US 20230186600A1

(19) **United States**

(12) **Patent Application Publication** (10) **Pub. No.: US 2023/0186600 A1**

NGUYEN et al. (43) **Pub. Date: Jun. 15, 2023**

(54) **METHOD OF CLUSTERING USING ENCODER-DECODER MODEL BASED ON ATTENTION MECHANISM AND STORAGE MEDIUM FOR IMAGE RECOGNITION**

(71) Applicant: **VINAI ARTIFICIAL INTELLIGENCE APPLICATION AND RESEARCH JOINT STOCK COMPANY**, Ha Noi (VN)

(72) Inventors: **Xuan Bac NGUYEN**, Ha Noi (VN); **Duc Toan BUI**, Ha Noi (VN); **Hai Hung BUI**, Ha Noi (VN)

(21) Appl. No.: **17/894,988**

(22) Filed: **Aug. 24, 2022**

(30) **Foreign Application Priority Data**

Dec. 9, 2021 (VN) 1-2021-07930

Publication Classification

(51) **Int. Cl.**

G06V 10/762	(2006.01)
G06V 10/77	(2006.01)
G06V 10/82	(2006.01)
G06F 7/24	(2006.01)

(52) **U.S. Cl.**

CPC ***G06V 10/762*** (2022.01); ***G06V 10/7715*** (2022.01); ***G06V 10/82*** (2022.01); ***G06F 7/24*** (2013.01)

(57) **ABSTRACT**

A method of clustering using encoder-decoder model based on attention mechanism extracts image features, clusters to form image feature vector clusters, and based on the cosine similarity score between the image feature vectors to arrange each image feature vector cluster into an image feature vector sequence. The image feature vector sequence includes cosine distance encoding vectors concatenated with respective image feature vectors and is used as the input data sequence in encoder and decoder neural network models to generate an output data sequence from the input data sequence. The output data sequence is a binary sequence having values of 1 or 0 at a position denoting that the image corresponding to the position is or is not in the same cluster with respect to the center image of the cluster.

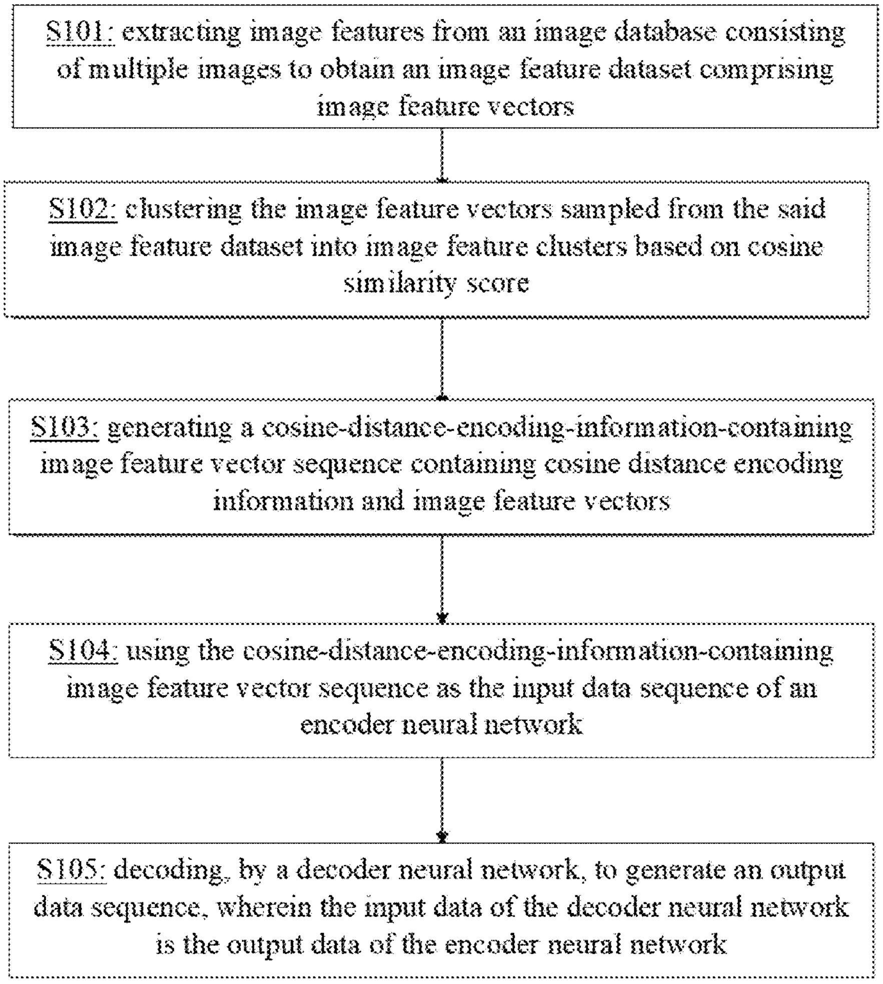

S101: extracting image features from an image database consisting of multiple images to obtain an image feature dataset comprising image feature vectors
S102: clustering the image feature vectors sampled from the said image feature dataset into image feature clusters based on cosine similarity score
S103: generating a cosine-distance-encoding-information-containing image feature vector sequence containing cosine distance encoding information and image feature vectors
S104: using the cosine-distance-encoding-information-containing image feature vector sequence as the input data sequence of an encoder neural network
S105: decoding, by a decoder neural network, to generate an output data sequence, wherein the input data of the decoder neural network is the output data of the encoder neural network

