O Modelo HTM e Sua Aplicação No Processamento de Linguagem Natural

Felipe Rayel

***Resumo*—A Memória Temporal Hierárquica (HTM) é uma teoria de aprendizagem de máquina que tenta modelar o neocórtex cerebral, órgão responsável pelas tarefas cognitivas mais complexas nos mamíferos. Neste trabalho é apresentado como o aprendizado ocorre em uma rede HTM e sua aplicação prática na solução do problema de *grounding* de palavras isoladas no processamento de linguagem natural.**

***Palavras-Chave*—Hierarquical Temporary Memory, conexionismo, predição, Numenta, NuPIC, aprendizagem de máquina, Cortical.IO, representação esparsa distribuída, linguagem natural.**

I. Introdução

Para o ser humano, realizar certas tarefas sempre foi muito simples e natural, como por exemplo, o reconhecimento de padrões visuais, compreensão de linguagem, reconhecimento pelo tato. Já os computadores, atualmente possuem grande dificuldade em realizar estas atividades. A dificuldade está em entender a forma como esta inteligência deve ser programada.

O uso de redes neurais artificiais trouxe um grande avanço nesta área, tornando possíveis certas coisas que antes eram impraticáveis [1]. Porem a máquina ainda está longe de alcançar um desempenho próximo ao do ser humano.

O que dá ao ser-humano esta grande capacidade de reconhecimento e adaptação é o neocórtex cerebral. O neocórtex é a área mais evoluída do córtex e recobre os lobos frontais dos primatas e todos os mamíferos. Tem espessura de uns dois milímetros e é composto por aproximadamente 30 bilhões de neurônios distribuídos em seis camadas [2].

Sabe-se que o neocórtex tem um padrão de conexões neurais e há evidências que sugerem que um conjunto de algoritmos distintos são usados para executar as diferentes funções da mente humana [3].

Inspirado biologicamente no neocórtex, Jeff Hawkins descreveu em seu livro "*On Intelligence*" em 2004 uma teoria para o funcionamento do cérebro baseado em um algoritmo de aprendizagem cortical (CLA).

Este algoritmo foi usado para definir uma tecnologia de aprendizagem de máquina chamada de Memória Temporal Hierárquica (HTM).

A HTM tem a pretensão de tornar possível a construção de máquinas que se aproximem ou até mesmo superem o desempenho humano para a realização de tarefas cognitivas.

Em 2005 Jeff Hawkins criou a Numenta. Uma empresa que tem a missão de implementar a tecnologia HTM e explorar seu uso comercial e científico.

II. Memória Temporal Hierárquica

Todo sistema inspirado no modelo de arquitetura do neocórtex pode ser considerado uma rede neural. Portanto as redes HTM podem ser consideradas uma nova forma de rede neural, seguindo o paradigma conexionista [4].

A. Neurônios

Na rede HTM, os neurônios são chamados de células e estão organizadas em colunas, camadas e regiões distribuídas em uma hierarquia [4].

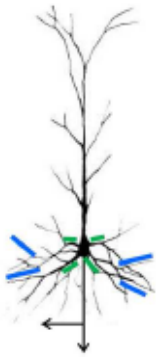

Fig. 1. Modelo de um neurônio piramidal típico do neocórtex.
Fonte: Numenta (2011). [4]

Como demonstra a figura 1, o neurônio biológico possui dois tipos de dendritos onde ocorrem as sinapses. As proximais e as distais.

Nas sinapses proximais (marcadas em verde) a conexão ocorre verticalmente com os neurônios da mesma coluna. São poucos e de grande influencia na ativação do neurônio.

Nas sinapses distais (marcadas em azul) a conexão ocorre de forma horizontal com outras colunas. São numerosas e tem pouca influência na ativação do neurônio por estarem mais distantes. Se um número suficiente de sinapses distais se ativarem simultaneamente, poderá ocorrer um pico de ativação. Pode-se dizer então que estes dendritos atuam como detectores de limites de coincidência [5].

