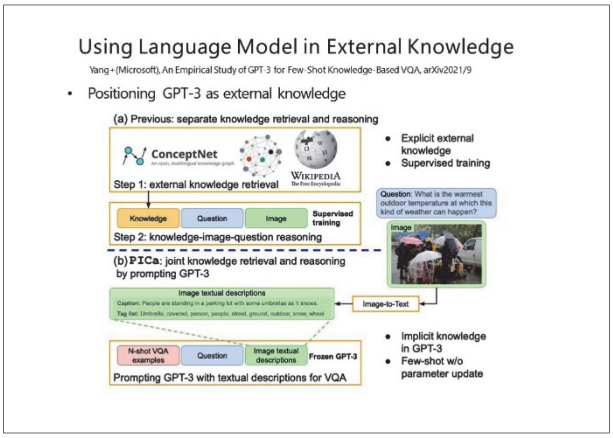
1.5 million nodes filter from ConceptNet only, then embeded using GCN

VILBERT
Faster R-CNN
Vision-Language Representation
BERT word embedding
Knowledge Graph embedding
Concept-Language Representation
Aggregator
Classifier
Concept-Language representation

Figure 1: Model architecture of the proposed ConceptBert.

Slide 48

I have talked briefly to you about the current and latest research and what we are doing.



Using Language Model in External Knowledge

Yang+(Microsoft), An Empirical Study of GPT-3 for Few-Shot Knowledge-Based VQA, arXiv2021/9

- Positioning GPT-3 as external knowledge

(a) Previous: separate knowledge retrieval and reasoning

Step 1: external knowledge retrieval

Step 2: knowledge-image-question reasoning

(b) PICa: joint knowledge retrieval and reasoning by prompting GPT-3

Prompting GPT-3 with textual descriptions for VQA

- Explicit external knowledge
- Supervised training

Question: What is the warmest outdoor temperature at which this kind of weather can happen?

Image textual descriptions

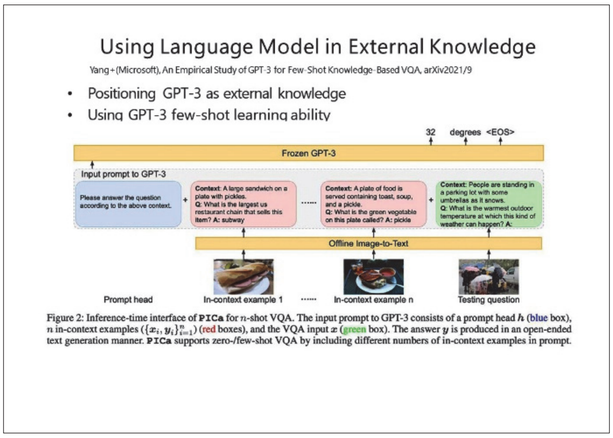
Caption: People are standing in a parking lot with some umbrellas as it snows.

Tag list: Umbrella, covered, person, people, street, ground, outdoor, snow, wheel

- Implicit knowledge in GPT-3
- Few-shot w/o parameter update

Slide 49

These slides show some likely projects using massive language modes being researched in the last few months, but I cannot describe them due to time constraints.



Using Language Model in External Knowledge

Yang+(Microsoft), An Empirical Study of GPT-3 for Few-Shot Knowledge-Based VQA, arXiv2021/9

- Positioning GPT-3 as external knowledge
- Using GPT-3 few-shot learning ability

Input prompt to GPT-3

Please answer the question according to the above context.

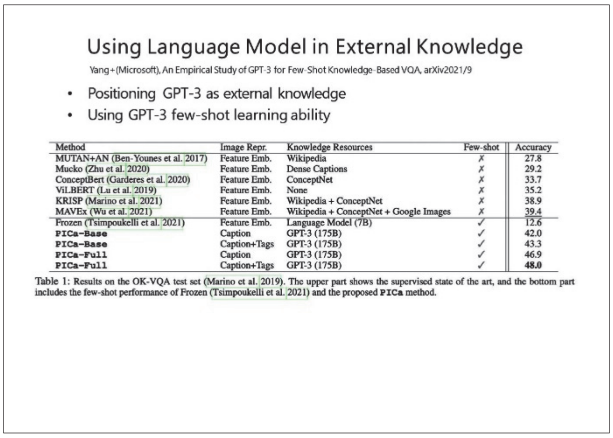
Context: A large sandwich on a plate with pickles.
Q: What is the largest us restaurant chain that sells this item? A: subway