FIG. 1

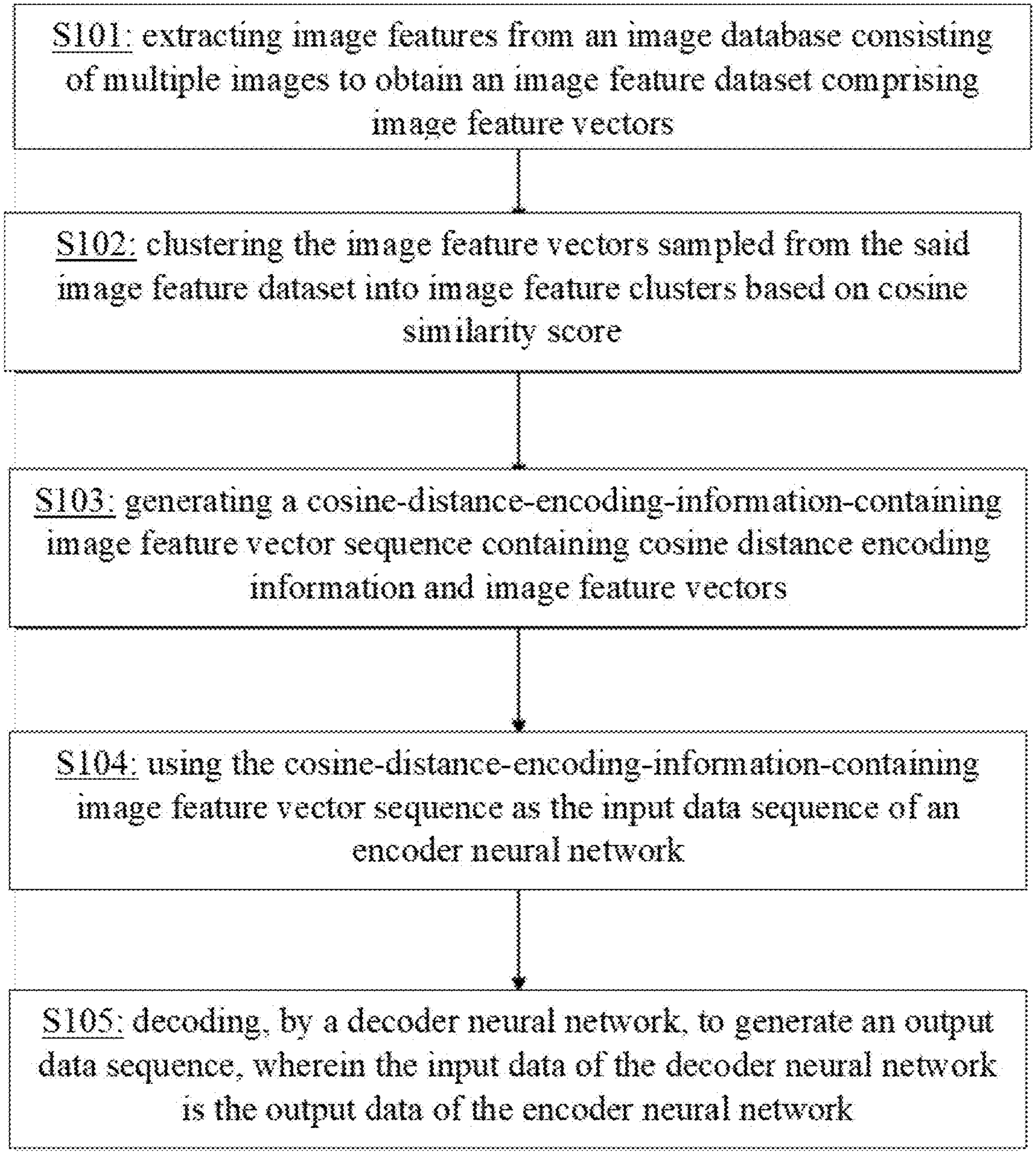

S101: extracting image features from an image database consisting of multiple images to obtain an image feature dataset comprising image feature vectors
S102: clustering the image feature vectors sampled from the said image feature dataset into image feature clusters based on cosine similarity score
S103: generating a cosine-distance-encoding-information-containing image feature vector sequence containing cosine distance encoding information and image feature vectors
S104: using the cosine-distance-encoding-information-containing image feature vector sequence as the input data sequence of an encoder neural network
S105: decoding, by a decoder neural network, to generate an output data sequence, wherein the input data of the decoder neural network is the output data of the encoder neural network

