A Biologically Plausible SOM Representation of the Orthographic Form of 50,000 French Words

Claude Touzet, Christopher Kermorvant, Hervé Glotin

A Biologically Plausible SOM Representation of the Orthographic Form of 50,000 French Words

Claude Touzet[1], Christopher Kermorvant[2], and Hervé Glotin[3]

[1]Aix-Marseille University (AMU), Lab. de Neurosciences Intégratives et Adaptatives, LNIA UMR-CNRS 7260, Pôle Cerveau-Comportement-Cognition, Marseille, France
[2]A2iA SA (Analyse d'Image & Intelligence Artificielle), Paris, France
[3]Institut Univ. de France (IUF) & Univ. Toulon (UTLN), Lab. des Sciences de l'Information et des Systèmes (LSIS), UMR CNRS 7296, Toulon, France
claude.touzet@univ-amu.fr,
christopher.kermorvant@a2ia.com,
glotin@univ-tln.fr

Abstract. Recently, an important aspect of human visual word recognition has been characterized. The letter position is encoded in our brain using an explicit representation of order based on letter pairs: the open-bigram coding [15]. We hypothesize that spelling has evolved in order to minimize reading errors. Therefore, word recognition using bigrams — instead of letters — should be more efficient. First, we study the influence of the size of the neighborhood, which defines the number of bigrams per word, on the performance of the matching between bigrams and word. Our tests are conducted against one of the best recognition solutions used today by the industry, which matches letters to words. Secondly, we build a cortical map representation of the words in the bigram space — which implies numerous experiments in order to achieve a satisfactory projection. Third, we develop an ultra-fast version of the self-organizing map in order to achieve learning in minutes instead of months.

Keywords: Handwriting recognition, word recognition, open-bigram coding, orthographic representation, cortical representation

1 Introduction

Visual handwritten word recognition is an active field, attracting hundreds of researchers [1], starting as early as 1929. A huge amount of ideas have been implemented and tested including algorithms (such as dynamic programming [2]) or holistic approaches (such as considering only the global characteristics of the word [3]), statistical methods (such as hidden Markov models (HMM) [4]), contextual approaches (such as contextual character geometry [5]), and artificial neural networks (such as multiple layer perceptron (MLP) [6] and error-backpropagation training [7] or self-organizing maps [8]).

Since 2009, connectionist models such as multi-dimensional LSTM (Long Short-Term Memory) recurrent neural networks [9-10], deep feed-forward neural

networks [11] and various mixtures of these have won several international connected handwriting competitions (such as the International Conference on Document Analysis and Recognition) without any prior knowledge about the various languages (French [17], Arabic [24]) to be learned. GPU-based deep learning methods for feed-forward networks were the first artificial pattern recognizers to achieve human-competitive performance [12] on the famous MNIST handwritten digits problem [13].

Such results support the claim that we are currently experiencing a second Neural Network ReNNaissance (the first one happened between 1985 and 1993). In many applications, deep NNs are now outperforming all other methods, including support vector machines (SVM).

Deep and recurrent neural networks refer explicitly to the brain architectures, and mimic some of the principles that are known about the way that human brain implement word reading. Dehaene *et al.* have proposed a biologically plausible model of the cortical organization of reading [14] that assumes seven successive steps of increasing complexity — from the retinal ganglion cells to a cortical map of the orthographic word forms.

Cognitive psychology has done a tremendous amount of work relatively to reading, one among the most important cognitive abilities. However, these discoveries have not been considered by pattern recognition researchers, most certainly because of field boundaries between soft and hard science. One of the most recent successes of experimental psychology was the demonstration that human visual word recognition uses an explicit representation of letter position order based on letter pairs: the open-bigram coding [15].

In its simplest form, an open-bigram (OB) coding assumes a limit of 2 intervening letters (see Bigrams 2 in Fig. 1). For example, TABLE bigrams amount to 9: TA, TB, TL, (not TE), AB, AL, AE, BL, BE, LE. The weighting of each bigram is 1 if present (0 otherwise) in binary OB models. In graded OB models, weights decrease with the distance between letter positions.

1.1 Why bigrams are better

Various measures of distance can be used to ascertain the orthographic proximity of two words.

