A Neural Field Model of Sequence Perception

Andrew P. Valenti (andrew.valenti@tufts.edu)

Bradley Oosterveld (bradley.oosterveld@tufts.edu)

Matthias J. Scheutz (matthias.scheutz@tufts.edu)

Human-Robot Interaction Laboratory, Tufts University, 200 Boston Ave.
Medford, MA 02155 USA

Abstract

We show how temporal and spatial information can be represented as stable patterns in a dynamical system. It is hypothesized that human perception and knowledge representation arise from such patterns. We describe how these patterns can be used in a neurologically-inspired model of speech perception, and show how word recognition arises when phonemes and their position in a word are mapped onto activation patterns. These activation patterns are used to identify the set of words whose prefix is the corresponding sequence, consistent with the "cohort"-based model of word recognition.

Keywords: dynamical systems; neural fields; speech perception; neurological plausibility; word-recognition; Cohort model

Introduction

The brains of all animals encode and process sensory input acquired from the environment. Sensory input, regardless of modality, is encoded as temporal and spatial patterns, and a superior form of pattern processing has evolved in humans coinciding with the expansion of the neocortex. In this brain structure, several essential cognitive processes (e.g., visual, auditory, speech) engage in processing (Koch, 2004; Mattson, 2014) which includes not only recognizing patterns, but classifying them (Grossberg, 2005). In addition, different members of a particular sensory input category, e.g., the phoneme "ə', are mapped to the same pattern to allow for invariance in speech perception across multiple speakers (Kleinschmidt & Jaeger, 2015). Consistent with these hypotheses, our model uses patterns of activation in a neural field to represent sequences of states, in this case in the context of perceiving words.

Model Architecture

The human neocortex consists of six layers of tissue containing approximately 10^{10} neurons. A column of tissue could be represented mathematically as a neural field which forms patterns of activation through interaction with other fields. From this cortical information cognitive processing emerges (Amari, 1977).

In a previous model of speech perception (Valenti, Brady, Scheutz, Holcomb, & Pu, 2016), a single neural field layer was modeled as a $32x32$ grid of bidirectionally connected units whose dynamics allow the field to fall into a stable equilibrium pattern. In the present model, we use a neural field of similar size and design.

Input is presented to our model (Figure 1) as a one-hot vector that is the size of the number of atomic categories in the input domain, i.e., 61 phonemic labels. This input vector (I) is fully connected to the neural field layer (F) by the input weights (W_i). F, described in the next section, is connected to the output vector (O) by the fully connected output weights (W_o).

State representation

Our model represents a state, or position in a known sequence, as the activation of F, which contains neurons connected to form a neural field. Each neuron in F is connected to each of its neighbors with weights (W_{mh}) based on a Mexican hat function whose input is the Euclidean distance between a pair of units. These weights create an on-center off-surround activation pattern, where the closest neighbors provide a positive influence on activation, further neighbors a negative influence, and the furthest no influence. As described in (Amari, 1977), after repeated updates units connected as such will fall into a stable equilibrium and subsequent updates will not result in a change in activation for any of the units in the layer.

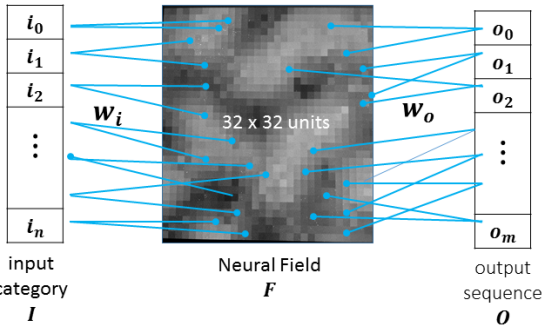

Figure 1: Diagram of model architecture.

The model represents sequences in the input domain through combinations of these unique equilibrium states. As our model receives input, it combines its current equilibrium, which represents the inputs it has seen up to this point, with an equilibrium that represents the next input character. These combined equilibriums then settle to form a new equilibrium. The process can be described as follows:

$$F_t = \sigma(settle(W_{mh}, (I \cdot W_i + F_{t-1})))$$

The previous and next equilibriums are summed and their result is allowed to settle into a stable equilibrium. The settling process is a sequence of repeated activations, where each neuron's activation is updated based on its neighbors until the updates no longer cause a change in activation.

The activation of the settled state is then normalized to the range $[0, 1]$ through the application of the squashing function σ: $\sigma(x) = max(0, \frac{x}{x+1})$.

This normalized state F_t represents the sequence of inputs that the model has received up to this point. For a given sequence of inputs a single state in the field is produced, and that state can only be reached by that sequence of inputs. A visualization of the combination process is show in Figure 2.

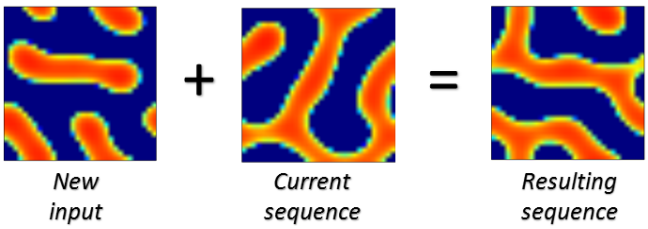

Figure 2: The addition of an input equilibrium state to a previous equilibrium state, after settling, produces a third equilibrium representing the sequence of states up to that point.

State interpretation

For each unit in the input domain, we choose an arbitrary equilibrium generated from a random seed and train the input weights to produce that equilibrium given the corresponding input.

At any given point, the activation of the units in the neural field F represents the sequence of inputs that the model has been presented. This activation pattern must somehow be grounded in a human-interpretable fashion. This grounding depends on the information that the model receives and what that information means to humans.