Context: A plate of food is served containing toast, soup, and a pickle.
Q: What is the green vegetable on this plate called? A: pickle

Context: People are standing in a parking lot with some umbrellas as it snows.
Q: What is the warmest outdoor temperature at which this kind of weather can happen? A:

Figure 2: Inference-time interface of PICa for n-shot VQA. The input prompt to GPT-3 consists of a prompt head h (blue box), n in-context examples ($\{x_i, y_i\}_{i=1}^n$) (red boxes), and the VQA input x (green box). The answer y is produced in an open-ended text generation manner. PICa supports zero-/few-shot VQA by including different numbers of in-context examples in prompt.

Slide 50



Using Language Model in External Knowledge

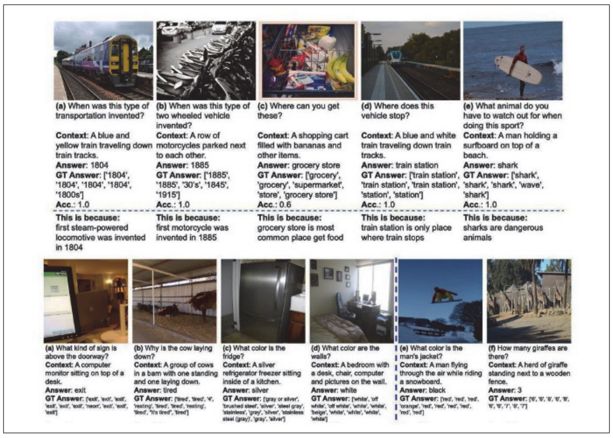
Yang+(Microsoft), An Empirical Study of GPT-3 for Few-Shot Knowledge-Based VQA, arXiv2021/9

- Positioning GPT-3 as external knowledge
- Using GPT-3 few-shot learning ability

Method	Image Repr.	Knowledge Resources	Few-shot	Accuracy
MUTAN+AN (Ben-Younes et al. 2017)	Feature Emb.	Wikipedia	✗	27.8
Mucko (Zhu et al. 2020)	Feature Emb.	Dense Captions	✗	29.2
ConceptBert (Garderes et al. 2020)	Feature Emb.	ConceptNet	✗	33.7
ViLBERT (Lu et al. 2019)	Feature Emb.	None	✗	35.2
KRISP (Marino et al. 2021)	Feature Emb.	Wikipedia + ConceptNet	✗	38.9
MAVEx (Wu et al. 2021)	Feature Emb.	Wikipedia + ConceptNet + Google Images	✗	39.4
Frozen (Tsimpoukelli et al. 2021)	Feature Emb.	Language Model (7B)	✓	12.6
PICa-Base	Caption	GPT-3 (175B)	✓	42.0
PICa-Base	Caption+Tags	GPT-3 (175B)	✓	43.3
PICa-Full	Caption	GPT-3 (175B)	✓	46.9
PICa-Full	Caption+Tags	GPT-3 (175B)	✓	**48.0**

Table 1: Results on the OK-VQA test set (Marino et al. 2019). The upper part shows the supervised state of the art, and the bottom part includes the few-shot performance of Frozen (Tsimpoukelli et al. 2021) and the proposed PICa method.

Slide 51



(a) When was this type of transportation invented?
Context: A blue and yellow train traveling down train tracks.
Answer: 1804
GT Answer: ['1804', '1804', '1804', '1804', '1800s']
Acc.: 1.0
This is because:
first steam-powered locomotive was invented in 1804

(b) When was this type of two wheeled vehicle invented?
Context: A row of motorcycles parked next to each other.
Answer: 1885
GT Answer: ['1885', '1885', '30's', '1845', '1915']
Acc.: 1.0
This is because:
first motorcycle was invented in 1885

(c) Where can you get these?
Context: A shopping cart filled with bananas and other items.
Answer: grocery store
GT Answer: ['grocery', 'grocery', 'supermarket', 'store', 'grocery store']
Acc.: 0.6
This is because:
grocery store is most common place get food

