Co-operative neural networks and 'integrated' classification

Khurshid Ahmad, Bogdan Vrusias, and Mariam Tariq

k.ahmad@surrey.ac.uk, b.vrusias@surrey.ac.uk, m.tariq@surrey.ac.uk

University of Surrey, UK

Abstract - 'Integrated' classification refers to the conjunctive or competitive use of two or more (neural) classifiers. A co-operative neural network system comprising two independently trained Kohonen networks and co-operating with the help of a Hebbian network, is described. The effectiveness of such a network is demonstrated by using it to retrieve images and related texts from a multi-media database. Preliminary results of such an approach appear to be encouraging.

1 INTRODUCTION

Single neural networks are used in a range of tasks including classification, prediction and control. Multi-net neural systems are used for tasks that a single net cannot solve, or tasks that can be more effectively solved by the multi-net. Task decomposition is used extensively in artificial intelligence, artificial life, and in learning systems. Amongst the complex tasks are those that involve different modalities of communication: for example, in natural language processing one has to deal with speech and text; in multi-media systems there could be the different modalities of text and images. Our previous work comprises research in the area of language development where a co-operative neural network architecture involved both supervised and unsupervised learning algorithms to simulate the learning of concepts and words, conceptual and semantic relations as well as simple word-order rules [1]. A similar architecture was used in the simulation of language disorders using co-operative modular connectionist networks, in which semantic-lexical and phonological knowledge are instantiated using self-organising Kohonen maps, while connections between them are implemented using Hebbian networks [5].

Sharkey [11] has categorised multi-nets as systems that either comprise *ensembles* or *modular* combinations of neural networks: 'In an *ensemble* combination, the component nets are redundant in that they each provide a solution to the same task, or task component, even though this solution might be obtained by different means. By contrast, under a *modular* approach, the task or problem is decomposed into a number of subtasks, and the complete task solution requires the contribution of all the several modules'.

Classification can be defined as the task of identifying and arranging entities into classes, based on a set of characteristics. The complexity of the task can be related to the complexity of the data set that has to be classified. If the data is consistent, then classification using a single-net system could be the best solution. More frequently than not, the data set content is inconsistent and can contain ambiguities thereby making the execution of the classification almost impossible. Decomposing the problem into subtasks can be of some help here: Instead of training one

large classifier on all data, we divide it into smaller classifiers trained on the individual data sources, and combine the outputs. Even when the data is consistent and can be trained with a single classifier, one can argue that if instead two or more classifiers are used, each trained on different parts of the data, then training time could decrease.

Neural networks literature has many examples where a single monolithic network has been used in classification. The modular approach suggests otherwise: Given that a typical classification task can be decomposed into subtasks, then an individual neural network may be trained to learn the solution of each subtask. The individual networks then learn to co-operate with each other in order to perform the classification task. We describe how classification of annotated images, for the purpose of indexing and subsequently retrieving the images from a database, can benefit from a modular neural network system. An annotated image, for us, is one that is accompanied by a textual description typically in unconstrained free text. The image classification task can be decomposed into three subtasks, each of which is performed by an individual neural network. For instance, one Kohonen map may be trained to learn to classify the individual images on purely visual features, for instance, colour, shape, texture and so on; another Kohonen map may be trained to learn to classify the accompanying text on the basis of keywords in context; and, the third network in our system is a Hebbian network that learns to establish connections between the most active visual features of a given image to the most active textual/keyword related features of the accompanying text.

2 TRADITIONAL IMAGE-ANNOTATION SYSTEMS

The purely visual information within an image cannot constrain the image effectively for storage or retrieval. Keyword-based searching has been amongst the most popular ways for retrieving images. Images were sometimes placed under specific categories, which were then organized into a hierarchy that users could navigate and browse to search for relevant images. Images were generally annotated and/or categorized by humans, which may be time-consuming if the image repository is large. Furthermore, such annotation and categorisation are subjective. The content-based image retrieval paradigm was conceived in order to address these limitations. Here retrieval is based on the visual properties of an image such as colour and texture. Though successful to a certain extent this method has its own limitations.

Srihari [13] is one of the first researchers to contemplate the notion of a more effective understanding and retrieval of pictures if the two modalities of vision and language are

combined. For humans, the correlation of visual information with speech/text is a given in that such a 'correspondence' expedites image understanding and text understanding. Srihari's PICTURE system [14] demonstrates how to use text associated with images, known as *collateral text*, in scene understanding by defining visual semantics. The collateral text is a kind of an annotation for the image. This text is analysed and processed at the different linguistic levels: lexical, syntactic and semantic levels as well as by interpreting spatial prepositions. The contents of a collateral text are a source of *a priori* knowledge for the image classification process. Srihari's outline of her *Show&Tell* system [15] includes open-ended questions such as semantic relationships between terms that may give generic or specific description of objects within the images (*recreational facility* may refer to an *athletic track*, *swimming pool*, or *playing fields*).