1. For example, the orthographic distance (D1) between two words (X, the number of shared letters):

$$D1\ (word1, word2) = (2\ X)\ /\ (word\_length1 + word\_length2)$$

Distance D1 is an increasing arithmetic function of X. This distance is a logical choice when using a letters coding model.

2. Another possibility is the distance (D2):

$$D2\ (word1, word2) = (X * (X+1))\ /\ (word\_length1 * (word\_length1 + 1) + word\_length2 * (word\_length2 + 1))$$

Distance D2 is an increasing geometric function of X. This distance is a logical choice for OB coding since the number of bigrams in common between two words is given by:

$$(X * (X+1)) / 2$$

A geometric increase in distance between words is interesting because it allows to take into account the respective length of the words. For example, in the case of two words of respective length 5 and 8 letters, sharing 3 letters, D1 = 6/13 and D2 = 12/102. In the case of two words of respective length 3 and 10 letters sharing 3 letters, D1 remains unchanged (6/13), where D2 = 12/122.

Using D2, when the number of shared letters is equivalent, lower ratios (word_length1 / word_length2) are privileged. In a representation that takes into account the distance between neighbors, D2 privileges neighbor words with the same length. To resume, the bigram representation (resp. to a 'letter' representation) allows for a greater continuity of the representation when the length of the words is also taken into account.

The Levenshtein distance (Edit-distance) takes into account the position of the letters in the word. Therefore, it is less biologically plausible.

1.2 How many bigrams per word?

Using the RIMES data-set [16] (7400 words) and the letters extracted by A2iA [17] (first proposal) we test the influence of the size of the bigram set over the word recognition. It is important to note that the 'poor' quality of the letter extraction only allows a Word Recognition Rate of 28%.

When we use a nearest neighbor convergence with distance D1 (because we know the whole vocabulary), a performance of 44% is achieved.

Fig2. shows that the performance using bigrams are better, depending on the size of the bigram set. We vary this size from a bigram set with no intervening letter (bigrams 0: TABLE = TA, AB, BL, LE) to the whole letters of the word (TA, TB, TL, TE, AB, AL, AE, BL, BE, LE).

Letters	Bigrams 0	Bigrams 1	Bigrams 2	Bigrams 3	Bigrams (whole)
44%	45%	48%	49%	50%	51%

Fig. 1. A bigram representation of the word — in the case of the RIMES data-set — allows a much better performance in recognition (improvement from 44% to 51%).

Because the bigram representation is an over-coding, missing or wrongly labeled letters have less impact on the recognition procedure. Bigrams increase the size of the representation (compared to a letter representation), which allows to resist to failures. This seems to imply that existing words in the language (French) have evolved in order for this bigram over-coding to be pertinent (at least more than a pure letter representation is).

2 Cortical map model

A Kohonen map (also known as a Self-Organized Map - SOM [20]) is a model of the cortical map. We will use it to implement a biologically plausible representation of the orthographic form of words.

2.1 Not uniform representation despite uniform frequency

The following figure illustrates the performances of the SOM learning of 25 (French) words (uniform frequency distribution for all words). α and β (learning coefficients for the winner and its four neighbors) are initially set to 0.6 and 0.15, and decrease with the number of iterations (by 1/total_nb_of_iterations to 0.1 and 0.05 resp.). Each node has four neighbors (North, South, East and West), nodes on the border of the map have only three neighbors, nodes at the corners have only two neighbors. The size of the map is 25 nodes (5 x 5). The number of iterations is set to 50. Learning samples selection is random. Number of input dimensions: 193 (binary OB). Non-null inputs average only a few dozens per sample. Figure 2 displays the nodes associated to each word.

The words (Fig. 2) represented by the same node are similar, but nevertheless quite different. In particular, the length of the words may be very different, and it seems that the short words (e.g., "action") are somewhat pulled by the long ones. This comes from the fact that only a fraction of the inputs are non null (e.g., 15 out of 193 in the case of "action"), and the impact of the (null) input weights are important.

	amélio-		ALBESTROFF	
allée		adapté adaptée		acheter
	aboie aboiements aboit		agents	affectant annexe
Abonné	Alors	allemand animal		accidenté accidentée action
abonnés	Ainsi aisé		ACTUELLEMENT	accompagnée amenée

Fig. 2. SOM learning of 25 words using their bigram representations. Several nodes have no matching correspondence with any words of the learning base, when at the same time several nodes are the prototypes for several words (such as: "accidenté accidentée action").

