ESANN'2005 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 27-29 April 2005, d-side publi., ISBN 2-930307-05-6.

Learning to Classify a Collection of Images and Texts

P. Saragiotis[1], B. Vrusias[1], K. Ahmad[1]

1 - Department of Computing, University of Surrey, Guildford, Surrey, GU2 7XH, UK

A single net system based on Kohonen's Feature map was trained using a combined vector that contains visual features of an image and its collateral keywords. The performance of the single net was compared with a multinet system, comprising two SOMs, one trained with visual features and the other on keywords, in the presence of a Hebbian network that learns to associate visual features with keywords. The multi-net system performs better than the single net. Similar results were obtained when Grossberg's ART networks were used instead of SOMs.

1 Introduction

Recent work on content-based image retrieval (CBIR) systems has sought inspiration from text-retrieval literature [13]. It has been argued that the all-important visual features can help in the classification (and retrieval) of images but only up to a certain extent: In some cases, the visual features have to be supplemented by an image-external linguistic description of the contents of the images in a CBIR system [8]. The collateral linguistic description, for example salient words, describing key objects in the image is then used to annotate the image, and in querying for the image subsequently either by using visual features or keywords, or a mixture of the two [2]. The salient words can still be ambiguous as experts describing the same image sometimes use different words [10]. Linguistic ambiguity notwithstanding, neurobiological studies indicate that multi-sensory perception is used in a range of cognitive tasks including attention and pattern recognition in noisy environments [4]. This cross-modal interaction has helped in establishing systems that use information in one modality to classify or retrieve information in another – keywords may be used to retrieve images and vice versa. Cross modality may facilitate the development of systems that can learn to automatically attach keywords to images that have no collateral description – *auto-annotation*- and to automatically illustrate a set of keywords with images – *auto-illustration* [15].

We have been investigating whether or not the addition of linguistic features to a visual feature vector will improve the performance of a system that will learn to classify a set of images without *a priori* knowledge of the categories [1]. One way in which this investigation has proceeded has been to train a neural computing system, for example Kohonen's *Self Organising Feature Maps* (*SOM*, [6]) *or* Grossberg's *Adaptive Resonance Theory*-based networks (*ART*, [5]), using a vector that comprises both visual and linguistic features. The performance of this system, trained on a

monolithic vector, is then compared with a *multi-net* system that comprises one network that is trained using vectors with visual features only, and another network trained with vectors with linguistic features only. The two systems are trained independently but in the presence of a third system that learns the association between the most activated nodes in the two independents. For the multinet based on SOMs we have used a Hebbian network for the association [15], and for ART networks we have used the so-called *map field* of the fuzzy ARTMAP network [3]. The connections amongst the two SOMs are fully bi-directional: each node in one SOM is connected to all the nodes in the other SOM. For the fuzzy ARTs, connected through the map field, the connections are not bi-directional.

A commercially available 50,000-image database, that has keywords associated with each image, (the Hemera Photo Objects www.hemera.com) was chosen for training and testing. Each of the images is that of a pre-segmented single-object and belongs to one of the 100-annotator chosen categories. Ten of these categories were arbitrarily selected and over 100 images per category were randomly selected from the collection. In all, 1036 images were selected for training the network and an additional 115 were used for testing. The category information attached with each image was **not** used in training the networks.

Currently, our results indicate that a CBIR system that is based on a multi-net architecture performs better than the one of a single net architecture. These findings tend to confirm the claims that a combination of autonomous, possibly specialist, neural networks may reduce model complexity; fuse the output of the constituent autonomous networks; improve generalisation; and restrict over-fitting [11].