FIG. 2

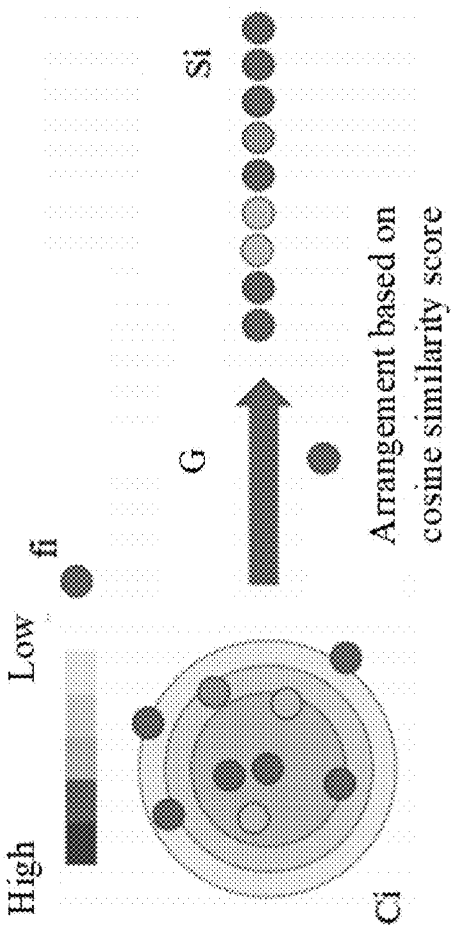
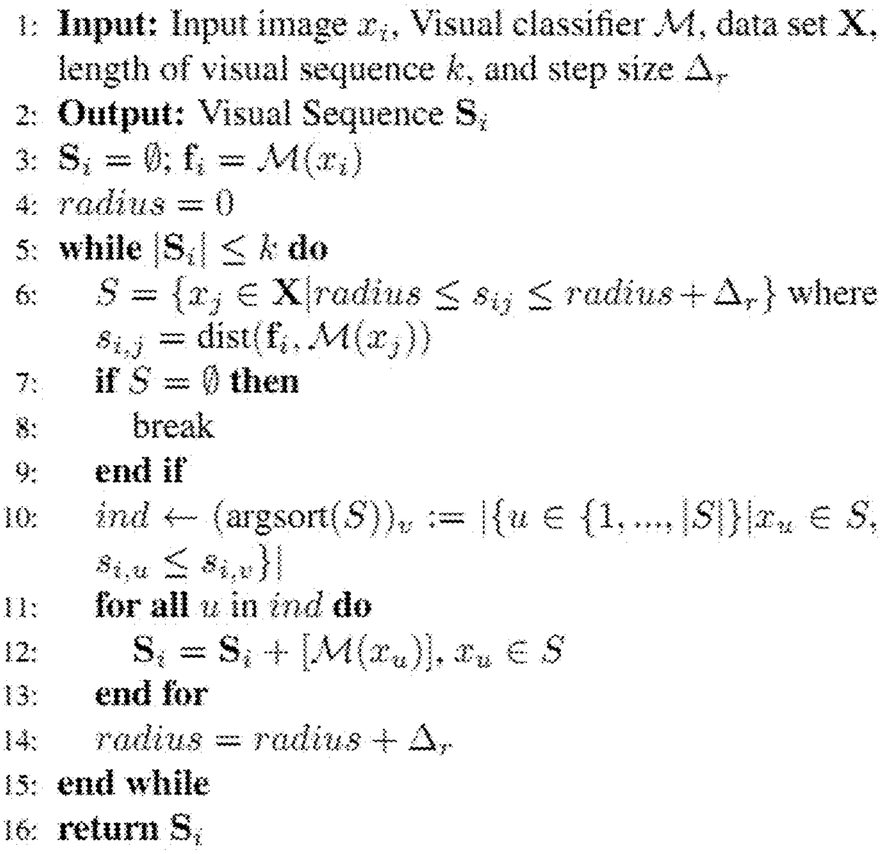
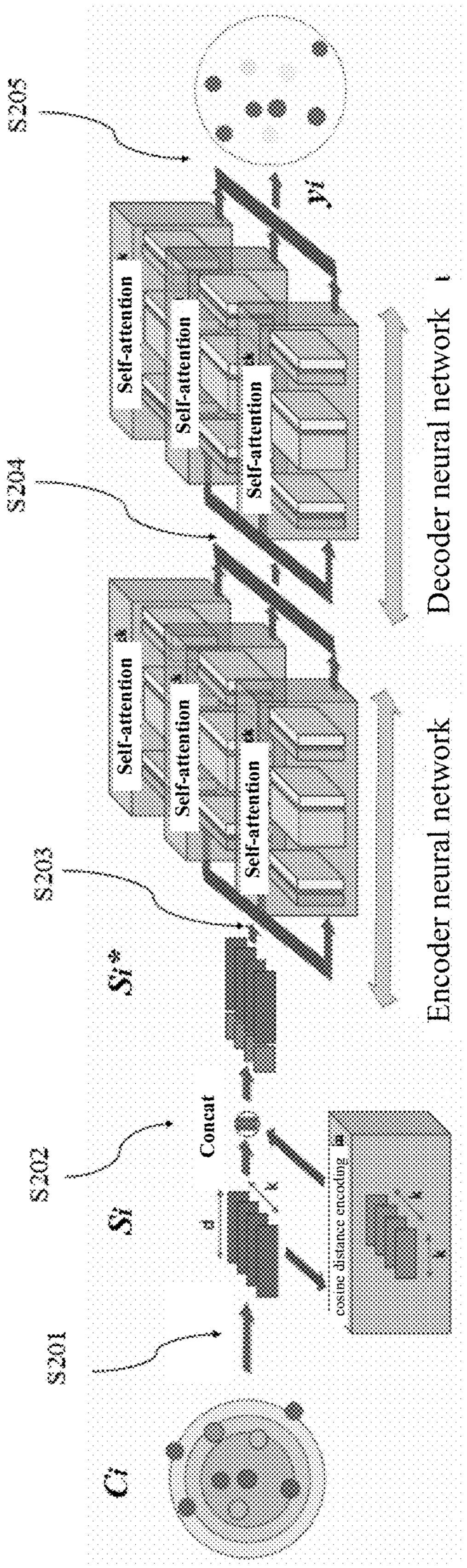

High
Low
fi
G
Si
Ci
Arrangement based on cosine similarity score

FIG. 3



```
Input: Input image $x_i$, Visual classifier $\mathcal{M}$, data set $\mathbf{X}$, length of visual sequence $k$, and step size $\Delta_r$
Output: Visual Sequence $\mathbf{S}_i$
$\mathbf{S}_i = \emptyset$; $\mathbf{f}_i = \mathcal{M}(x_i)$
$radius = 0$
while $|\mathbf{S}_i| \leq k$ do
    $S = \{x_j \in \mathbf{X} | radius \leq s_{ij} \leq radius + \Delta_r\}$ where $s_{i,j} = \text{dist}(\mathbf{f}_i, \mathcal{M}(x_j))$
    if $S = \emptyset$ then
        break
    end if
    $ind \leftarrow (\text{argsort}(S))_v := |\{u \in \{1, ..., |S|\} | x_u \in S, s_{i,u} \leq s_{i,v}\}|$
    for all $u$ in $ind$ do
        $\mathbf{S}_i = \mathbf{S}_i + [\mathcal{M}(x_u)], x_u \in S$
    end for
    $radius = radius + \Delta_r$
end while
return $\mathbf{S}_i$
```

FIG. 4


C_i
S201
S_i
d
k
S202
Concat
cosine distance encoding
S_i*
S203
Self-attention
Self-attention
Self-attention
Encoder neural network
S204
Self-attention
Self-attention
Self-attention
Decoder neural network
t
y_i
S205

METHOD OF CLUSTERING USING ENCODER-DECODER MODEL BASED ON ATTENTION MECHANISM AND STORAGE MEDIUM FOR IMAGE RECOGNITION

CROSS-REFERENCE TO RELATED APPLICATION

[0001] The present application claims priority from Vietnamese Application No. 1-2021-07930 filed on Dec. 9, 2021, which is incorporated herein by reference in its entirety.