2.2 Long words pull shorter ones

We introduce a difference among the non-null and null inputs by using different values of α and β when the weight update relates to null inputs. They are fixed during all the learning and set to 0.05 and 0.01 respectively. If these coefficients were set to 0 then the weights associated to these null inputs could not be updated, which ends-up with a bias (favoring long words) since these connections are nevertheless updated from time to time by non-null inputs. As shown in Fig. 3, the lengths of the various words belonging to the same node are closer. However, as in the previous case (Fig.2), there are numerous non-used nodes, and an exaggeration of the distance between nodes.

amenée			allée allemand animal	
				ACTUELLEMENT
amélio-	Alors		accompagnée adapté adaptée	
	aboie Ainsi aisé	aboiements		ALBESTROFF
annexe	Abonné abonnés	aboit action agents	accidenté accidentée	acheter affectant

Fig. 3. α and β associated to null inputs are fixed set to 0.05 and 0.01 respectively. The lengths of the various words belonging to the same node are more similar (e.g., "aboit action agents").

2.3 Equi-selection of the winners

To alleviate the defect shown on the previous Fig. 3, we modify the learning algorithm in order to impose that the each node wins as often as any other, only once per iteration (Fig. 4).

affectant	Alors	action		accidenté accidentée
aisé	Ainsi	acheter		adapté adaptée
allée	ACTUELLEMENT	ALBESTROFF		aboie
animal	amélio-		aboit agents	aboiements
accompagnée		allemand amenée annexe		Abonné abonnés

Fig. 4. Forcing the learning on each node has improved the occupancy of the map. The number of unused nodes is reduced by a factor of 2 (compared to Fig. 3). However, there are still errors in the sense that a node may represent several words (e.g., "allemand amenée annexe").

2.4 Increasing map size to add flexibility

One possibility that would explain this overuse of several nodes — and non-use of several others — may be related to the fact that the distribution of the words (and their bigrams) is highly constrained by the size of the map (25 nodes for 25 words). A larger map helps to spread the words without losing the neighborhood property (Fig. 5).

	amenée		aboit	aboie	
ACTUELLEMENT		acheter annexe		aboiements	
	ALBESTROFF		adapté adaptée		amélio-
Ainsi aisé		allemand		action	Alors
	abonnés	allée		affectant	
Abonné	animal	agents	accompagnée		accidenté accidentée

Fig. 5. A 36 nodes map (6 x 6) representing the 25 learning samples. The number of learning iterations has increased to 100 (instead of 50), in order to allow the same amount of modifications per weight. The larger map allows a better separation between words that are not true neighbors. Only 2 nodes representing more than one word ask for explanations: "Ainsi aisé" and "acheter annexe".

2.5 A correct cortical map representation

Continuing with the idea of extending the map in order to separate what is different, it is of tremendous importance to clearly see the frontiers between the words (in order to implement an efficient word recognition system: one node/one word). Fig. 6 displays the word associated to each node (not just the winning node associated to a given input). A given word may now be represented by several nodes.

Nodes 1 to 10	11 – 20	21-30	31-40	41-50
aboit	aboie	allée	allée	amélio-
aboit	aboit	Abonné	amélio-	amélio-
aboit	Abonné	Abonné	Abonné	amélio-
abonnés	abonnés	Abonné	abonnés	agents
abonnés	aisé	aisé	adapté	agents
abonnés	aisé	adapté	adapté	adaptée
abonnés	aisé	aisé	adaptée	adaptée
abonnés	aisé	aisé	aisé	adaptée
abonnés	abonnés	aisé	accidenté	**accidentée**
abonnés	allée	Abonné	accidenté	accidenté
51-60	61-70	71-80	81-90	91-100
allée	allemand	allemand	accompagnée	accompagnée
amélio-	animal	aboiements	aboiements	annexe
Ainsi	Ainsi	aboiements	aboiements	annexe
Ainsi	Alors	acheter	amenée	amenée
agents	acheter	acheter	ALBESTROFF	amenée
agents	adaptée	acheter	ALBESTROFF	ACTUELLEMENT
adaptée	**accidentée**	affectant	ACTUELLEMENT	ACTUELLEMENT
accidentée	**accidentée**	affectant	affectant	affectant
accidentée	**accidentée**	**accidentée**	affectant	amenée
accidentée	action	action	action	action