2 Background and Motivation

Modern CBIR systems are expected to deal with the deluge of images that are becoming available due to the Internet and due to digital photography [14]. Such systems have to classify images into fairly complex categories for facilitating systematic storage and efficient retrieval. However, these categories are either not known in advance or do not reflect the ontology of a given domain as understood by less knowledgeable members of the domain. This has motivated us to chose learning systems that do not require an *ab initio* description of categories. It has been observed that visual features will underconstrain the description of an image and that such features have to be supplemented by information in other modalities – especially modalities like language that facilitate the articulation of categories. We have chosen a linguistic feature vector to supplement the information available in the visual description of images.

Self-Organising Maps have been used extensively in a number of neural computations, including organising large collections of documents [6], and in building CBIR systems [7]. The Kohonen maps, using a competitive learning algorithm, obtain ‘a small set of important features’ by a non-linear method based on a layer of adaptive units that gradually develop into an array of feature detectors. The output map of a trained SOM does not cluster the input vectors *per se*, rather similar input vectors are placed close to each other on the map. The literature on SOMs

indicates that some authors use a sequential clustering approach [12]: the SOM arranges the input vectors into the output map and then a statistical clustering algorithm (for example '*k*-means') identifies the cluster boundaries.

The Adaptive Resonance Theory was developed by Grossberg and his colleagues as a way to overcome the so-called stability-plasticity dilemma [5]. ART networks are capable of performing on-line incremental clustering of the input data into an arbitrary number of categories depending on the value of an internal parameter called *vigilance*. ART networks have also been used extensively in a number of classification problems, including text clustering [9].

3 Method: Architecture, Training and Testing

3.1 Architecture

The choice of the number of nodes in the input and output layers of the various networks was determined by reference to the image collection under consideration.

Input: The dimensionality of the visual feature vectors can more or less be fixed with reference to conventional CBIR systems: visual attributes related to shape, edges, texture and colour have been outlined in some detail. A 67 dimensional vector was used – 21 colour features were used together with 19 edge features, 20 from textures and 7 for shape.

The dimensionality of the linguistic vector is not as easy to determine: for the Hemera Collection, a total of 10,018 descriptors were attached to the 1151 images we had selected. On average 8.8 terms were attached to each image by Hemera experts. Each image was treated as a 'document' –comprising keywords only- and information retrieval metrices of term-frequency/inverse document frequency ($t_f id_f$) were used to select amongst the terms for each of the 10 Hemera categories. Rarer terms will not be as representative of a category as compared with more frequent terms. Frequency and $t_f id_f$ values were used to create a 30 dimensional vector. The use of more keywords leading to 50 and 100 dimensional vectors, did not improve performance.

Output: The dimensions of the output map for the SOMs were computed by trial-and-error: a 15X15 network offered the best balance between performance and training time when compared to smaller (e.g. 10X10) or larger (e.g. 50X50) map.

The output map of the trained SOMs for both single and multinet experiments was clustered in *n*-classes using the '*k*-means' algorithm. The number *n* was chosen to be equal to the number of classes into which the humans classify the data.

The number of nodes on the output layer of an ART network (category nodes) depends on the vigilance parameter. For the single ART experiments we varied the vigilance parameter from a minimum value of $\rho_{min} = 0.001$ to a maximum of $\rho_{max} = 0.300$ in order to achieve a number of category nodes close to *n*. For the fuzzy ARTMAP experiments again we varied vigilance to constrain the number of nodes in the second fuzzy ART module close to *n* while allowing the category nodes in the first module to expand.

3.2 Training

Single Net Training: The SOMs and the fuzzy ARTs were each trained on a 97 dimensional vector (67 visual features and 30 keyword-related features). The SOM's were trained for a 1,000 epochs. The training epochs for the fuzzy ARTs were determined by the complexity of the data sets.

Multi-net Training: The two SOMs were trained respectively on a 67 dimensional visual feature vector and a 30 dimensional keyword-based vector. The mediating Hebbian network helped to strengthen (or weaken) the connections between the nodes of the two SOMs that were simultaneously active (or inactive) during one training cycle.

