Image Recognition | Deep Learning | Neural Network

Technology Reports

Deep Learning-based Image Recognition Applications

Image recognition services with machine learning have been expanding in recent years. Recognizing abstract concepts from an image by conventional technology is a fundamental task, as in determining the categories of objects appearing in an image (i.e. "food," "flower," etc.). Deep learning has also been gaining in popularity among machine learning applications. NTT DOCOMO has developed an image recognition system based on deep learning technology and has publically released a recognition API. This system enables the building of high-accuracy image recognition models to attach various tags to images simply by training image data prepared beforehand.

Service Innovation Department *Toshiki Sakai*
Xinyu Guo

1. Introduction

Deep learning has been increasing in use and has been enjoying success in a variety of fields. Enterprises such as Google and Facebook in the United States and Baidu in China have established research laboratories and acquired deep learning startups since 2013. For example, Google had come to use deep learning in 47 services including image recognition[*1] and speech recognition as of March 2015 [1].

In image recognition, deep learning has brought significant improvements in accuracy [2] and has progressed greatly in a variety of tasks (**Figure 1**).

NTT DOCOMO previously released an Application Programming Interface (API)[*2] for image recognition using conventional image recognition technology [3] [4]. This API can be used to recognize an object in an image if that object has a definite shape such as a "product package." However, it is incapable of being used for tagging different types of images taken by a user with a smartphone. For this reason, NTT DOCOMO developed image recognition technology using deep learning that can recognize abstract concepts in an image such as the type of scene (wedding ceremony, field day, etc.) or category of an object (food, flower, etc.), names of objects having indefinite shape such as bread and curry rice, and features that have heretofore been dependent on human sensitivities such as the color and pattern of fashion items. This image recognition technology enabled NTT DOCOMO to develop a set of recognition engines capable of high-accuracy tagging. NTT DOCOMO trained these image recognition engines for scenes, fashion items, food, flowers, etc. using a large image dataset and released them in the form of an API in November 2015 [3].

[*1] **Image recognition:** A technology that uses image-processing and machine-learning techniques to mechanically understand images and extract meaning (such as names of objects appearing in an image, type of scene, etc. that a human being could infer from an image).

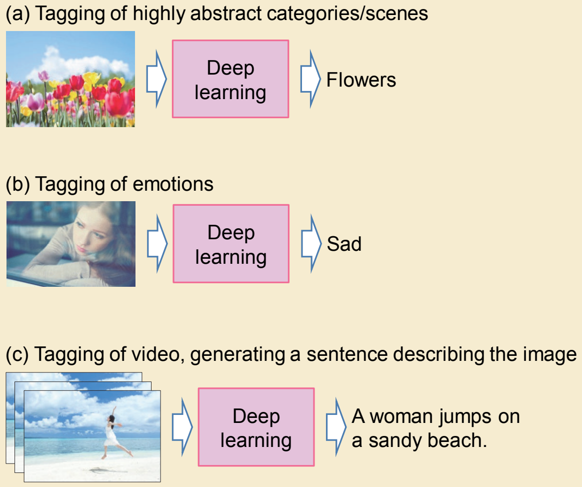

(a) Tagging of highly abstract categories/scenes
Deep learning
Flowers
(b) Tagging of emotions
Deep learning
Sad
(c) Tagging of video, generating a sentence describing the image
Deep learning
A woman jumps on a sandy beach.

Figure 1 Image recognition using deep learning

In this article, we present an overview of deep learning technology. We then describe the differences between conventional technology and image recognition using deep learning. Next, we list the issues resolved by deep learning. Finally, we define the features of an image recognition API service developed and offered by NTT DOCOMO and introduce applications using this API.

2. Overview of Deep Learning

Deep learning is a branch of machine learning*3 technology using multi-layer neural networks. A neural network is a machine learning technique inspired by the information processing mechanism of biological neural networks. Neural networks have been used since the 1950s [5] and have been applied, for example, to the classification task of dividing multidimensional data such as vector data or images into classes (**Figure 2** (a)). Multi-layer neural networks, that is, neural networks with several intermediate layers, can perform more complex classification and recognition tasks (Fig. 2 (b)). Multi-layer neural networks were popular in the 1980s and 1990s and were used in several types of image recognition tasks. It was shown in 1979 that they could be used to achieve a recognition rate of 98.6% for handwritten numerals [6]. However, classical multi-layer neural networks suffer from the problem that increasing the number of layers makes learning much harder and extremely time consuming. For this reason, difficulties in solving a complex recognition task that needs many layers have prevented multi-layer neural networks from reaching a practical level.