Fig. 6. A 10x10 map (100 nodes) representing the 25 samples of the learning base. Due to space constraints, the map has been cut in two equal parts. In fact, there is only one map of 10 columns. Again, due to the increase of the map size, the number of iterations has been set to 200. As we can see, the frontiers between the various words have a clear semantics. The respective occupancy size (measured using the number of nodes associated to a given word) contains also some information. Similar words (e.g., "**accidentée**" (in bold) and accidenté") occupy larger regions than "isolated" words (such as "animal" or "Alors").

3 Ultra-fast building of the SOM

We try to build a SOM (with 4 neighbors per neuron) using a D2 (bigram) distance for the 50 000 words of the French (using the eManulex database [18]). Computing requirements are huge, since a matrix of the D2 measures (50 000 x 50 000 — about 20 Go of RAM) must be computed and kept into memory [19]. An on-the-fly computing does not solve the problem because each iteration requires about 2.5 GFLOPS, and several thousands of iterations are required. It would take about 6 months on a standard PC using a Python written software to generate the SOM representing the 50 000 French words. Obviously, acceleration procedures must be found.

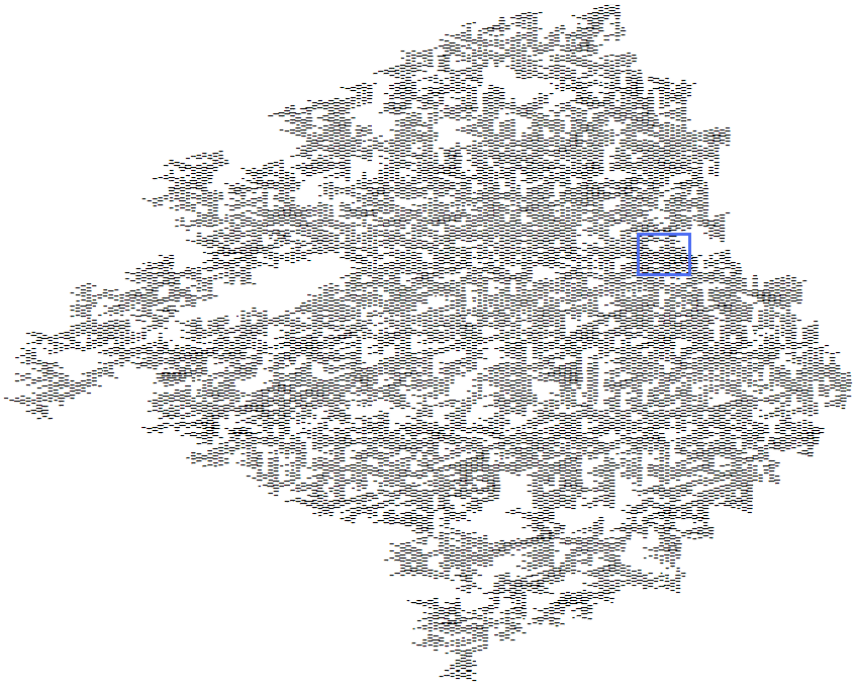

Fig. 7. A cortical map coding 50 000 French words. The size of the map (if every word was readable — font size "6") is about three meters long. The complete map is accessible at: http://www.touzet.org/Claude/Cognilego/FSOM44025-3.pdf ; an animated version is accessible at: http://www.touzet.org/Claude/Cognilego/FSOM44025-3.avi . Zones in white are regions without words. They constitute frontiers among regions of similar words. Following our initial hypothesis that the orthographic form of words optimizes bigrammic recognition, it is tempting to make the hypothesis that these zones are available for future creations of words which will be easily recognized — except that we must be remembered that 50 000 is just a fraction of all already existing French words (total number supposed to be around 200 000 words). Each experiment generates a different map, but the textures of the maps look similar (black and white patterns).