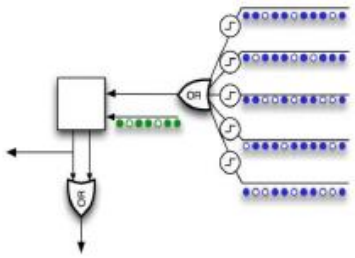

Fig. 2. Modelo de um neurônio HTM.
Fonte: Numenta (2011). [4]

Em uma célula HTM (figura 2) existe apenas um único dendrito proximal. A ativação da célula ocorre através da soma linear destas sinapses.

As conexões distais são armazenadas em uma lista de segmentos em cada célula HTM.

Cada um destes segmentos funciona como um detector de limite, ou seja, se o número de sinapses que estão ativas em qualquer um dos segmentos armazenados está acima de um limiar pré-especificado, o segmento se ativa e as células associadas a este segmento entram em estado preditivo [4].

As células HTM podem ter três estados de saída: ativo pela entrada de alimentação (ativação), ativo pela entrada lateral (predição) e inativo.

A sinapse em uma célula HTM tem peso binário que e é definido através dos conceitos de sinapse potencial e da permanência. A sinapse potencial representa todos os axônios que passam perto o suficiente de um segmento dendrítico ao ponto de formar uma sinapse. A permanência é representada por um valor escalar atribuído a cada sinapse. Esse valor varia de 0,0 (totalmente desconectada) a 1,0 (totalmente conectada). O aprendizado envolve aumentar e decrementar a permanência de uma sinapse [4].

As regras utilizadas para aumentar ou decrementar a permanência de uma sinapse segue as regras de aprendizagem Hebbiana [6]. Por exemplo, se uma célula está em um estado ativo por consequência de um segmento dendrítico receber entradas acima de um limiar, então as sinapses ativas que contribuíram para a ativação da célula tem sua permanência aumentada. Já as sinapses que estão ativas e não contribuíram tem sua permanência diminuída [4].

B. Hierarquia

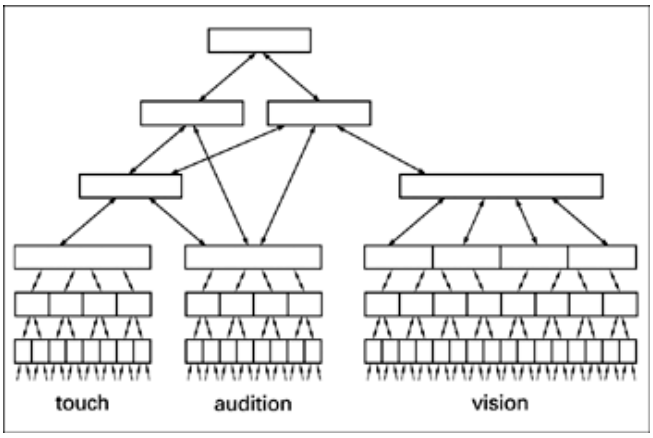

Fig. 3. Composição hierárquica das regiões HTM.
Fonte: On Intelligence (2004). [14]

Uma rede HTM é composta por regiões organizadas em uma hierarquia de camada (figura 3). Esta forma de organização a torna mais eficiente, pois reduz o tempo de treinamento e utilização da memória. Padrões aprendidos em níveis inferiores são combinados de novas formas em níveis mais altos. Por exemplo, ao aprender uma nova palavra, você não precisa reaprender as letras [4].

É possível combinar várias redes HTM, caso existam dados de mais de uma fonte ou sensor. Por exemplo, em um nível inferior, uma rede pode processar informações auditivas e a outra pode processar informações visuais. Mas em um nível mais alto, estes padrões podem se juntar para gerar um significado mais amplo.