In our development of the model we have focused on speech as our input. Speech is composed of a small set of atomic units. These units are combined into a much larger set of sequences, where each unique sequence has a unique meaning. In our evaluation we consider speech at the phonemic and word levels, so this structure fits nicely with the representational capabilities of our model.

We draw the inspiration of the grounding in our system from the Cohort model of word recognition (Marslen-Wilson, 1987). In the Cohort model, as a sequence of speech segments is heard by the listener, every word that begins with that prefix of segments is activated. This set of activated words is referred to as the cohort of relevant words. As more of the word is recognized the size of the cohort shrinks until a single word remains.

We train the W_o so that the output of *O* behaves in a similar fashion. *O* is a vector that is the size of the total number of words, (unique sequences of phones), with which the model has been presented. Each element in *O* represents a word, and a positive activation at a given element means that the current state of *F* represents the word, or a prefix of the word, at the given index. As the sequence of inputs increases in length, the number of words that share the unique prefix decreases.

This behavior is achieved through the training of W_o. We use a version of the perceptron learning rule, seen below, to train the single layer perceptron whose activation is found in *O*.

$$\Delta W_o = \eta \cdot F_t \cdot (Target - F_t \cdot W_o)$$

Where η is the learning rate and *Target* is the ground truth, a vector the size of *O* with positive activations in the elements that correspond to the set of words for which the sequence represented by the F_t is either a prefix or the word itself.

Our model represents the incremental recognition of a word, presented as a sequence of phonemes, by outputting the set, (i.e., cohort), of known words for which the current sequence of phonemes is a prefix. As more words are presented, the cohort of positive activations in *O* shrinks. When the input to our model contains multiple words, e.g., a sentence, the model is also able to determine the boundaries between those words. Since it is trained on word level sequences, the model is able to detect when a sequence of incrementally presented phonemes no longer represents a word in its lexicon. As phonemes are presented, the size of cohort of activated units in *O* does not increase. To detect the end of a word the model looks for values of *O* whose activations are not a subset of the activation of *O* from the previous input. In these cases the model knows that end of the word has been reached, and *F* is reset to its initial starting state; the most recent input is presented again, starting the next word.

Model Evaluation

The TIMIT corpus (Garofolo, Lamel, Fisher, Fiscus, & Pallet, 1993) provides a set of sentences which are annotated at the word and phoneme level; we use these annotations in the evaluation of our model. TIMIT uses 61 phonemes as the atomic units in its transcriptions. In our evaluation, we define *I* as a 61 unit vector each of whose elements represents one of these phonemes. There are $7,368$ words in the TRAIN subset of the corpus, so *O* is defined as a $7,368$ element vector, with each element representing a word. We chose the TRAIN set as our lexicon, since it is larger than the TEST set, and since we are modeling perception and not lexicon acquisition, our evaluation is done on that same set.

The weights of the input and the decoder perceptrons were trained until the error terms fell below a threshold value. These thresholds were chosen by hand by the authors, and a more rigorous investigation on how optimal values may be reached is needed.

Results

Once the model was trained, the entirety of the TRAIN set was presented. The $65,529$ phonemes were presented in the order in which they were found in the corpus. After the presentation of a phoneme, the activation of *O* was compared to the ground truth. For 99.1% of the words in the lexicon, the activation of *O* matched the ground truth for every phoneme in that word.

The model was artificially reset to a default state at the end of every word so that errors in the perception of one word did not affect the perception of other words. However, each time before the model was reset, the first phoneme of the next word was presented. In 82.5% of cases, the model detected the word level transition by testing if the activation of *O* was a subset of the activation of *O* from the previous phoneme.

Conclusion

We have constructed a model in which the representation of discrete sequences as patterns of activation modeled in a more neurologically plausible fashion than the representations in previous models e.g., as a vector of neurons. We have evaluated the model in the speech perception domain and noted that it recognizes sets of words consistent with the access stage of the Cohort model of word recognition (Marslen-Wilson, 1987). Our model is consistent with the hypothesis that the neocortex receives and processes patterns of information in the same way regardless of whether the sensory input is visual, auditory, or tactile. Thus we believe our model is applicable to these cognitive domains as well.

The model assumes a perfect, invariant presentation of input data which is believed to occur in later stages of cognitive processing, e.g., the IT layer in the visual cortex. Further development of the model will introduce uncertainty in the input during training to assess how well the model can generalize when presented with lower level input.

References

Amari, S. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological cybernetics*, *27*(2), 77–87.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., & Pallet, D. S. (1993). DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon technical report n*, *93*.

Grossberg, S. (2005). Adaptive resonance theory. In L. Nadel (Ed.), *The encyclopedia of cognitive science* (1st ed.). Wiley.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, adapt to the novel. *Psychologicial Review*, *122*(2), 148–203.

Koch, C. (2004). *The Quest for Consciousness: a Neurobiological Approach* (1st ed.). Englewood, CO: Roberts and Company Publishers.

Marslen-Wilson, W. D. (1987). Functional parallelisim in spoken word-recognition. *Cognition*, *25*, 71–102.

Mattson, M. (2014). Superior pattern processing is the essence of the evolved human brain. *Frontiers in Neuroscience*, *8*(265). doi: 10.3389/fnins.2014.00265

Valenti, A. P., Brady, M. C., Scheutz, M. J., Holcomb, P. J., & Pu, H. (2016). A neural field model of word repetition effects in early time-course ERPs in spoken word perception. In A.Papafragou, D. Grodner, D. Mirman, & J. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 2765–2770).