To eliminate this problem, technological improvements were made in multi-layer neural network algorithms such as by developing parameter initialization techniques and training techniques to prevent overfitting. At the same time, the parallel distributed processing using General Purpose computing on Graphics Processing Units (GPGPU)*4 dramatically improved learning speed. As a result of these efforts, learning with deep layers became feasible and deep learning grabbed attention once again in the latter half of the 2000s. In the field of image recognition, deep learning based method competed in object recognition accuracy at ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC2012). They gained a recognition rate approximately 10% better than conventional image recognition technology, which only improved 2% from 2010 to 2011. This achievement marked a turning point for deep learning in the field of

*2 **API:** An interface that enables software functions to be used by another program.

*3 **Machine learning:** A technology that enables a computer to acquire knowledge and decision-making/action-taking criteria from data much like a human being acquires the same from sensory perception and experiences.

*4 **GPGPU:** The use of GPUs generally used for rendering and other types of image processing in computers for other types of applications. GPGPU excels at parallel distributed processing.

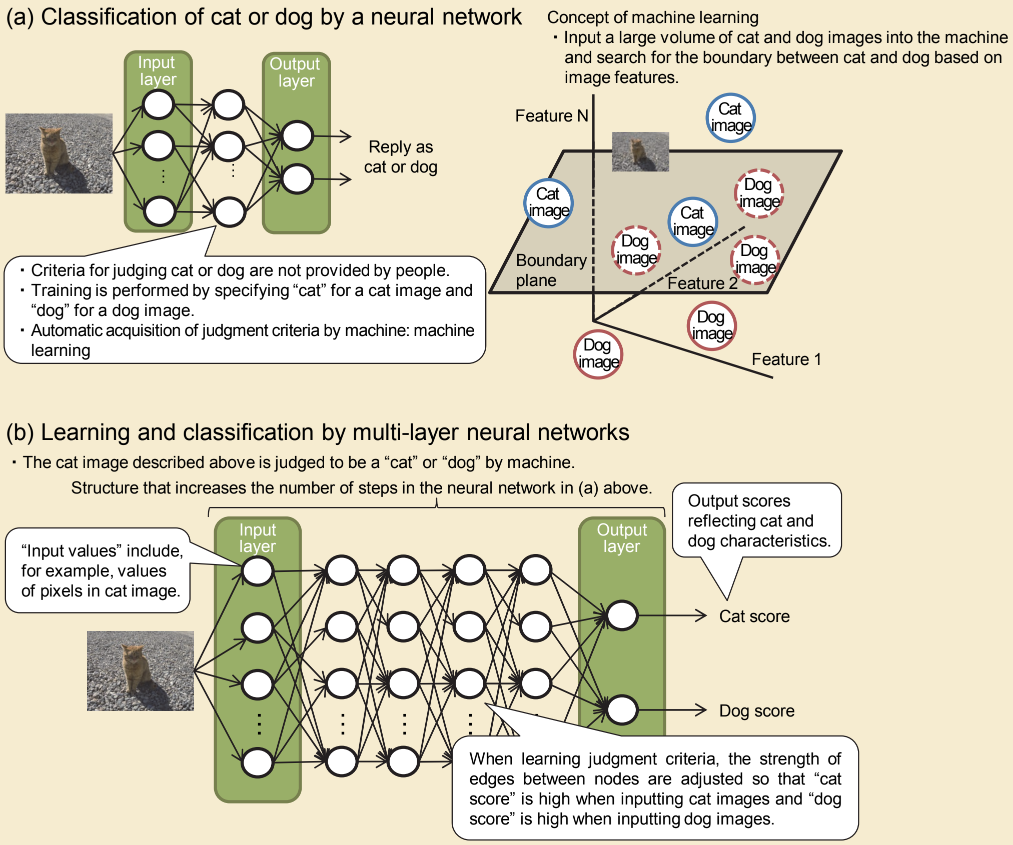

(a) Classification of cat or dog by a neural network
Input layer
Output layer
Reply as cat or dog
・Criteria for judging cat or dog are not provided by people.
・Training is performed by specifying "cat" for a cat image and "dog" for a dog image.
・Automatic acquisition of judgment criteria by machine: machine learning
Concept of machine learning
・Input a large volume of cat and dog images into the machine and search for the boundary between cat and dog based on image features.
Feature N
Cat image
Dog image
Boundary plane
Feature 2
Feature 1
(b) Learning and classification by multi-layer neural networks
・The cat image described above is judged to be a "cat" or "dog" by machine.
Structure that increases the number of steps in the neural network in (a) above.
"Input values" include, for example, values of pixels in cat image.
Input layer
Output layer
Output scores reflecting cat and dog characteristics.
Cat score
Dog score
When learning judgment criteria, the strength of edges between nodes are adjusted so that "cat score" is high when inputting cat images and "dog score" is high when inputting dog images.