Paek et al. [8] use a different methodology for the integration of linguistic and visual-based approaches for the labelling and classification of indoor/outdoor photographs in the domain of terrorism. They use a traditional metric in information retrieval, a product of term frequency and inverse document frequency ($TF \cdot IDF$), to classify texts. Paek et al use a similar metric for images, a product of the frequency of an object in an image and an inverse image frequency ($OF \cdot IIF$). By combining these two vectors they reach a classification accuracy of around 12% over that of other existing methods.

Similarly Sclaroff et al [10] have proposed a system that combines visual and textual statistics in the form of a single vector that can be used to search for images on the Web. The novel idea here is to use Latent Semantic Indexing (LSI) on the text side, and colour and orientation histograms for the images. LSI provides some advantages over the classical keyword method used by most search engines in that it implicitly covers the issues of synonyms, word senses and term omissions. The text used is either taken from the URL and/or heuristically determined from its proximity to the image on a web document. Their experiments showed that a better performance was achieved by a combination of the text and visual vectors.

The research discussed above mainly focuses on integrated indexes and not knowledge structures that model both visual and linguistic features. Benitez et al [5] have developed a knowledge representation framework called *MediaNet* that represents both semantic and perceptual information related to multimedia data. A semantic network is used where the concepts (nodes) refer to the semantic notion of what an object is. Each concept may have a number of text representations such as "man," "human," "homo" as well as audio and visual representations. Concepts can be linked together by various relationships such as hypernymy, meronymy and entailment. Concepts with their textual representations were created using *WordNet*, which also generated all the senses and synonyms for each word.

WordNet was also used to generate automatically all the required relationships. Visual representations were automatically generated using colour and texture feature extraction tools.

3 CO-OPERATIVE NEURAL NETWORKS AND INTEGRATED IMAGE/TEXT CLASSIFIERS

Amari [3] and Willshaw & Malsburg [17] have suggested that neural network architectures can be used to generate a topographic map from a pre-synaptic two-dimensional array into a post-synaptic two-dimensional array. The neurons in the pre-synaptic (input) layer are connected to the post-synaptic (output) layer through the use of the Hebbian rule. Our architecture is based on that of Abidi & Ahmad [1] who extended this work to a multi-net format by successfully connecting a Hebbian network between two SOFMs in a simulation of the development of mappings between concepts and words in children. Here the two SOFMs can be compared to Willshaw & Malsburg's pre- and post-synaptic sheets.

The collateral use of textual description of an image may, in a restricted sense, be construed as a *Hebbian* link: a link, or more precisely an *association*, which evolves over time [6]. This association may be bi-directional. Note that humans develop text and image analysis facilities more or less independently of each other. There is some argument about the validity of this independence in that there are specialist areas of human brain dealing with images and text. There is some evidence also that concepts and words, learnt independently, are associated experientially (see [1]).

3.1 Multi-net architecture for multi-net classification

We propose a multi-net system comprising SOFMs modules that combine their solution to solve a problem.

Figure 3.1. Basic system architecture.

The input, represented by vectors, is created automatically and that input trains the artificial neural network modules of the system. Each neural network is evaluated, after the training, for its consistency and performance. Each module learns the data provided and for each input it creates a

Hebbian link between the winner nodes of the different output maps. In such a way the modules learn to co-operate on query requests in order to produce results (Figure 3.1).

The artificial neural networks are Kohonen self-organising feature maps (SOFM) and their architecture is based on the SANN (Surrey Artificial Neural Network Classifier) system developed at the University of Surrey [2].

3.2 Associating neural networks

The input to the system is a vector that has been created based on the characteristics of the data. It is separated on two different vectors that each feed the respective neural network module (Figure 3.2).

Figure 3.2. Multi-net connectionism based on Hebbian learning.

The two neural networks are connected with a Hebbian Learning Rule to form a modular and co-operative multi-net. Each time an image and accompanying descriptive text are input to each neural network respectively, the two winning nodes are linked and reinforced using the Hebbian Learning Rule (Figure 3.3).

Figure 3.3.a.

Figure 3.3.b.

Figure 3.3. The two neural networks dynamically create a link between the winning neurons of the image and text input (a). An image is related to all text that falls on the area of the linked node (b).

The result is a co-operative classification of images and texts that both contribute to the output.