TECHNICAL FIELD

[0002] The present invention belongs to the field of artificial intelligence and refers to a method of clustering using encoder-decoder model based on attention mechanism, and to a storage medium comprising a computer program to perform the method. More particularly, the method of clustering using encoder-decoder model based on attention mechanism generates an input data sequence from the image feature clusters using the information on the cosine similarity score, for decoding into an output data sequence through encoder and decoder neural networks, wherein each position in the output data sequence may correspond to one image, and based on a value at a position in the output data sequence to recognize or classify the image.

RELATED ART

[0003] The image recognition technique, more particularly, classification (or clustering) of a human face image or a landmark image (which may be generally referred to as visual classification), has been gaining great concerns in machine learning. The solutions therefor, such as human face image or landmark image classification, may be divided into three main groups, including non-supervised learning visual classification, semi-supervised learning visual classification, and supervised learning visual classification.

[0004] Based on that it is easy to collect features of the visual data, in practice it is thus possible to access huge databases of visual images. However, it could be said that the exploitation of the information from these visual images is relatively difficult, such as in notation (e.g., extraction of the features in an image for presentation as image-associated information, image recognition, image classification, image clustering, or the like), for example, the reason being that there are too many complicated factors that may influence the visual images, e.g., brightness, shooting poses, depended on practical shooting circumstances. Therefore, it is essentially important and necessary to study, propose, and provide parameterized models for performing information exploitation of the visual images, e.g., visual image classification, whose performance is substantively enhanced.

[0005] One of the widely known models includes GCN networks (Graph Convolutional Network), which solve the problem of visual classification by way of non-supervised learning. The GCN networks uses the same similarity concepts in spectral graph theory to design parameterized extractors as suitable in the CNN networks (Convolutional Neural Network), and has been shown as one of the most efficient methods in solving the classification of complicated samples. Some examples of GCN networks were disclosed in the journals of the titles "Learning to cluster faces via confidence and connectivity estimation", by Lei Yang et al, published in the Proceedings of IEEE Conference on computer vision and pattern recognition, 2020; and "Learning to cluster faces on an affinity graph", by Lei Yang et al, published in the Proceedings of IEEE Conference on computer vision and pattern recognition, 2019.

[0006] In general, GCN networks aim at generating affinity graphs, using the image feature vectors sampled in the visual image database as the vertices, wherein the adjacent vertices are joined together based on the cosine similarity score between the image feature vectors. The graph with said similarity is usually a large-scale graph, which may contain millions of graph vertices, and thus the GCN networks are assumed to have a large computational volume, and require high memory usage. In addition, these networks are quite sensitive to hard and noisy samples.

[0007] Therefore, there is a demand for an improved solution in association with image recognition, which may minimize the requirements of memory usage and computational volume, and achieve great results even with hard and noisy samples.

SUMMARY

[0008] The object of the present invention is to provide a method of clustering using encoder-decoder model based on attention mechanism, which may overcome one or some of the above-mentioned problems.

[0009] Another object of the present invention is to provide a method of clustering using encoder-decoder model based on attention mechanism, which may technically reduce the requirements of memory usage and computational volume, and technically achieve great results even with hard and noisy samples.

[0010] It should be understood that the present invention is not limited to the above-described objects. In addition to these objects, the present invention may also include others that will be obvious to the ordinary person, specified or encompassed in the description below.

[0011] To achieve one or some of the above objects, the present invention provides a method of clustering using encoder-decoder model based on attention mechanism, the method comprising:

[0012] extracting image features from an image database X consisting of multiple images x_i, by an image feature extracting model, to obtain an image feature dataset comprising image feature vectors f_i, wherein each image feature vector f_i corresponds to one image x_i in the said image database X;

[0013] clustering the image feature vectors sampled from the said image feature dataset into image feature clusters C_i based on the cosine similarity scores $s_{i,j}$ between the image feature vectors f_i and f_j, wherein each image feature cluster C_i has a center image feature vector;

[0014] arranging the image feature vectors f_i in each image feature cluster C_i into an image feature vector sequence S_i in an ascending or descending order based on the cosine similarity scores $s_{i,j}$ of the image feature vectors compared with the center image feature vector in the same said image feature cluster C_i;

[0015] generating cosine distance encoding vectors e_t from the components such as the cosine similarity scores $s_{i,j}$ where each cosine distance encoding vector e_t corresponds to an image feature vector f_t in the image feature cluster C_i and the cosine similarity scores $s_{i,j}$ forming the cosine distance encoding vector e_t are the cosine similarity scores between the image feature vectors and the image feature vector f_t in the same said image feature cluster C_i;

[0016] concatenating the image feature vector f_t with the respective cosine distance encoding vector e_t to form a respective cosine-distance-encoding-information-containing image feature vector f_t*;

[0017] generating a cosine-distance-encoding-information-containing image feature vector sequence S_i* in correspondence with each image feature cluster C_i, where the components of the cosine-distance-encoding-information-containing image feature vector sequence S_i* are the cosine-distance-encoding-information-containing image feature vectors f_t* in correspondence with the image feature vectors f_t belonging to a respective image feature cluster C_i, and the cosine-distance-encoding-information-containing image feature vectors f_t* are arranged in an ascending or descending order based on the cosine similarity scores $s_{i,j}$ of the image feature vectors compared with the center image feature vector in the same said image feature cluster C_i;

[0018] using the cosine-distance-encoding-information-containing image feature vector sequence S_i* as the input data sequence of an encoder neural network, wherein the encoder neural network is configured to generate a respective encoded representation for each input in the input data sequence by using an attention mechanism, which shows the attention in the encoded representations of the input data sequence;

[0019] decoding, by a decoder neural network, to generate an output data sequence, wherein the decoder neural network is configured to receive the encoded representations as the input data for decoding into the output data sequence.