In the fuzzy ARTMAP network, two fuzzy ART modules are linked together by a map field. According to the literature, fuzzy ARTMAP is mainly used for supervised learning where the first fuzzy ART module receives an input vector, while the second the corresponding target vector. In our case we did not use a target vector, but instead the second modality vector. This way each fuzzy ART is presented with an input vector and once both modules determine the winning node, the map field learns the association between the nodes of the two fuzzy ARTs. In case there is a mismatch because of previous associations, a new node has to be selected in the first module. This procedure is repeated for all input vectors until no more mismatches occur.

3.3 Testing

The testing procedure for the single net systems is as follows: (a) present to the network an unseen testing vector; (b) find the best matching unit for the testing vector; (c) if using SOMs then find the class given to that node by '*k*-means'; (d) calculate the evaluation measures using the information about the true (expert's) classes and the classes assigned by the '*k*-means' algorithm (or the category representation layer of the ART network). The testing procedure for the multi-net systems is the same as that for the single-net systems except for the fact that the winning node in one network (linguistic/visual) stimulates a corresponding node in the other network (visual/linguistic).

The performance of the single- and multi-net systems was measured using so-called *F-measure* which, in turn, depends on precision (fraction of retrieved images that are relevant) and recall (fraction of relevant images that are retrieved) statistics; $F_\beta = (\beta^2 + 1) \cdot p \cdot r / (\beta^2 \cdot p + r)$, where p is precision, r is recall, and β a weighting parameter between precision and recall, which in our case was set to 1.

4 Experimental Results

The performance of the three systems, single-net system, multi-net (auto-illustration) system and multi-net (auto-annotation) system was compared using the 115 test vectors. The performance of single net SOM is comparable with that of a multi-net

SOM when the input to the multi-net is a visual feature vector and the constituent visual SOM activates the node belonging to the correct category node in the keyword SOM. However, when the multinet is presented with a keyword feature vector then its performance is about 30% better than that of the single net SOM. (see Table 1).

For the fuzzy ARTMAP, however, we have mixed results. The performance of the single-net fuzzy ART is much poorer than that of a multinet fuzzy ARTMAP that receives a visual feature stimulation and retrieves the 'correct' text node – a 100% increase in performance when two fuzzy ARTs connected with a map field are used instead of a single monolithic fuzzy ART. When a keyword feature vector is presented to the fuzzy ARTMAP, the network cannot converge to a solution because it tries to link similar high level concepts (keyword feature vector) to dissimilar low level visual properties while forcing the number of category nodes in the output layer of the image sub-module to remain close to the number of classes into which the humans classify the data. In the opposite scenario (visual features as input) the network managed to converge because we do not limit the number of category nodes for the highly variable visual features in the input sub-module.

TABLE 1 F-MEASURES FOR THE PERFORMANCE OF SINGLE NET AND MULTINET SYSTEM FOR IMAGE RETRIEVAL USING A COMBINED 97-DIMENSIONAL TEST VECTOR THE SINGLE NET SYSTEM AND 30 & 67 DIMENSIONAL TEST VECTORS FOR THE MULTINET SYSTEM. .

System	Input	Output	SOMs	Fuzzy ART(MAP)s
Single Net	Monolithic vector	Monolithic vector	**0.36**	**0.17**
MultiNet				
AUTO-ANNOTATION	Visual Feature Vector	Keyword Feature Vector	**0.38**	**0.35**
AUTO-ILLUSTRATION	Keyword Feature Vector	Visual Feature Vector	**0.48**	**No convergence**

We have examined the visual feature vectors and the keyword feature vectors in detail. The categories produced by the visual feature vectors are considerably diffused than is the case for the keyword feature vectors. This result is not surprising – keywords are a direct expression of a category in that, for example, the term *mammal*, for example, will represent a whole class of hairy animals that feed their infants. However, an equivalent visual feature that captures animals of all colours, shapes, and textures is well nigh impossible to conceive.