Therefore, we have proposed and developed various optimizations / simplifications such as to keep only the best 100 scores (D2) for each word (instead of 50 000 score values), and to compute the self-organizing map using a stochastic crystal growing algorithm, instead of the classical but costly Kohonen algorithm:

1. A first word is selected and associated to the node at the center of the map.
2. One of its neighbor nodes is randomly selected. The best matching word is found: its distance to the already placed neighbor words is minimum (the summation of all D2 distances). Its weights are adjusted (alpha = 1.0, beta = 0.0).
3. Repeat from 2 until last word.

The final result is not the result of a global optimization process, but the duration of the (self-) learning is reduced to 40 minutes for the 50 000 words.

Using this ultra-fast learning map, we build a bigram representation of the 50 000 words (Fig. 7 & 8). Note that we also changed the number of neighbors, from 4 to 6 (hexagonal lattice), following the original formulation of the SOM [20]. This allows for a more compact map, with less frontiers and discontinuities. Also the hexagonal lattice appears to be more biologically plausible, and more efficient. Our ultra-fast SOM shares a number of similarities with the SOM of symbol strings [25], a much earlier work. However, among other differences, where the SOM of symbol strings involves successive training and growing phases, our proposal integrates learning and growing in one step.

Fig. 8. Enhancement of the (fig. 7) cortical map. The neighborhood size is 6, e.g., the word "sémantique" has 6 neighbors: "satanique, océaniques, mécaniques, cinémathèque, sémantiques, quasiment".

4 Conclusion

Open bigrams (OB) allow an over-coding of the orthographic form of words that facilitates recognition. OB coding favors same length words (i.e., neighbors of similar lengths). Using OB description, a cortical map has been built in order to visualize (the most frequent) 50 000 French words. This visualization of the cortical representation of (OB) words is highly pedagogic, allowing to really appreciate the fact that neighbor words are somewhat different from what we would naively think (i.e., letter-based distance). In future work, we may consider weigthed metrics in the bigramic space, taking into consideration their uncertainty. The incertainty of a bigram may simply be defined by a bayesian approach based on the counts of the bigram and the letter frequency. The most informative bigrams(x,y) are the ones with small probability that y follows x. Then, if our assumption is correct, languages may have evolved to separate words in the bigrammic space according to distance mostly based on the most informative bigrams. A weighted cosine metrics shall still be fast enought to compute such soft bigrammig map.

A realistic developmental database that takes into account the order of presentation of the words to the children would certainly generates a different kind of maps [19], less optimal (because neighbor words may be seen at different ages and end up in very different locations on the map), but closer to biological cortical map.

It is important to remember that the ultra-fast learning allows only for a local optimization and does not take into account the sampling frequency of the learning samples (each sample is represented on the map). Last, but not least, this ultra-fast learning is very important since it allows to consider the implementation of the Theory of neuronal Cognition [21-23]. In this case, the difficulty is no more in the learning duration, but in the availability of the learning data for each of the 500 cortical maps.

Acknowledgements. Work supported by the French Research Agency ANR 2010-CORD-013 "Cognilego — From pixels to semantics: a cognitive approach".

References

1. Impedovo, S.: More than twenty years of advancements on Frontiers in Handwriting Recognition. Pattern Recognition. In Press, Available online 12 June 2013
2. Chen, W., Gader, P., Shi, H.: Lexicon-driven handwritten word recognition using optimal linear combinations of order statistics. IEEE Trans. Pattern Anal. Mach. Intell., 21 (1), 77-82 (1999)
3. Salome, J., Leroux, M., Badard, J.: Recognition of cursive script words in a small lexicon. In: Proc. of ICDAR 2011, pp. 774-782 (1991)
4. Cho, W., Lee, S., Kim, J. H.: Modeling and recognition of cursive words with hidden Markov models. Pattern Recognition, 28 (12), 1941-1953 (1995)
5. Xue, H., Govindaraju, V.: Incorporating Contextual Character Geometry in Word Recognition. In: Proceedings of the Eighth International Workshop on Frontiers in Handwriting Recognition, pp. 123-127 (2002)