[0020] According to an embodiment, the step that the encoder neural network generates a respective encoded representation for each input in the input data sequence by using an attention mechanism, which shows the attention in the encoded representations of the input data sequence comprises:

[0021] projecting the cosine-distance-encoding-information-containing image feature vector sequence S_i* into at least one sub-space, wherein for each sub-space, perform the operations of:

[0022] determining first, second, and third trainable matrices;

[0023] projecting the cosine-distance-encoding-information-containing image feature vector sequence S_i* into first, second, and third super spaces to generate first, second, and third super space features based on the first, second, and third trainable matrices, respectively;

[0024] calculating the attention scores $r_{i,j}$ between the cosine-distance-encoding-information-containing image feature vector f_i* and the cosine-distance-encoding-information-containing image feature vectors f_j* in the cosine-distance-encoding-information-containing image feature vector sequence S_i* based on the first super space feature of the cosine-distance-encoding-information-containing image feature vector f_i* and the second super space features of the cosine-distance-encoding-information-containing image feature vectors f_j*; and

[0025] generating a sub-space output of the cosine-distance-encoding-information-containing image feature vector f_i* by calculating a weighted sum of the third super space features of the cosine-distance-encoding-information-containing image feature vectors f_j*, wherein the weights assigned to the third super space features are respective attention scores $r_{i,j}$;

[0026] linearly transforming a concatenation result of the sub-space outputs of the cosine-distance-encoding-information-containing image feature vector sequence S_i* in correspondence with each sub-space to obtain an attention output; and

[0027] generating the encoded representations based on the attention output.

[0028] Preferably, the output data sequence of the decoder neural network is a binary sequence y_i with a length in correspondence with that of the cosine-distance-encoding-information-containing image feature vector sequence S_i*, where the value at the t^{th} position of the binary sequence y_i being 1 denotes the t^{th} image feature vector has the same label as the center image feature vector in the same said image feature cluster C_i, and the value at the t^{th} position of the binary sequence y_i being 0 denotes the t^{th} image feature vector does not have the same label as the center image feature vector in the same said image feature cluster C_i.

[0029] The cosine distance encoding vector e_t is determined through the following expression:

$$e_t = \{s_{t,i}\}_{i=1}^{k}$$

[0030] where $s_{t,i}$ is the cosine similarity score between the i^{th} image feature vector and the t^{th} image feature vector.

[0031] The cosine-distance-encoding-information-containing image feature vector f_t* is determined through the following expression:

$$f_t^* = \text{concat}(f_t, e_t)$$

[0032] where concat is a function that concatenates two vectors into one vector.

[0033] The said encoder neural network and decoder neural network are trained using a target function which is determined through the following expression:

$$\mathcal{L}_i(\hat{y}_i, y_i) = -\sum_{t=1}^{k}\left[y_i^t \times \log\left(\sigma(\hat{y}_i^t)\right) + \left(1 - y_i^t\right) \times \log\left(1 - \sigma(\hat{y}_i^t)\right)\right]$$

[0034] where σ is the sigmoid function.

[0035] Preferably, the said trained image feature extracting model uses two datasets consisting of a labeled dataset D_L, and an unlabeled dataset D_U.

[0036] In another aspect, the present invention provides a storage medium comprising a computer program which includes instructions that, when executed, will cause the computer to perform the said method of image cluttering.

BRIEF DESCRIPTION OF THE DRAWINGS

[0037] FIG. **1** is a flowchart showing a method of clustering using encoder-decoder model based on attention mechanism according to a preferred embodiment of the present invention;

[0038] FIG. **2** is a schematic diagram showing a way to rearrange the image feature clusters into a sequence according to a preferred embodiment of the present invention;

[0039] FIG. **3** is a screenshot representing a routine showing a sorter G; and

[0040] FIG. **4** is a block diagram showing a method of clustering using encoder-decoder model based on attention mechanism according to a preferred embodiment of the present invention.

DETAILED DESCRIPTION

[0041] Below, the advantages, effects, and substance of the present invention may be explained through the detailed description of preferred embodiments with reference to the appended figures. However, it should be understood that these embodiments are only described by way of example to clarify the spirit and advantages of the present invention, without limiting the scope of the present invention according to the described embodiments.

[0042] In general, as described below, the method of clustering using encoder-decoder model based on attention mechanism aims at classifying the images or visual data, e.g., recognizing whether the images belong to the same cluster, more particularly, whether human face photos are of the same shooting poses, whether the landmark images are of the same photos of lakes, old castles, etc., for example. However, it should be understood that the techniques or principles in accordance with the present invention are not limited to image or visual data classification, but may be applied in a variety of image recognition applications, such as annotation or labeling for images or visual data.

[0043] FIG. **1** represents a method of clustering using encoder-decoder model based on attention mechanism according to a preferred embodiment of the present invention.

[0044] As shown in the figure, the method of clustering using encoder-decoder model based on attention mechanism according to the preferred embodiment comprises the steps described below.

[0045] Step S**101**: extracting image features from an image database to obtain image feature vectors.

[0046] Herein, for ease of description and representation, the image database is referred to as the image database X consisting of multiple images x_i, the image feature vectors are referred to as the image feature vectors f_i, with each image feature vector f_i in correspondence with one image x_i in the said image database X.

[0047] In this step, the extracting of the image features from the image database X consisting of multiple images x_i may be performed through an image feature extracting model M. Using the image feature extracting model M, the input image x_i belonging to the image database (e.g., with the dimension of h×w×3, wherein h denotes the height, and w denotes the width of image) is introduced into the image feature extracting model M to extract or capture visual features.

[0048] In a particular example, the visual features are image feature vectors f_i with the dimension of 1×d, wherein d is the feature dimension of each extracted image by the image feature extracting model M. For convenience, the image feature vectors f_i may be represented as $f_i=M(x_i)$.

[0049] In general, the image feature extracting models are already known, and widely used, such as CNN models, for example. Specific description of the image feature extracting models is intended for omission to focus into allegedly more important contents of the present invention.

[0050] According to a preferred embodiment, the use of the image feature extracting model M is maximized on efficiency by using two datasets consisting of a labeled dataset D_L, and an unlabeled dataset D_U during training for the image feature extracting model M. First, the image feature extracting model M is trained using the labeled dataset D_L by way of typical supervised learning. Then, the image feature extracting model M after being trained by the labeled dataset D_L is used to trigger extracted training samples in the unlabeled dataset D_U. The training may correspond to semi-supervised learning, wherein the unlabeled dataset D_U is much greater than the labeled dataset D_L.

[0051] Step S**102**: clustering the image feature vectors sampled from the said image feature dataset into image feature clusters based on the cosine similarity score.