(d) Where does this vehicle stop?
Context: A blue and white train traveling down train tracks.
Answer: train station
GT Answer: ['train station', 'train station', 'train station', 'station', 'station']
Acc.: 1.0
This is because:
train station is only place where train stops

(e) What animal do you have to watch out for when doing this sport?
Context: A man holding a surfboard on top of a beach.
Answer: shark
GT Answer: ['shark', 'shark', 'shark', 'wave', 'shark']
Acc.: 1.0
This is because:
sharks are dangerous animals

(a) What kind of sign is above the doorway?
Context: A computer monitor sitting on top of a desk.
Answer: exit
GT Answer: ['exit', 'exit', 'exit', 'exit', 'exit', 'exit', 'neon', 'exit', 'exit', 'exit']

(b) Why is the cow laying down?
Context: A group of cows in a barn with one standing and one laying down.
Answer: tired
GT Answer: ['tired', 'tired', '4', 'resting', 'tired', 'tired', 'resting', 'tired', "it's tired", 'tired']

(c) What color is the fridge?
Context: A silver refrigerator freezer sitting inside of a kitchen.
Answer: silver
GT Answer: ['gray or silver', 'brushed steel', 'silver', 'steel gray', 'stainless', 'gray', 'silver', 'stainless steel (gray)', 'gray', 'silver']

(d) What color are the walls?
Context: A bedroom with a desk, chair, computer and pictures on the wall.
Answer: white
GT Answer: ['white', 'off white', 'off white', 'white', 'white', 'beige', 'white', 'white', 'white', 'white']

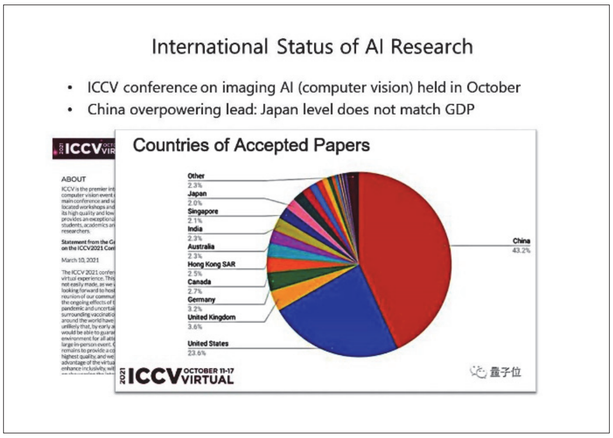
(e) What color is the man's jacket?
Context: A man flying through the air while riding a snowboard.
Answer: black
GT Answer: ['red', 'red', 'red', 'orange', 'red', 'red', 'red', 'red', 'red', 'red']

(f) How many giraffes are there?
Context: A herd of giraffe standing next to a wooden fence.
Answer: 3
GT Answer: ['6', '6', '8', '6', '8', '6', '6', '7', '8', '7']

Slide 52

6 State of AI Research and Human Talent

During a recent consultation, I spoke on the current state of AI research. In summary, Japan is working on AI research.



International Status of AI Research

- ICCV conference on imaging AI (computer vision) held in October
- China overpowering lead: Japan level does not match GDP

Slide 53

The most recent example was the International Conference on Computer Vision (ICCV) held virtually in October 2021. Probably only people with a deep interest in computer science will know about it but now the Conference Proceedings are the focus of attention.

In the general science and engineering research fields, publishing an article in a journal is considered a key accomplishment, but in computer science, speaking at a conference is best. Speaker selection is very strict and only 10% or 20% of presentations are chosen. With a one in five success rate, everybody posts articles online wherever they can.