Figure 2 Classification and recognition using machine learning technology based on neural networks

image recognition [2].

3. Differences between Conventional Technology and Image Recognition Using Deep Learning

1) Conventional Technology

Image recognition technology before deep learning had a basic two-step configuration as shown in **Figure 3** (a). In step 1, instead of using the image as-is, characteristics of an image is converted into quantifiable features (such as a histogram that represents what colors appear at what frequency or how brightness is distributed in the image). Then, in step 2, the image is classified and/or recognized based on those features. The judgment criteria for performing classification and recognition is usually acquired through machine learning (hereinafter, the module that performs classification and recognition by learning judgment criteria is referred to as a "recognizer"). After this learning step, the recognizer recognizes and/or classifies input images based on image features and learned criteria.

In such conventional technology, the image features in step 1 above were manually designed for each recognition task, such as image features appropriate for the detection of people, the recognition of human faces, etc. This process is usually called feature engineering. In the feature engineering step, researchers and developers should consider what features to focus on to give good classification results and what kind of algo-

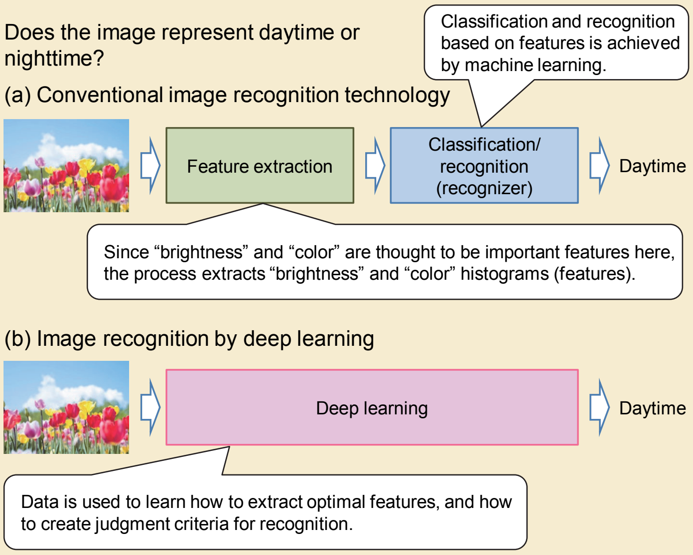

Does the image represent daytime or nighttime?
(a) Conventional image recognition technology
Classification and recognition based on features is achieved by machine learning.
Feature extraction
Classification/ recognition (recognizer)
Daytime
Since "brightness" and "color" are thought to be important features here, the process extracts "brightness" and "color" histograms (features).
(b) Image recognition by deep learning
Deep learning
Daytime
Data is used to learn how to extract optimal features, and how to create judgment criteria for recognition.

Figure 3 Differences between conventional technology and image recognition technology using deep learning

rithm is optimal.

It is difficult to engineer appropriate features for some tasks, as in the recognition of abstract concepts such as type of scene (wedding ceremony, field day, etc.) or category of object in the image ("food," "flowers," etc.). This situation made it tough to improve recognition accuracy.