[0052] Herein, for ease of description and representation, the image feature cluster is referred to as an image feature cluster C_i, the cosine similarity score is referred to as a cosine similarity score $s_{i,j}$, the cosine similarity score is the cosine similarity score $s_{i,j}$ between the image feature vectors f_i and f_j.

[0053] The cosine similarity score between two vectors is a known mathematical feature between two vectors. For example, the cosine similarity score between two vertices v_i, v_j of the similarity graph represented by a adjacency matrix W, which is the cosine of the angle between two vectors consisting of a vector denoted by the i^{th} row and j^{th} row of the adjacency matrix W, denoted as W^i, W^j.

[0054] The cosine similarity score is determined as follows:

$$\sigma_{ij} = \frac{W^i . W^j}{\|W^i\| \|W^j\|}.$$

[0055] According to a preferred embodiment, clustering of the said image feature vectors is performed by using a k-nearest neighbors algorithm based on the cosine similarity score, referred to as a k-nearest neighbors model K.

[0056] Each image feature cluster C_i has a center image feature vector f_i, and may be represented as $C_i=K(f_i, F, k)$, wherein F=M(X) is a feature subset extracted from the image database X, and k is the number of nearest neighbors.

[0057] The image feature clusters C_i form a set of image feature clusters C, which may be represented as $C=\{C_i\}_{i=1}^{N}$.

[0058] Step S**103**: generating a cosine-distance-encoding-information-containing image feature vector sequence consisting of cosine distance encoding information and image feature vectors.

[0059] According to a preferred embodiment, a cosine-distance-encoding-information-containing image feature vector sequence consisting of cosine distance encoding information and image feature vectors is a cosine-distance-encoding-information-containing image feature vector sequence S_i* in correspondence with each image feature cluster C_i, wherein the components of the cosine-distance-encoding-information-containing image feature vector sequence S_i* are the cosine-distance-encoding-information-containing image feature vectors f_t* in correspondence with the image feature vectors f_t belonging to a respective image feature cluster C_i, and the order of the cosine-distance-encoding-information-containing image feature vectors f_t* is based on the ascending or descending cosine similarity scores $s_{i,j}$ of the image feature vectors compared with the center image feature vector in the same said image feature cluster C_i.

[0060] In order to generate cosine-distance-encoding-information-containing image feature vectors f_t*, firstly the cosine distance encoding vectors e_t are formed from the components such as the cosine similarity scores $s_{i,j}$, wherein each cosine distance encoding vector e_t corresponds to an image feature vector f_t in the image feature cluster C_i, and the cosine similarity scores $s_{i,j}$ forming the cosine distance encoding vector e_t are the cosine similarity scores between the image feature vectors and the image feature vector f_t in the same said image feature cluster C_i; Then, concatenating the image feature vector f_t with the respective cosine distance encoding vector e_t to form a respective cosine-distance-encoding-information-containing image feature vector f_t*.

[0061] According to a preferred embodiment, the cosine-distance-encoding-information-containing image feature vector sequence S_i* is generated by:

[0062] arranging the image feature vectors f_t in each image feature cluster C_i into an image feature vector sequence S_i in an ascending or descending order based on the cosine similarity scores $s_{i,j}$ of the image feature vectors compared with the center image feature vector in the same said image feature cluster C_i, and

[0063] concatenating the image feature vector f_t with the respective cosine distance encoding vector e_t, at each position of the t^{th} image feature vector in the image feature vector sequence S_i, to form a cosine-distance-encoding-information-containing image feature vector f_t* to form the said cosine-distance-encoding-information-containing image feature vector sequence S_t*.

[0064] It should be understood that the present invention is not limited to the preferred embodiments, however, the cosine-distance-encoding-information-containing image feature vectors S_i* may be generated without generating the image feature vector sequence S_i, e.g., the cosine-distance-encoding-information-containing image feature vectors f_t* may be generated first, and then arranged into a sequence to form a cosine-distance-encoding-information-containing image feature vector sequence S_i*, for example.

[0065] According to a preferred embodiment, as shown in FIG. **2**, the image feature vector sequence S_i generated from the image feature cluster C_i consists of an image feature vectors f_i as the center through arrangement by sorter G, represented as S_i=G(C_i).

[0066] According to the preferred embodiment, the cosine similarity scores $s_{i,j}$ between the image feature vectors f_j in a cluster with the image feature vector f_i as the center of the cluster are calculated, and the image feature vectors f_j will be arranged in a descending order of the cosine similarity scores to form the sequence.

[0067] In order to provide more coherent information, a screenshot of the routine presenting the sorter G is shown in FIG. **3** for reference.

[0068] According to a preferred embodiment, the cosine distance encoding vector e_t is determined through the following expression:

$$e_t = \{s_{t,i}\}_{i=1}^{k}$$

[0069] wherein $s_{t,i}$ is the cosine similarity score between the i^{th} image feature vector and the t^{th} image feature vector.

[0070] The cosine-distance-encoding-information-containing image feature vector f_t* is determined through the following expression:

$$f_t^* = \text{concat}(f_t, e_t)$$

[0071] where concat is a function that concatenates two vectors into one vector.

[0072] Step **S104**: using the cosine-distance-encoding-information-containing image feature vector sequence as the input data sequence of an encoder neural network.