The overall poor performance of all the networks reported above is to a large extent due to the poor discrimination power of the visual features used in the study. *F-measures* obtained for a SOM (or a fuzzy ART) that only classifies keyword vectors and is tested on it, are in the region of 0.8, whereas the corresponding *F-measure* for visual feature vectors is around 0.2.

5 Conclusions

The above results confirm a long-held opinion in the image retrieval community that visual features invariably constrain an image and that conjunctive use of the two

modalities is more beneficial for image retrieval. What the multinet system also demonstrates is that when one modality of information cannot discriminate between two objects successfully, then another modality needs to be *cued* in to improve the overall performance.

References

[1] Ahmad K., Casey M., Vrusias B., Saragiotis P. (2003). "Combining Multiple Modes of Information using Unsupervised Neural Classifiers". Proc. 4th International Workshop on Multiple Classifier Systems (MCS 2003, Guildford, UK) LNCS-2709, Heidelberg: Springer Verlag. pp 236-245.

[2] Barnard K., Duygulu P., de Freitas N., Forsyth D., Blei D., Jordan M.I. (2003). "Matching Words and Pictures". Journal of Machine Learning Research, vol 3, pp 1107-1135.

[3] Carpenter G. A., Grossberg S., Markuzon N., Reynolds J. H., Rosen D. B. (1992). "Fuzzy ARTMAP: A Neural Network Architecture for Incremental Supervised Learning of Analog Multidimensional Maps". IEEE Transactions on Neural Networks vol. 3, No. 5, pp. 698-713.

[4] Driver J., Spence C. (2000). "Multisensory perception: beyond modularity and convergence". Current Biology. Vol. 10, pp 12731-735.

[5] Grossberg S. (1980). "How Does the Brain Build a Cognitive Code?". Psychological Review, vol 87 pp. 1-51.

[6] Kohonen T. (1997). "Self-Organizing Maps". 2nd Ed. Berlin, Heidelberg, New York: Springer-Verlag.

[7] Laaksonen J. T., Koskela J. M., Laakso S. P., Oja E. (2000). "PicSOM - content-based image retrieval with self-organizing maps". Pattern Recognition Letters, Vol 21(13-14): pp 1199-1207.

[8] Li, J and Wang, J.Z. (2003). 'Automatic Linguistic Indexing of Pictures by a Statistical Modeling Approach'. IEEE Trans. Pattern Analysis and Machine Intelligence. Vol 25 (No.10), pp 1-14.

[9] Massey L. (2003). "On the quality of ART1 text clustering". Neural Networks vol. 16, pp. 771-778.

[10] Ogle V.E., Stonebraker M. (1995). "Chabot: retrieval from a relational database of images", IEEE Computer Magazine, Vol. 28(9), pp 40-48.

[11] Sharkey A.J.C. (2002). "Types of Multinet System". In (eds.) F. Roli & J. Kittler. Proc. Of the Third International Workshop on Multiple Classifier Systems (MCS 2002). Berlin, Heidelberg, New York: Springer-Verlag, pp 108-117.

[12] Slonim N., Friedman N., Tishby N. (2002). "Unsupervised document classification using sequential information maximization". Proc SIGIR'02, 25th ACM International Conference on Research and Development of Information Retrieval, Tampere, Finland, ACM Press, New York, USA.

[13] Squire McG.D, Muller W., Muller H., Pun T. (2000). "Content-Based Query of Image databases: Inspirations from Text Retrieval", Pattern Recognition Letters 21. Elsevier, pp 1193-1198.

[14] Veltkamp C. R., Tanase M. (2002). "Content-Based Image Retrieval Systems: A Survey". A revised and extended version of Technical Report UU-CS-2000-34, October 2000.

[15] Vrusias B. (2004). "Combining Unsupervised Classifiers: A Multimodal Case Study". Unpublished PhD thesis, University of Surrey, Guildford, Surrey, UK.