2) Deep Learning

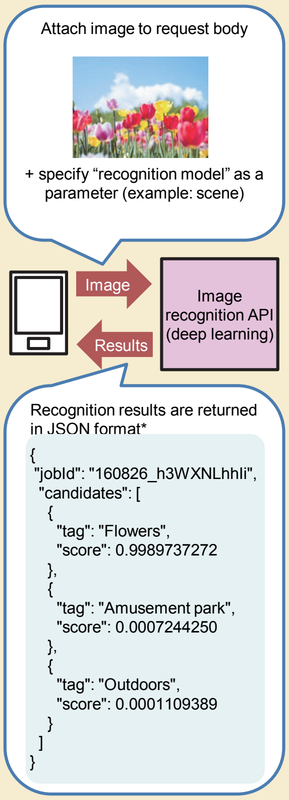
In contrast, image recognition using deep learning learns both appropriate features and recognition rules, as shown in Fig. 3 (b). Optimizing features to be used in recognition and creating recognition criteria based on those features are automatically done in the learning process. This approach enables the recognition of abstract concepts when it is not clear to decide which features to focus on and extract.

On the other hand, data is used not only to learn classification criteria in the final stage but also to learn feature extraction in the initial stage. This requires a huge amount of data for learning, which is a drawback of image recognition using deep learning. Various techniques have come into use to deal with this issue, including pre-training in which a deep learning recognizer is trained beforehand using a common large-scale image database such as ImageNet [7] and data augmentation in which the amount of training data is artificially increased.

4. Image Recognition API and Applications

In November 2015, "docomo Developer support" [2] publically released an image recognition API using the deep learning-based image recognition technologies mentioned above. This API provides several image recognition models, such as the model for scene recognition, for fashion recognition that can identify type, pattern, and color of a fashion item, and for other kinds of recognition. These models were trained from a huge amount of image data gathered by NTT DOCOMO and can predict suitable tags even for images which has problem to design appropriate features using conventional methods. docomo Developer support is a service which provides useful functions for developing applications and services. Anyone can use a number of APIs including the image recognition API based on deep learning by becoming a registered member of docomo Developer support and submitting a usage application.

Figure 4 shows how a developer of applications and services can use the image recognition API (category recog-


Attach image to request body
+ specify "recognition model" as a parameter (example: scene)
Image
Results
Image recognition API (deep learning)
Recognition results are returned in JSON format*

```
{
 "jobId": "160826_h3WXNLhhli",
  "candidates": [
    {
      "tag": "Flowers",
      "score": 0.9989737272
    },
    {
      "tag": "Amusement park",
      "score": 0.0007244250
    },
    {
      "tag": "Outdoors",
      "score": 0.0001109389
    }
  ]
}
```

*JavaScript Object Notation (JSON): A data description language based on object description in JavaScript.

Figure 4 Using the deep learning API in a service/application (example: scene recognition)

nition) released by docomo Developer support and **Figure 5** shows types of images that can be recognized by the API.

In the image recognition API (category recognition), docomo Developer support provides several trained deep learning models for each "recognition type" such as scene or fashion. Developers who would like to incorporate image recognition in their applications or services can select which models to use. In preparing such a model, NTT DOCOMO collected more than 1,000 images per tag for training purposes (here, a name or category such as "wedding ceremony" returned as a result of image recognition is called a "tag").

Users of docomo Developer support can immediately incorporate image recognition functions based on deep learning in their applications and services. Because training of these models has already been completed based on the large volume of image data, there is no need for users themselves to gather training data.

4.1 Applications of Scene Recognition

The scene recognition function can recognize the scene displayed in the image (such as wedding ceremony, field day, and birthday) and object categories (such as flower and food).

We can envision a variety of applications using this function, such as an application for saving images in cloud storage, an application for managing images on a smartphone, and an application for automatically creating a photo album. Recognizing images taken by a user and automatically attaching tags to those images simplifies image management for users.

Additionally, using this recognition function on image posting sites and Social Network Sites (SNSs) can reduce the workload in attaching tags when users post images.

4.2 Applications of Fashion Recognition

NTT DOCOMO uses image recognition technology based on deep learning as described above to achieve rapid fashion recognition. Given an input image, the fashion item category can be recognized and the image tagged accordingly.

The following four fashion recognition models are currently provided:

(1) Type: coat, cardigan, etc.

(2) Pattern: plain, border, etc.

(3) Color: pink, yellow, etc.

(4) Style: business, casual, etc.