[0073] In this step, the encoder neural network is configured to generate a respective encoded representation for each input in the input data sequence by using an attention mechanism, which shows the attention in the encoded representations of the input data sequence. In order to capture the attention in the encoded representations of the input data sequence, the cosine-distance-encoding-information-containing image feature vector sequence S_i* is projected into at least one sub-space. For each sub-space, the first, second, and third trainable matrices are determined. The first, second, and third trainable matrices may, respectively, be referred to as query matrix $W^Q \in R^{d \times d'}$, key matrix $W^K \in R^{d \times d'}$, and value matrix $W^V \in R^{d \times d'}$. Then, the cosine-distance-encoding-information-containing image feature vector sequence S_i* is projected into first, second, and third super spaces (referred to as query, key and value super spaces, respectively) to generate first, second, and third super space features (referred to as query super space feature Q, key super space feature K and value super space feature V, respectively) based on the first, second, and third trainable matrices, respectively, according to Equations:

$$Q = S_i * W^Q, Q \in R^{d \times d'}$$

$$K = S_i * W^K, K \in R^{d \times d'}$$

$$V = S_i * W^v, V \in R^{d \times d'}$$

[0074] Then, the attention scores $r_{i,j}$ between the cosine-distance-encoding-information-containing image feature vector f_i* and the cosine-distance-encoding-information-containing image feature vectors f_j* in the cosine-distance-encoding-information-containing image feature vector sequence S_i* are calculated based on the first super space feature of the cosine-distance-encoding-information-containing image feature vector f_i* and the second super space features of the cosine-distance-encoding-information-containing image feature vectors f_j* according to Equation:

$$r_{i,j} = \frac{\textcircled{?}}{\sum_{j=1}^{k} \textcircled{?}}$$

$\textcircled{?}$ indicates text missing or illegible when filed

[0075] The sub-space output Z_i of the cosine-distance-encoding-information-containing image feature vector f_i* is generated by calculating a weighted sum of the third super space features V_j of the cosine-distance-encoding-information-containing image feature vectors f_j*, wherein the weights assigned to the third super space features are respective attention scores $r_{i,j}$:

$$Z_i = \Sigma_{j=1}^{k} r_{i,j} \cdot V_j.$$

$$Z = Att(Q, K, V) = \{Z_i\}_{i=1}^{k}$$

[0076] If the sub-space number is m, then the feature dimension of each sub-space is

$$d' = \frac{d}{m}.$$

The sub-space outputs of the cosine-distance-encoding-information-containing image feature vector sequence S_i* in correspondence with each sub-space $Z_{s,i}$ are calculated as follows:

$$Z_{s,i}=Att(Q_{s,i},K_{s,i},V_{s,i}),1\leq i\leq m$$

[0077] A concatenation result of the sub-space outputs of the cosine-distance-encoding-information-containing image feature vector sequence S_i* in correspondence with each sub-space is linearly transformed to obtain an attention output Z_M:

$$Z_M=\text{concat}(Z_{s,1}, \ldots ,Z_{s,m})\cdot W^M$$

[0078] Where W^M is an additional weight matrix.

[0079] The encoder neural network may generate encoded representations based on the attention output Z_M. According to an embodiment of the present invention, the encoder neural network includes a point-wise feed forward network (FFN) to receive the attention output Z_M.

[0080] Step S**105**: decoding, by a decoder neural network, to generate an output data sequence, wherein the input data of the decoder neural network are the output data of the encoder neural network.

[0081] Herein, the input data of the decoder neural network or the output data of the encoder neural network are encoded representations.

[0082] According to a preferred embodiment, the output data sequence of the decoder neural network is a binary sequence y_i with a length in correspondence with that of the cosine-distance-encoding-information-containing image feature vector sequence S_i*, where the value at the t^{th} position of the binary sequence y_i being 1 denotes the t^{th} image feature vector has the same label as the center image feature vector in the same said image feature cluster C_i, and the value at the t^{th} position of the binary sequence y_i being 0 denotes the t^{th} image feature vector does not have the same label as the center image feature vector in the same said image feature cluster C_i.

[0083] The said encoder neural network and decoder neural network are trained using the target function which may be determined through the following expression:

$$\mathcal{L}_i(\hat{y}_i,y_i) = -\sum_{t=1}^{k}\left[y_i^t\times\log(\sigma(\hat{y}_i^t)) + (1-y_i^t)\times\log(1-\sigma(\hat{y}_i^t))\right]$$

[0084] where σ is the sigmoid function.

[0085] In general, the said encoder and decoder neural networks are already known and may be similarly applied as the encoder and decoder neural networks used in attention-based encoder-decoder models. An example of models in this form is provided in a journal with the title "Attention is all you need", by Ashish Vaswani et al. Another example of models in this form is provided in U.S. Ser. No. 10/452,978 B2, U.S. Ser. No. 10/719,764 B2, U.S. Ser. No. 10/839,259 B2, U.S. Ser. No. 10/956,819 B2. The whole contents of the documents are intended to be introduced herein for reference and may be incorporated into the solution provided in accordance with the present invention by any known means.

[0086] The features and operational principles of the encoder and decoder neural networks of the present invention are completely similar to those of the encoder and decoder neural networks provided or used in said journals and patent documents. Thus, specific descriptions of the encoder and decoder neural network is intended for omission to focus into allegedly more important contents of the present invention.

[0087] As shown in FIG. **4**, the method of clustering using encoder-decoder model based on attention mechanism is illustrated through steps S**201**-S**205**, described in greater detail below.

[0088] In step S**201**, the image feature cluster C_i (including the image feature vectors f_i with the dimension of 1×d) is rearranged into an image feature vector sequence S_i.

[0089] Next, in step S**202**, at each position in the image feature vector sequence S_i, the cosine distance encoding vector e_t (with the dimension of 1×k) will be concatenated with a respective image feature vector f_t to form a cosine-distance-encoding-information-containing image feature vector sequence S_i*, wherein each component of the cosine-distance-encoding-information-containing image feature vector sequence S_i* is a cosine-distance-encoding-information-containing image feature vector f_t* with the dimension of 1×(k+d), which is concatenation of the cosine distance encoding vector e_t and a respective image feature vector f_t.

[0090] In step S**203**, the cosine-distance-encoding-information-containing image feature vector sequence S_i* is used as the input data sequence of the encoder neural network to generate a respective encoded representation for each input in the input data sequence by using an attention mechanism (self-attention), which shows the attention in the encoded representations of the input data sequence.

[0091] Next, in step S**204**, the encoded representations generated in step S**203** above are used as the input data of the decoder neural network, to generate an output data sequence y_i as a binary sequence.

[0092] Finally, in step S**205**, the output data sequence y_i is combined to form a recognized output image feature cluster.