6. Oh, I-S., Suen, C. Y.: A class-modular feedforward neural network for handwriting recognition. Pattern Recognition, 35 (1), 229-244 (2002)
7. Senior, A.W., Fallside, F.: An off-line cursive script recognition system using recurrent error propagation networks. In: Proc. Third Intl. W. F. Hand-writing Recog, 132-141 (1993)
8. Laaksonen, J.: Subspace classifiers in recognition of handwritten digits, PhD thesis, Helsinki University of Technology (1997)
9. Graves, A., Schmidhuber, J.: Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks. In: Bengio, Y. Schuurmans, D. Lafferty, J. Williams, C. K. I., Culotta, A. (eds.), Advances in Neural Information Processing Systems 22 (NIPS'22), Vancouver, BC, pp. 545-552 (2009)
10. Graves, A., Liwicki, M., Fernandez, S., Bertolami, R., Bunke, H., Schmidhuber, J.: A Novel Connectionist System for Improved Unconstrained Handwriting Recognition. IEEE Trans. Pattern Analysis and Machine Intelligence, 31 (5), 855-868 (2009)
11. Ciresan, D. C., Meier, U., Gambardella, L. M., Schmidhuber, J.: Convolutional Neural Network Committees For Handwritten Character Classification. In: Proc. of ICDAR 2011, Beijing, China, pp. 1135-1139 (2011)
12. Ciresan, D. C., Meier, U., Schmidhuber, J.: Multi-column Deep Neural Networks for Image Classification. In: IEEE CVPR, pp. 3642-3649 (2012)
13. LeCun, Y., Bottou, L., Bengio, Y., Haner, P.: Gradient-based learning applied to document recognition. Proc. IEEE, 86, pp. 2278-2324 (1998)
14. Dehaene, S., Cohen, L., Sigman, M., Vinckier, F.: The neural code for written words: A proposal. Trends in Cognitive Sciences, 9, 335-341 (2005)
15. Whitney, C., Bertrand, D., Grainger, J.: On coding the position of letters in words: A test of two models. Experimental Psychology. 59(2), 109-114 (2012)
16. RIMES : Reconnaissance et Indexation de données Manuscrites et de fac-similÉS / Recognition and Indexing of handwritten documents and faxes `http://www.rimes-database.fr`
17. Menasri, F., Louradour, J., Bianne-Bernard, A-L., Kermorvant, C.: The A2iA French handwriting recognition system at the Rimes-ICDAR2011 competition. In Document Recognition and Retrieval Conference, Edited by Chien, L-C.; Lee, S.-D.; Wu, M. Hsien. Proceedings of the SPIE, Volume 8297, 8 pp. (2012)
18. Ortga, É., Lété, B.: eManulex: Electronic version of Manulex and Manulex-infra databases. (2010). `http://www.manulex.org`
19. Dufau, S., Lété, B., Touzet, C., Glotin, H., Ziegler J., Grainger, J.: Developmental Perspective on Visual Word Recognition: New Evidence and a Self-Organizing Model. European Journal of Cognitive Psychology, 22:5, 669-694 (2010)
20. Kohonen, T.: Self-organizing maps, Third Extended Edition, Springer, 2001.
21. Touzet, C.: Why Neurons are Not the Right Level of Abstraction for Implementing Cognition. BICA 2012 : Annual Int. Conf. on Biologically Inspired Cognitive Architectures, Palermo, Italy, pp. 317-318 (2012)
22. Touzet, C.: The Illusion of Joy. In: J. Schmidhuber, K.R. Thórisson, Looks M. (Eds.). Artificial General Intelligence 2011. Springer-Verlag LNAI 6830, pp. 357-362 (2011)
23. Touzet, C.: Consciousness, Intelligence, Free-Will? The answers from the Theory of neuronal Cognition. La Machotte Ed., Auriol, France (2010) (in French).
24. Bluche, T., Louradour, J., Knibbe, M., Moysset, B., Benzeghiba, F., Kermorvant, C.: The A2iA Arabic Handwritten Text Recognition System at the OpenHaRT2013 Evaluation. Submitted (2014)
25. Kohonen T. and Somervuo P.: Self-Organizing Maps of Symbol Strings with Application to Speech Recognition. Proc. of WSOM'97, Espoo, FI, pp. 2-7 (1997)