This fashion recognition technology can be used to tag query images (images submitted by users) by the four models mentioned above and search for images with similar items (similarity search) based on these tags (type, pattern, color, etc.). The whole process is shown in **Figure 6**. Developers prepare beforehand a group of images of fashion items and attach tags (color, pattern, etc.) to each image. These images and tags are stored in a fashion database for displaying the results of a similarity search. Fashion recognition results are tags of a query image. Similar fashion items can

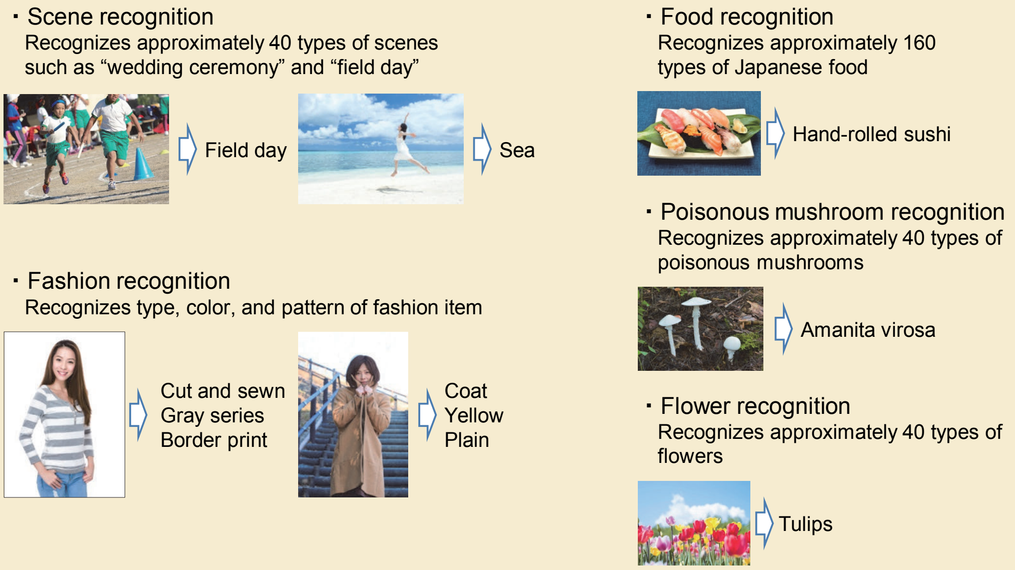

・Scene recognition
Recognizes approximately 40 types of scenes such as “wedding ceremony” and “field day”
Field day
Sea
・Fashion recognition
Recognizes type, color, and pattern of fashion item
Cut and sewn
Gray series
Border print
Coat
Yellow
Plain
・Food recognition
Recognizes approximately 160 types of Japanese food
Hand-rolled sushi
・Poisonous mushroom recognition
Recognizes approximately 40 types of poisonous mushrooms
Amanita virosa
・Flower recognition
Recognizes approximately 40 types of flowers
Tulips

Figure 5　Examples of image recognition when using the deep learning API in a service/application

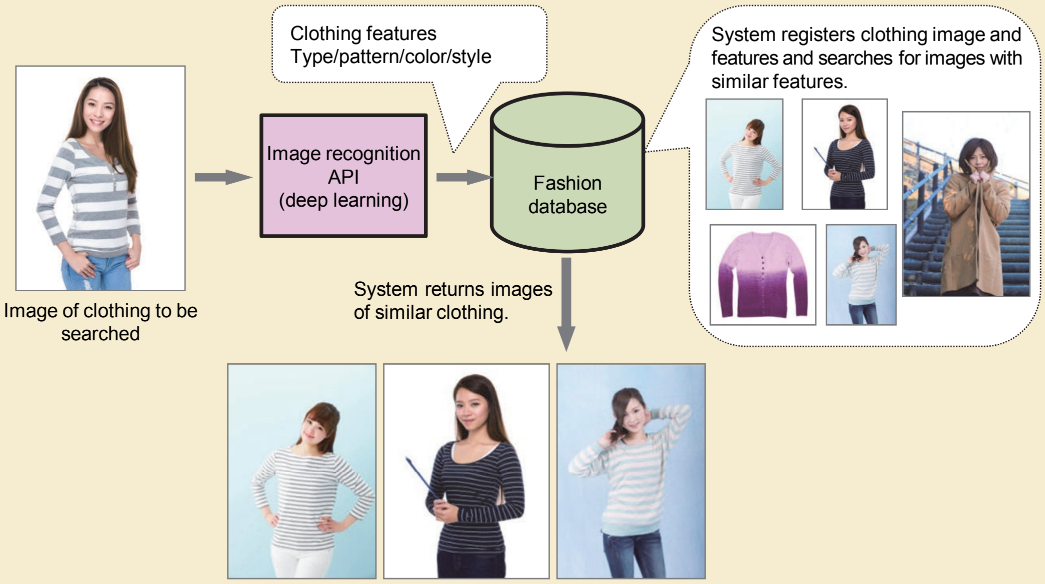

Clothing features
Type/pattern/color/style
System registers clothing image and features and searches for images with similar features.
Image recognition API (deep learning)
Fashion database
Image of clothing to be searched
System returns images of similar clothing.