[0093] Regarding the method of clustering using encoder-decoder model based on attention mechanism described above, the image recognition models, which use the method of clustering using encoder-decoder model based on attention mechanism provided by the present invention, may be understood as a class of models to perform image recognition (e.g., image classification), including image feature extracting models, encoder neural networks, decoder neural networks, and relevant components to perform the functions of image recognition, such as sorter G for performing arrangement, cosine distance encoders to perform cosine distance encoding, memories, and calculators, for example.

[0094] From the above, the present invention has been described in detail according to the preferred embodiments. It is obvious that a person of ordinary may easily generate variations and modifications to described embodiments. Thus, these variations and modifications do not fall outside the scope of the present invention as determined in the appended claims.

What is claimed is:

1. A method of clustering using encoder-decoder model based on attention mechanism, the method comprising:

extracting image features from an image database X consisting of multiple images x_i, by an image feature extracting model, to obtain an image feature dataset comprising image feature vectors f_i, wherein each image feature vector f_i corresponds to one image x_i in the said image database X;

clustering the image feature vectors sampled from the said image feature dataset into image feature clusters C_i based on cosine similarity scores $s_{i,j}$ between the image feature vectors f_i and f_j, wherein each image feature cluster C_i has a center image feature vector;

arranging the image feature vectors f_i in each image feature cluster C_i into an image feature vector sequence S_i in an ascending or descending order based on the cosine similarity scores $s_{i,j}$ of the image feature vectors compared with the center image feature vector in the same said image feature cluster C_i;

generating cosine distance encoding vectors e_t from the components such as the cosine similarity scores $s_{i,j}$, where each cosine distance encoding vector e_t corresponds to an image feature vector f_t in the image feature cluster C_i and the cosine similarity scores $s_{i,j}$ forming the cosine distance encoding vector e_t are the cosine similarity scores between the image feature vectors and the image feature vector f_t in the same said image feature cluster C_i;

concatenating the image feature vector f_t with the respective cosine distance encoding vector e_t to form a respective cosine-distance-encoding-information-containing image feature vector f_t*;

generating a cosine-distance-encoding-information-containing image feature vector sequence S_i* in correspondence with each image feature cluster C_i, wherein the components of the cosine-distance-encoding-information-containing image feature vector sequence S_i* are the cosine-distance-encoding-information-containing image feature vectors f_t* in correspondence with the image feature vectors f_t belonging to a respective image feature cluster C_i, and the cosine-distance-encoding-information-containing image feature vectors f_t* are arranged in an ascending or descending order based on the cosine similarity scores $s_{i,j}$ of the image feature vectors compared with the center image feature vector in the same said image feature cluster C_i;

using the cosine-distance-encoding-information-containing image feature vector sequence S_i* as the input data sequence of an encoder neural network, wherein the encoder neural network is configured to generate a respective encoded representation for each input in the input data sequence by using an attention mechanism, which shows the attention in the encoded representations of the input data sequence; and

decoding, by a decoder neural network, to generate an output data sequence, wherein the decoder neural network is configured to receive the encoded representations as the input data for decoding into the output data sequence.

2. The method according to claim **1**, wherein the step that the encoder neural network generates the respective encoded representation for each input in the input data sequence by using an attention mechanism, which shows the attention in the encoded representations of the input data sequence, comprises:

projecting the cosine-distance-encoding-information-containing image feature vector sequence S_i* into at least one sub-space, wherein for each sub-space, perform the operations of:

determining first, second, and third trainable matrices;

projecting the cosine-distance-encoding-information-containing image feature vector sequence S_i* into first, second, and third super spaces to generate first, second, and third super space features based on the first, second, and third trainable matrices, respectively;

calculating the attention scores $r_{i,j}$ between the cosine-distance-encoding-information-containing image feature vector f_i* and the cosine-distance-encoding-information-containing image feature vectors f_j* in the cosine-distance-encoding-information-containing image feature vector sequence S_i* based on the first super space feature of the cosine-distance-encoding-information-containing image feature vector f_i* and the second super space features of the cosine-distance-encoding-information-containing image feature vectors f_j*; and

generating a sub-space output of the cosine-distance-encoding-information-containing image feature vector f_i* by calculating a weighted sum of the third super space features of the cosine-distance-encoding-information-containing image feature vectors f_j*, wherein the weights assigned to the third super space features are respective attention scores $r_{i,j}$;

linearly transforming a concatenation result of the sub-space outputs of the cosine-distance-encoding-information-containing image feature vector sequence S_i* in correspondence with each sub-space to obtain an attention output; and

generating the encoded representations based on the attention output.

3. The method according to claim **1**, wherein the output data sequence of the decoder neural network is a binary sequence y_i with a length in correspondence with that of the cosine-distance-encoding-information-containing image feature vector sequence S_i*, where the value at the t^{th} position of the binary sequence y_i being 1 denotes the t^{th} image feature vector has the same label as the center image feature vector in the same said image feature cluster C_i, and the value at the t^{th} position of the binary sequence y_i being 0 denotes the t^{th} image feature vector does not have the same label as the center image feature vector in the same said image feature cluster C_i.

4. The method according to claim **1**, wherein the cosine distance encoding vector e_t is determined through the following expression:

$$e_t = \{s_{t,i}\}_{i=1}^{k}$$

where $s_{t,i}$ is the cosine similarity score between the i^{th} image feature vector and the t^{th} image feature vector, and

wherein the cosine-distance-encoding-information-containing image feature vector f_t* is determined through the following expression:

$$f_t^* = \text{concat}(f_t, e_t)$$

where concat is a function that concatenates two vectors into one vector.

5. The method according to claim **1**, wherein the said encoder neural network and decoder neural network are trained using a target function which is determined through the following expression:

$$\mathcal{L}_i(\hat{y}_i, y_i) = -\sum_{t=1}^{k} \left[y_i^t \times \log\left(\sigma(\hat{y}_i^t)\right) + \left(1 - y_i^t\right) \times \log\left(1 - \sigma(\hat{y}_i^t)\right) \right]$$

where σ is the sigmoid function.

6. The method according to claim **1**, wherein the said trained image feature extracting model uses two datasets consisting of a labeled dataset D_L, and an unlabeled dataset D_U.

7. A non-transitory computer readable storage medium comprising computer program instructions that, when executed, perform the method according to claim **1**.

* * * * *