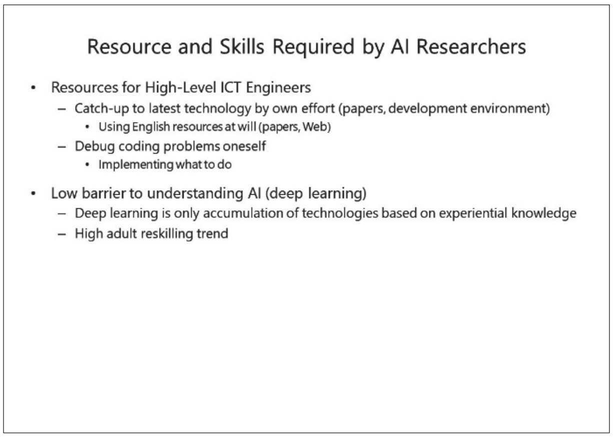
Especially in the AI field, the number of researchers in computer vision has increased more than tenfold over the last 10 or so years, and the field has become very competitive. I have presented two or three articles and students' doctoral materials at top conferences where well-known companies like Google and META as well as anonymous companies recruit for their AI and computer vision business. Everybody, including students, is desperate to make a presentation at these conferences.

The pie chart analyzes the number of articles presented at ICCV. I recollect vaguely that the total was several thousand.

As you can see, presentations from China were overwhelmingly dominant at twice the number from the USA. However, although US presenters announced their affiliations to American organizations, perhaps more than half had Chinese nationality. Therefore, we can say that Chinese researchers dominate in this research world. This is not only true in the field of computer vision, but I think it is also true in AI and other computer service fields.

Sadly, Japan only has this small presence despite having the world's third largest GDP. After these last 10 years, I'd like to at least come third, but it seems not to be. However, 2% still feels good. There are a number of reasons why but in simple terms, there is not enough money in the academic research world. Conversely, China has a huge number of researchers in well-funded universities, resulting in the above dominance.

Progress centered on the language model explained first and the Foundation Model has been amazing, and it feels to me like we are entering an SF world. In these circumstances, I think there are dangers for the future strength of the nation.



Resource and Skills Required by AI Researchers

- Resources for High-Level ICT Engineers
  - Catch-up to latest technology by own effort (papers, development environment)
    - Using English resources at will (papers, Web)
  - Debug coding problems oneself
    - Implementing what to do
- Low barrier to understanding AI (deep learning)
  - Deep learning is only accumulation of technologies based on experiential knowledge
  - High adult reskilling trend

Slide 54

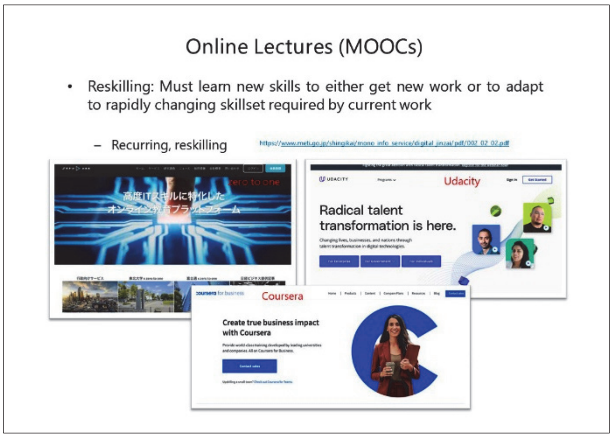
I write about these things in slide 54 as a researcher in this type of research field, looking at our students.

First, high-level ICT engineers are not simple programmers and work at very high software engineering levels. These people should be offered the resources necessary to keep-up with the latest advances. Of course, they must keep their knowledge and skills up-to-date by reading English academic papers and web articles.

Coding skills are obviously important for AI. Writing bad and computationally expensive code without high-level skills embodying the latest methods makes testing impossible in a continuously updated development environment. This ability is important and so are the engineers who have it.

On the other hand, getting a 'handle' on understanding AI or deep learning is not so difficult and is 'not rocket science'. Launching a rocket requires studying a lot of engineering papers and textbooks from the bottom up, which is not really the case for deep learning. The important thing is that there is no theory of deep learning. It is the accumulation of techniques based on limitless empirical knowledge. Of course, since the data scale is increasing, we need to better understand how to suppress the increase, which may be comparable with the threshold for studying quantum mechanics. At the same time, I think people cannot easily advance the field without the necessary resources mentioned earlier.

Obviously, human resources for these new fields go on increasing. Especially in Japan, universities still have some say about when students who have mastered these new fields can join their field, but some never do.