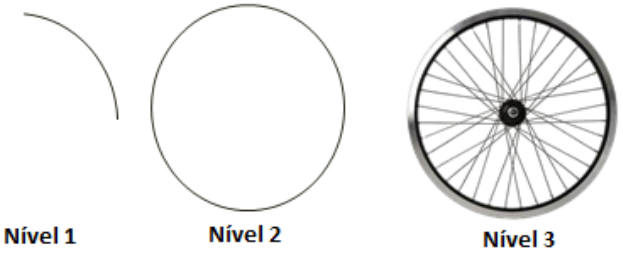

Fig. 4. Exemplo de padrões visuais em três níveis de complexidade que podem ser armazenados em uma camada de uma hierarquia HTM.

A figura 4 ilustra um exemplo de reconhecimento de uma roda. No nível mais baixo da hierarquia (nível um), estão armazenadas apenas informações de pequenas seções do campo visual como bordas, cantos etc. Já no nível médio (nível dois) estão as combinações dos padrões do nível anterior. Formas mais complexas podem ser reconhecidas como curvas, texturas, etc. No nível mais alto (nível três) estão as combinações dos padrões de nível médio e podem representar caraterísticas de objetos completos como uma cabeça, um carro, uma casa, etc. Desta forma, a hierarquia torna-se útil na medida em que permite aprender novos objetos de alto nível sem a necessidade de reaprender seus

componentes.

C. Representação Distribuída Esparsa

Sabe-se que no neocórtex, a informação é representada por uma pequena porcentagem de neurônios ativos [15]. Para que a informação possa ser representada de forma semelhante em uma rede HTM, é necessário que esta informação passe por um algoritmo que a transforme em uma representação distribuída esparsa. Este algoritmo é conhecido dentro da rede HTM como agrupador espacial.

Por exemplo, se a informação captada por um sensor gerar uma sequência de representações com 20.000 bits cada, a porcentagem de bits ativos pode variar entre elas, ou seja, hora podem aparecer 5.000 bits "1", outra hora pode aparecer 4.000, e assim por diante.

Para que esta porcentagem se mantenha constante entre ao longo do tempo, o agrupador espacial converteria esta entrada em uma representação interna de 10.000 bits, dos quais apenas 2% estariam ativos ao mesmo tempo, independente de quantos bits de entrada forem "1". Assim, com a variação do tempo, a representação interna sempre terá 200 bits ativos (2%) entre os 10.000 que poderiam estar.

Esta distribuição garante uma pequena variação de células ativas entre uma sequencia de entradas recebidas pela rede HTM longo do tempo.

Embora possa parecer que há grande perda de informação neste processo, os efeitos práticos desta redução são mínimos [4].

D. Regiões

Assim como o neocórtex é distribuído ao longo de uma grande folha de tecido neural com poucos milímetros de espessura, uma região HTM é composta por inúmeras células organizadas em uma matriz bidimensional de colunas.

Cada coluna em uma região é conectada a um subconjunto de bits de entrada de uma representação esparsa que pode sobrepor várias colunas (Figura 5). O resultado disto é a diferente ativação das colunas de acordo com a variação dos padrões de entrada. Colunas com maior ativação inibem colunas com menor ativação.

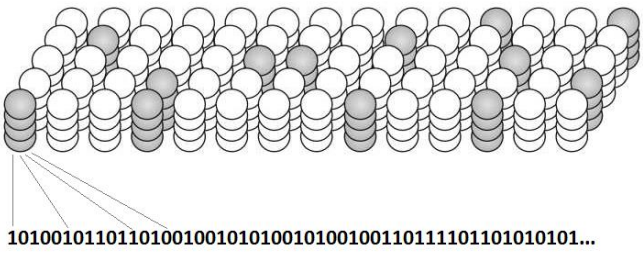

Fig.5. Uma região HTM com quatro células por coluna e conexão de uma representação esparsa à uma coluna. Fonte: Numenta (2011). [4]

Cada coluna é composta por várias células. Todas as células de uma coluna tem a mesma entrada de alimentação. Cada uma destas células em uma coluna pode estar ativa ou inativa. Desta forma pode-se representar a mesma entrada de várias formas diferentes, ou seja, a mesma entrada pode ser representada em diferentes contextos, apenas ativando diferentes células na mesma coluna.