Figure 6　Search for similar images

then be searched by comparing tags of the query image with tags of images in the fashion database.

A similarity search can be applied to various services. For example, let's assume that a person takes an interest in a clothing item appearing in a magazine or catalog, shown in a photo-sharing application (Instagram, etc.) for smartphones, or worn by the main character in a TV drama. Using a similarity search service, he/she only needs to take a photo of the clothing to search for a similar item without clarifying the clothing details. If the URL of Electronic Commerce (EC) site*5 for this item can also be provided, it can significantly reduce users' search time.

A similarity search can also be used to discover a new way of coordinating clothes. This could benefit users who would like to select clothes that match well with what they already have, users who always purchase the same type of clothes, or users are not sure how to match and dress in appropriate ways thereby having little chance to wear the clothes they bought. When a user takes a photo for a clothing item, the similarity search service finds similar clothes and recommends a wide variety of clothing items that coordinate properly. Users can decrease wasteful purchasing and greatly reduce time spent in searching for compatible clothing items. It's also an enjoyable way for users to select clothes at home, in a shopping mall, or even on a train.

5. Expanded Application of Image Recognition by Deep Learning

In future image recognition based on deep learning, we envision the expanded use of tags beyond the simple ones used in the recently released API. For example, studies are being performed on predicting the emotion (such as anger or sadness) evoked on seeing an image (Fig. 1 (b)) [8]. Researches are also progressing on video recognition—a technique for attaching a descriptive sentence as a tag to video has been proposed (Fig. 1 (c)) [9].

6. Conclusion

In this article, we presented an overview of deep learning technology, explained how it differs from conventional image recognition technology, described the features of an image recognition API service developed and offered by NTT DOCOMO, and introduced applications using this API.

Research on applying deep learning to fields other than image recognition is moving forward. There are studies on using deep learning in natural language processing and machine translation, in marketing, and in content recommendation on the Web. Deep learning is becoming an essential technology in all sorts of scenarios involving the analysis and use of data.

In the future, NTT DOCOMO will undertake a successive expansion of objects which can be recognized by image recognition APIs using deep learning. NTT DOCOMO will also focus on the development of recognition techniques targeting data other than images and new recognition technologies which combine images and other types of data.

References

[1] J. Dean: "Large Scale Deep Learning." http://on-demand.gputechconf.com/gtc/2015/presentation/S5817-Keynote-Jeff-Dean.pdf

[2] A. Krizhevsky, I. Sutskever and G. E. Hinton: "ImageNet Classification with Deep Convolutional Neural Networks," Advances in Neural Information Processing Systems, 25, pp.1097-1105, 2012.

[3] NTT DOCOMO: "Image Recognition｜docomo Developer support｜NTT DOCOMO," (in Japanese). https://dev.smt.docomo.ne.jp/?p=docs.api.page&api_docs_id=102

[4] H. Akatsuka et al.: "High-speed, Large-scale Image Recognition and API," NTT DOCOMO Technical Journal, Vol.17, No.1, pp.10-17, Jul. 2015.

[5] F. Rosenblatt: "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain," Psychological Review, Vol.65 (6), pp.386-408, Nov. 1958.

[6] K. Fukushima: "Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position," IEICE Transactions A, Vol.J62-A, No.10, pp.658-665, Oct. 1979 (in Japanese).

[7] Stanford Vision Lab, Stanford University, Princeton University: "ImageNet." http://image-net.org/

[8] K. Peng, T. Chen, A. Sadovnik and A. Gallagher: "A mixed bag of emotions: Model, predict, and transfer emotion

*5 **EC site:** A web site that sells products and/or services.

distributions," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.860-868, 2015.

[9] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko and T. Darrell: "Long-term Recurrent Convolutional Networks for Visual Recognition," CoRR, abs/1411.4389, 2014.