Online Lectures (MOOCs)

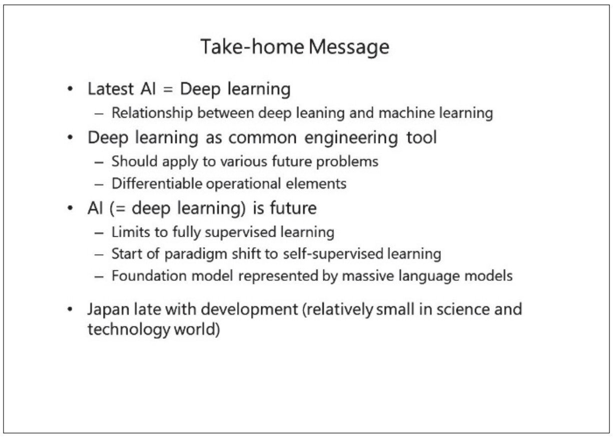
- Reskilling: Must learn new skills to either get new work or to adapt to rapidly changing skillset required by current work
  - Recurring, reskilling https://www.meti.go.jp/shingikai/mono_info_service/digital_jinzai/pdf/002_02_02.pdf

高度ITスキルに特化したオンライン学習プラットフォーム
zero to one
Udacity
Radical talent transformation is here.
Coursera
Create true business impact with Coursera

Slide 55

There have always been ways to learn new skills (re-skilling) and online courses in the AI field have been increasing for the last 10 years. I also have a relationship with this Japanese startup called zero to one (slide 55) and I'd like to tell you a short story about the amazing advances in the USA about teaching various things.

Coursera and Udacity are two well-known online teaching businesses. For example, since the fee for an online study course in autonomous vehicle technology is about ¥500,000 ($4000), the course is used both by private and corporate students. This type of engineering re-education is growing very quickly. In principle, Coursera courses are free, but there are some fees for unit completion and course graduation certificates. Although these are online courses, completing the course proves skill and ability in the subject and these courses are being used by more people in the USA, particularly when changing jobs. This approach has yet to catch-on in Japan but there are still several companies like this zero to one taking a similar approach.



Take-home Message

- Latest AI = Deep learning
  - Relationship between deep leaning and machine learning
- Deep learning as common engineering tool
  - Should apply to various future problems
  - Differentiable operational elements
- AI (= deep learning) is future
  - Limits to fully supervised learning
  - Start of paradigm shift to self-supervised learning
  - Foundation model represented by massive language models
- Japan late with development (relatively small in science and technology world)

Slide 56

So, to summarize–what is the main take-home message from this talk?

I think the word AI is really used as a ruse. It feels as if things like simple data science and Excel spreadsheets are now being called AI. In fact, 'leading-edge AI' is really 'deep learning' and deep learning itself could be called 'machine learning', which is just one technique or one field in a wider range.

In this sense, it is quite different from conventional methodologies and, cannot be understood as a single field. Supervised learning is extremely important in traditional mathematics for creating methods based on mathematical logic.

In contrast, deep learning is different from this type of conventional sense of values and is actually engineering. A network is constructed to solve a problem by brute force using trial and error and something like experience. This means there are both many tools and many problems that can only be solved using deep-learning performance.

Consequently, as well as being one AI methodology, I think deep learning is–and should become–a general engineering tool, and university engineering departments should reconsider it as a subject for teaching.

Although I have not explained this fully today, neural networks are quite unlike the conventionally held image. The general image of a neural network is a biological system of nerves in living things, but here it means an elaborate computer network system built for a specific task and used for learning by computations differentiating outputs from inputs. As a result, neural networks are now used as a hybrid-like network for solving various problems, such as the feel of materials.

Earlier in this talk, I spoke about AI now and in the future as well as about the paradigm shift from conventional, successful but restricted fully-supervised learning to self-supervised learning using massive language models and how current developments have a slightly unpredictable feeling.

Irrespective of this situation, I am sad to report that Japan is well behind compared to the rest of the scientific world.

This concludes today's talk; thank you for your time.

Lecturer Biography

Takayuki Okatani

1999	Department of Mathematical Engineering and Information Physics, Tokyo University, PhD
1999 ~	Graduate School of Information Sciences, Tohoku University, Professor (from 2013)
2016 ~	RIKEN Center for Advanced Intelligence Project