4 CLASSIFYING OVERVIEW, MIDRANGE OR CLOSE-UP IMAGES

The purpose of crime scene photographs is to document the scene of crime exactly as it was found. The photographs are used for detailed analysis by the investigation officers and

later as evidence in court. The SoCIS Project (Scene of Crime Information System), sponsored by the UK's Engineering and Physical Sciences Research Council, is investigating how images can be constrained through the use of collateral texts. Working in close co-operation of four UK Police Forces, we are developing a web-based system that will automatically index an image using available collateral text. This index will be used conjunctively with the visual features of the image for storage and retrieval from databases. The objective of the project is to develop a method that will be able to index and retrieve images in an arbitrary domain. The user can then automatically create reports and browse the image database.

Image classification is a complex task and is sometimes very application specific. Scene-of-crime image classification is a good example here. Typically three types of photographs are taken at a scene of crime: *Overview photographs* that show the entire scene; *Midrange photographs* that are taken to show interesting objects and their spatial relationships; and *Close-up photographs* that are taken of relevant objects, to show details of a single object or a single point of interest.

The SoCIS project is investigating how scene of crime officers (SoCOs) annotate the scene of crime photographs. A reference database of 66 images has been created and annotated by the SoCOs. In this paper we report on a simple one-line-per-image annotation (approximately 10 words per line).

5 CREATING FEATURE VECTORS

Each annotated image has two associated feature vectors: the physical features of the image itself, such as colour, texture, shape (of objects), brightness and so on; and the textual features of the annotation. There are two kinds of feature vectors that are generated simultaneously: the image representation vector and the text representation vector.

5.1 Creating Image Feature Vectors

There can be various kinds of information associated with an image: visual information that is related directly to the perceptual properties of the image such as the colours and shapes of objects present as well as non-visual information such as the identification of objects and events or the format of the image.

Some database management systems that support content-based image retrieval facilitate the storage of images of different types and formats as external files or binary large objects (BLOBs) directly in the database. A feature extractor function is provided by the systems to automatically generate the feature vectors of an image based on its visual characteristics such as colour, shape and texture. The feature vector is stored in the database as an array of numbers and is a combination of the following properties of an image:

PROPERTY	METRICATION
Colour content	of colours in an image.
Shape content,	of the relative orientation, curvature, and contrast of lines in an image.
Texture content,	on a small scale, of the flow and roughness of an image
Brightness structure,	of the brightness at each point in the image.
Colour structure	of the hue, saturation, and brightness at each point in the image.
Aspect ratio	the ratio of the width to the height in the image.

TABLE 1. Informix Image DataBlade feature vector properties.

5.2 Creating Text Feature Vectors

Each of the photographs in our reference database (of 66 images) has a key collateral text associated with it: an approximately 10-word description of the image. These descriptions are a good example of the specialist language used by forensic scientists; specialists of very different specialisms use their idiosyncratic language dominated by terms of the specialist domain and sentences that are largely declarative and imperative. This language can be differentiated from the day-to-day language of everyday usage – the so called general language.

It has been noted that whether or not a text is specialist or general language, the so-called closed-class words, like the determiners ('a', 'an' and 'the' in English), prepositions, conjunctions and certain verbs, may comprise as much as 25% of a given text. The relative frequency of these closed-class words or grammatical words, used to draw up the stop-lists used in information retrieval systems are usually the same in specialist and general language texts; it is the nominals (most terms are either single or compound nominals) that occur with different relative frequencies (the relative frequency of a word in a text is the ratio of the absolute frequency of the word to the total number of words in a text or corpus). A second ratio, that of relative frequency of the same term in a specialist text to that in general language texts can be used to quantify the lexical difference between the two texts. This ratio has been termed *weirdness* [1] to indicate how it measures the preponderance of words in specialist texts that would be unusual in general language. We have used the frequency information about general language words from the 100 million word British National Corpus [2].

Typically, before text documents are represented as vectors in order to act as the input to a text categorisation system, pre-processing takes the form of filters to remove words 'low in content' from the text (see the WEBSOM method [7]). We automatically remove punctuation, numerical expressions and closed-class words as a precursor of generating the feature set. Vectors representing the image description texts were created on the basis of a lexical profile of the training set of texts. This lexical profile was determined by two measures:

the frequency of a term; and, a weirdness coefficient describing the subject-specificity of a term, defined as:

$$\text{weirdness coefficient} = \frac{f_s/N_s}{f_g/N_g} \quad (1)$$

where f_s = frequency of term in a specialist corpus
 f_g = frequency of term in general language
 N_s = total number of terms in the specialist corpus
and N_g = total number of terms in the general language

The system automatically creates the feature set, by first removing the 25% of closed class words from the description and then choosing the relevant keywords from the remaining text. The choice is based on the frequency of the occurring words and the words with the highest weirdness coefficient. Subsequently, 70% of the highest weird words are selected, excluding spelling mistakes, and terms too infrequent to provide consistency within a domain are avoided. The text being analysed is descriptive, containing many relevant words, and thus generates a high proportion of keywords. A high value for the weirdness coefficient is indicative of a word which is uncommon in general language but common in the specialist corpus under examination and is thus a good candidate for a domain term or other word specific to that genre.