E. Aprendizagem

O tempo tem papel fundamental na aprendizagem dos dados captados pelos sentidos. É através da mudança de padrões durante um tempo, que a inferência é possível. Por exemplo, só conseguimos identificar objetos através do tato depois de manipula-lo por alguns segundos.

O aprendizado em uma rede HTM funciona da mesma forma. A rede aprende a reconhecer sequencias de padrões de entrada que mudam no tempo. Cria-se então um modelo destas transições e isto é armazenado de forma distribuída em cada célula [4].

Em uma rede HTM não é necessário haver um período para treinamento. O aprendizado é on-line, ou seja, assim como em um sistema biológico, o aprendizado é contínuo a partir de cada nova entrada.

Por outro lado, a fase de treinamento pode ser desativada nos níveis mais baixos da hierarquia, permitindo que haja aprendizagem apenas nos níveis mais elevados.

Ou seja, após a rede aprender a reconhecer as estruturas básicas do seu ambiente, torna-se desnecessário consumir tempo tentando aprender algo novo, sendo que todos os elementos mais simples já são conhecidos.

Se por acaso a rede receber novos padrões de baixo nível inéditos, o tempo para aprendê-los será muito maior.

Essa característica é vista nos seres humanos e é explicada pelo Neuro-Darwinismo de Edelman [7]. Por exemplo, após aprendermos as palavras em nosso idioma, torna-se mais difícil aprender idiomas estrangeiros que tenham letras ou fonemas desconhecidos.

Após os padrões serem aprendidos, a rede HTM pode realizar inferências sobre as novas entradas.

Quando a HTM recebe uma nova entrada, compara com os padrões reconhecidos em momentos anteriores e tenta localizar esta sequência nas aprendizagens feitas no passado.

Por exemplo, ao ouvir a primeira sílaba de uma palavra, a possibilidade de palavras que podem se formar é imensa. Ao ouvir a segunda sílaba, a possibilidade de palavras reduz significativamente, e assim por diante.

Todo reconhecimento sensorial tem um grande problema. A experiência sensorial sempre parece ser nova. Por exemplo, uma palavra dita por pessoas diferentes, nunca estimula o sistema auditivo da mesma forma. Sempre haverá diferença na pronúncia, timbre de voz, ruídos. Etc.

A região HTM também tem este problema, ainda mais por se tratar de um sistema onde os dados sensoriais precisam ser digitalizados. Ao captar as informações dos sentidos, as entradas nunca se repetirão exatamente.

A solução para este problema está no uso de representações

distribuídas esparsas. Com este tipo de representação, só há a necessidade de combinar uma parte do padrão reconhecido e assegurar que esta combinação seja significante [4].

F. Predição

Regiões HTM armazenam as transições entre representações distribuídas esparsas. Com isso, aprende-se a reconhecer sequências de padrões.

Quando uma HTM identifica uma sequência através da sua entrada atual e as últimas entradas, torna-se possível predizer os próximos padrões que estão por vir.

A predição é possível, pois quando uma célula se ativa, realiza conexões com as células que estavam ativas momentos antes (Figura 6). Este processo assemelha-se às sinapses distais que acontecem em um neurônio biológico.

Fig.6. Transição entre padrões e conexões de células que estavam ativas momentos antes.

Através destas conexões, as células HTM podem predizer quando elas se tornarão ativas buscando em sua lista de sinapses (Figura 7).

Fig.7. Padrão reconhecido com base em conexões anteriores resulta em predição de células

Quando uma coluna de uma região HTM torna-se ativa, é verificado em todas as células desta coluna se há alguma célula em estado preditivo (figura 8). Se houver, então apenas estas células se tornarão ativas, caso contrário todas se tornarão ativas. Isto significa que se uma entrada é esperada, o sistema confirma esta predição ativando estas células. Se o padrão de entrada é inesperado, o sistema ativa todas as células, tornando assim todas as possíveis interpretações válidas.