Having identified the feature set, the training vectors for each of the texts could then be generated. Each vector consisted of binary values indicating the presence or not of each of the feature words determined above, that is,

$$y(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \end{cases} \quad (2)$$

The SoCIS system was used to compute the frequency distribution of words in the descriptions entered by the expert. The user can choose the relevant terminology list to use as the reference for calculating the weirdness coefficient, which may be general language terms or domain specific terms.

6 TRAINING AND TESTING THE SYSTEM

The text and image feature vectors were computed respectively by image feature extraction and term extraction (see 5.1 and 5.2 above). Each vector was also annotated by one item of additional information: the image type – whether or not the image was *overview*, *mid-range* or *close-up*.

The SoCIS system uses n-dimensional vectors as inputs and produces a visual two-dimensional output map (SOFM), with details of the inputs' class and position on the map. The system can be configured to perform many kinds of training with different settings each time. Initially the settings are arranged to work as default with the most commonly used

training parameters, but the user can change them as he or she wishes.

For our case study, we chose two Kohonen maps with 15x15 nodes each to generate 50625 ($15^2 \times 15^2$) Hebbian links between the nodes. A self-organised network represents each map, and Hebbian connections between the maps allow for the coordination between them. We trained the system using 66 pairs of input vectors; 139-dimensional vectors for the text input and 87-dimensional vectors for the image input. The system appeared either to complete its training or be saturated in 1000 cycles; in each cycle we updated the Hebbian connections between the nodes from the two self-organised networks. An input pair (one input per map) was presented to the maps simultaneously and the resulting output activation reinforced the weights from the Hebbian links. It took less than 10 minutes to train the system (on a Pentium III, 866MHz, 265MB RAM system), but this speed is dependent not only on the machine's power but also on the dimensions of the input and output maps, and the amount of data.

For training purposes we randomly selected 90% of the data. Figure 6.1 shows the results of the two Kohonen maps after training – one for text (6.1.a) and one for image features (6.1.b). We have tested the system's performance with the remaining 10% of the data, by comparing it to the individual performances of each module separately. The text classification, when tested, had a performance of about 74% accuracy and the image classification had a performance of about 48%.

Figure 6.1.a.

Figure 6.1.b.

Figure 6.1. Results from SoCIS. On the left (a) is the output map of the text descriptions and on the right (b) the output map of the image features. The individual symbols in the output layer refer to the three types of photographs; plus sign is overview, circle is midrange and triangle is close-up.

Note that we have trained the Hebbian link whilst the two independent networks were being trained as well. This differs from our previous work where the Hebbian links were established after the two networks were fully trained [1, 18]. We believe the simultaneous training of the Hebbian links simulates what takes place in animal brains: for example, Singer [12] describes the use of a self-organisation process to

learn the binocular association between the two visual fields of a cat's vision.

Our initial estimation of linking the two maps shows that the accuracy of retrieval can be improved considerably by 10% and 40%. The 10% improvement relates to the performance of the co-operative system when compared with retrieval purely from textual feature; the 40% improvement comes from comparing with image features only. Note that the accuracy retrieved is for the three related categories: close-up, midrange and overall

7 CONCLUSIONS AND FUTURE WORK

We have attempted to present a method by which a multi-network system can co-operatively deal with two independent inputs and learn not only the underlying characteristics of the inputs but be able to collocate symbiotic inputs in the two images. Symbiosis here refers to the observation that one may be able to retrieve or match one of the independent inputs through the use of the other. The vectors representing the two inputs, one text one image, were chosen with a minimal bias. For text, we chose the keywords through a heuristic which can be implemented on a computer system: the heuristics based on different distributions of the same word in general and special languages, or the weirdness of the term. For the image, our primitive features are based on the experience of other image analysis researchers and, by and large, focus on the physical characteristics of the images. In this sense, the co-operative multinet neural architecture allows us to integrate the output of the two classifiers.

Each module of the multi-net system learns the output image of the other network, through the Hebbian link created under the training process. That establishes a dynamic link between the two modalities of image and text.