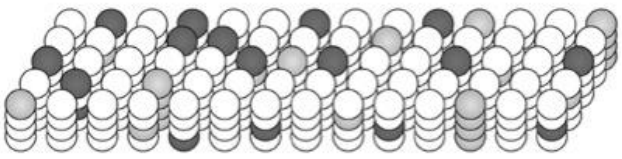

Fig. 8: Região HTM com neurônios pretos ativados pela confirmação da predição horizontal e cinzas ativados diretamente pelo input vertical.
Fonte: Numenta (2011) [4].

Resumindo, ao receber uma entrada de bits de uma representação esparsa, um conjunto de colunas se ativa. Dentro destas colunas, uma ou mais células se ativarão. Estas células farão com que outras células entrem em estado preditivo, pois possuem conexões aprendidas através de outras células da região. As ativações que ocorreram através destas conexões significam uma predição do que pode acontecer em seguida. Ao receber a próxima entrada, haverá uma ativação de outro conjunto de colunas. Se outra coluna tornou-se ativa inesperadamente, significa que ela não foi predita por nenhuma célula, então todas as células se ativarão. Se a coluna ativa tem ao menos uma célula em estado preditivo, apenas estas células se tornarão ativas [4].

A saída da região é um vetor que representa a atividade de todas as células, incluindo as células em estado preditivo.

Não há nenhuma central de armazenamento ou memória central para que os padrões sejam gravados. A memória é distribuída entre cada célula individualmente. Isto torna o sistema tolerante ao ruído e a erros.

III. Engine Cortical para Processamento de Texto

Uma das grandes dificuldades para que exista aprendizado real em maquinas, é a incapacidade de se extrair o significado de palavras e frases. O processamento de linguagem natural (NLP) é uma área de estudo que une esforços da ciência da computação, inteligência artificial e linguística com o objetivo de produzir interação entre computadores e linguagens humanas (naturais).

Com o propósito de tornar a linguagem natural compreensível para a máquina, foi inventado o *engine* cortical para processamento de texto [8].

Este *engine* foi baseado na teoria HTM e implementado pela empresa austríaca Cortical.IO. Atualmente é uma tecnologia proprietária e é oferecida na modalidade SaS (Software as Service) [8].

O *engine* cortical para processamento de texto fornece a possibilidade de criar expressões booleanas usando palavras baseadas em seu significado, além de permitir também calcular a similaridade semântica entre palavras.

Para que isto seja possível, o software de processamento utiliza um conceito chamado de *semantic fingerprint*, que é na verdade uma representação distribuída esparsa. Neste conceito, as palavras podem ser representadas atreves de uma impressão digital que captura todo o seu significado. Palavras diferentes que representem a mesma ideia ou a mesma coisa terão a mesma representação. Jamais coisas diferentes terão representações iguais, pois o *semantic fingerprint* é único [9].

Por exemplo, imagens e sons são representações semânticas que capturam diferentes dimensões de algum conceito. Se você estiver observando um jaguar ou ouvir seu rosnado, saberá que se trata de um animal. Mas ao ler a palavra jaguar, sem haver um contexto não será possível identificar de que se trata de um animal jaguar ou um carro Jaguar.

Na figura 9 fica clara a semelhança que existe entre a representação de duas palavras como Porsche e Jaguar. Ambas compartilham um contexto relacionado a carros, no entanto cada uma delas tem suas particularidades, por exemplo, o jaguar também representa um animal.