Each vector is an input to a different network, and the two are linked dynamically using the Hebbian rule. In contrast Paek et al. [8] used a single combined vector to integrate text and images. Keeping the vectors separated may be a more flexible way of querying text or image independently, as well as having the benefit of training individual networks that become experts on the specific domain.

The proposed multi-net architecture will be incorporated into the SoCIS prototype, so that the user (SoCO) can query the system to retrieve either images or text. For example the SoCO can search for images by inputting "browning pistol", or provide a sample image to get collateral text.

In a previous paper [2], we reported how we have used the Fisher Linear Discriminant Analysis to evaluate the efficiency of the clusters produced by the Kohonen map. We are currently extending this method to cope with clustering in two different networks.

REFERENCES

- [1] Abidi S.S.R., Ahmad K. "Conglomerate' Neural Network Architectures: The Way Ahead for Simulating Early Language Development". *Journal of Information Science and Engineering*, (Academia Sinica). Volume 13 (Number 2), pp 235-266, 1997.
- [2] Ahmad K., Vrusias B., Ledford A. "Choosing Feature Sets for Training Self-Organising Maps: A Case Study". *Neural Computing and Applications*. Vol. 10, pp 56-66, 2001.
- [3] Amari S. "Topographic Organization of Nerve Fields". *Bulletin of Mathematical Biology*, vol. 42, pp. 339-364, 1980.
- [4] Bechtel W., & Abrahamsen A. "Connectionism and the mind: An introduction to parallel processing in networks". Oxford, UK: Blackwell, 1993.
- [5] Benitez A.B., Smith J.R., and Chang S.-F., "MediaNet: A multimedia information network for knowledge representation," in Conference on Internet Multimedia Management Systems, vol. 4210, IST/SPIE'00, (Boston, MA), Nov. 6--7 2000.
- [6] Hebb D.O., "The Organization of Behavior", New York: Wiley, 1949.
- [7] Lagus, K. "Generalizability of the WEBSOM method to document collections of various types." In Proc. Of 6th European Congress on Intelligent Techniques & Soft Computing (EUFIT'98), 1, pp. 210-214, Aachen, Germany, 1998.
- [8] Paek S., Sable C. L., Hatzivassiloglou V., Jaimes A., Schiffman B. H., Chang S.-F., & McKeown K. R., "Integration of visual and text based approaches for the content labeling and classification of Photographs", ACM SIGIR'99 Workshop on Multimedia Indexing and Retrieval, Berkeley, CA, Aug. 19, 1999.
- [9] Rumelhart D.E., & McClelland J.L. "Parallel distributed processing: Explorations in the microstructure of cognition", vol. 1: Foundations. Cambridge, MA: MIT Press, 1986.
- [10] Sclaroff S., La Cascia M., Sethi S., "Unifying Textual and Visual Cues for content-Based Image Retrieval on the World Wide Web," Computer Vision and Image Understanding, Vol.75, Nos. 1-2, pp. 86-98, Academic Press, 1999.
- [11] Sharkey, A.J.C (Ed.) "Combining Artificial Neural Nets: Ensemble and Modular Multi-Net Systems, Springer-Verlag London Limited", 1999.
- [12] Singer, W. Ontogenetic Self-Organisation and Learning. In McGaugh, J.L., Weiberger, N.M. & Lynch, G. (Eds). *Brain Organisation and Memory: Cells, Systems, and Circuits*, pp.211-233, Oxford University Press, 1990.
- [13] Srihari R.K., "Computational Models for Integrating Linguistic and Visual Information: A Survey, Artificial Intelligence Review, special issue on Integrating Language and Vision", Volume 8, pp. 349--369, 1995.
- [14] Srihari R.K., "Use of Collateral Text in Understanding Photos, Artificial Intelligence Review, special issue on Integrating Language and Vision", Volume 8, pp. 409--430, 1995.
- [15] Srihari R.K. & Zhang Z., "Show&Tell: a Semi-Automated Image Annotation System," IEEE Multimedia, vol.7, no.3, pp.61-71. IEEE, USA, July-Sept. 2000.
- [16] Staggs S., "Crime Scene and Evidence Photographer's Guide", Staggs Pub, 1997.
- [17] Willshaw, D.J. & von der Malsburg, C. "How Patterned Neural Connections can be set up by Self-Organization", *Proceedings of the Royal Society, Series B*, vol. 194, pp. 431-445, 1976.
- [18] Wright, J.F. & Ahmad, K. 'The Connectionist of Aphasic Naming'. *Brain and Language*. Volume 59, pp 367-389, 1997.