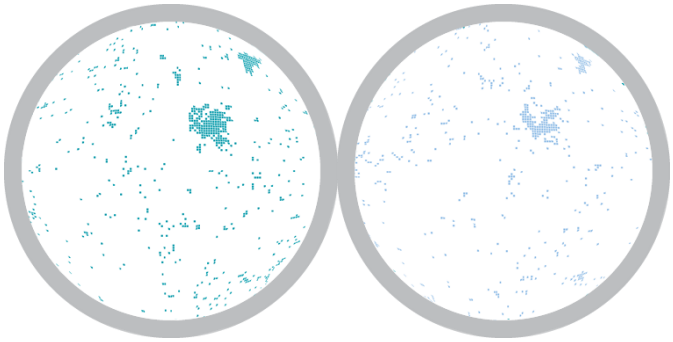

Fig.9. *Fingerprint* semântico das palavras Porsche e Jaguar.
Fonte: Cortical.IO (2014). [9]

A empresa Cortical.IO fornece (através de APIs) o acesso à sua base de *fingerpint* semântico. Todo o aprendizado já foi previamente realizado incialmente através da Wikipedia em vários idiomas (inglês, francês, alemão). Esta base de informações está em constante crescimento e atualização, garantindo sua evolução juntamente com evolução da linguagem [8].

IV. Experimento Aplicado ao Processamento de Linguagem Natural

Para o experimento, foi utilizado o software NuPIC (*Numenta Platform for Intelligent Computing*) que implementa uma rede HTM usando a linguagem Python e é distribuído sobre a licença GPL versão 3.

Este experimento com processamento de linguagem natural foi apresentado em um Hackaton organizado pela Numenta em 2013 na cidade de São Francisco. Como produto deste experimento, foi criado uma variação do NuPIC, chamado de NuPIC_NLP que une a capacidade de aprendizado e predição das redes HTM ao *engine* cortical para processamento de texto da empresa Cortical.IO através de chamadas remotas com a API CEPT [10].

A aplicação analisada consiste em associar pares de palavras baseadas em seu significado semântico codificado pelas suas representações esparsas.

Subutai Ahmad [10] utilizou o framework de associação definido pelo NuPIC_NLP para associar frases de três palavras e ensinar ao algoritmo cortical frases como "*cow eat grass", "dogs like sleep*". Após o treinamento de algumas sentenças, o software foi capaz de responder o que come uma raposa, sem nunca ter aprendido a palavra raposa anteriormente.

	Primeiro Termo	Segundo Termo	Terceiro Termo	Predição
1	cow	eat	grain	
2	elephant	eat	leaves	grain
3	goat	eat	grass	leaves
4	wolf	eat	rabbit	leaves
5	cat	likes	ball	
6	elephant	likes	water	ball
7	sheep	eat	grass	leaves
8	cat	eat	salmon	grass
9	wolf	eat	mice	rabbit
...				
30	fox	eat	?	rabbit

Fig.10. A predição do terceiro termo realizada em função do contexto aprendido pelos primeiros e segundos termos.

A figura 10 demonstra a predição como resposta da pergunta "o que come uma raposa", obtida após certa quantidade de treinamentos. Neste exemplo, o algoritmo cortical precisou associar semanticamente a palavra raposa com outras similares que foram treinadas anteriormente. Essa associação foi obtida pelo aprendizado feito pela rede HTM com o *fingerprint* semântico de cada palavra.

Ao tempo em que a predição era realizada, o algoritmo precisou aprender uma gramática primitiva de três palavras. Ao processar o verbo (Segundo termo) ele precisou lembrar-se do sujeito (primeiro termo).

A predição foi semanticamente correta ao analisar o verbo de cada frase. O algoritmo teve que prever a comida, se o verbo fosse "*eat*" e além disso, predizer o tipo certo de comida dependendo do animal.

Em um segundo exemplo (Figura 11) foi analisada a capacidade de distinção semântica entre carros e animais. Como resultado, a predição correta de que Ferrari é um carro e *Horse* é um animal.

	Primeiro Termo	Segundo Termo	Terceiro Termo	Predição
1	jaguar	is	car	
2	porsche	is	car	car
3	wolf	is	animal	car
4	fox	is	animal	animal
5	ferrari	is	?	car
6	horse	is	?	animal

Fig.11. A predição do terceiro termo evidenciada pela similaridade semântica dos primeiros termos.

V. Conclusões

O Modelo HTM é uma alternativa ao uso das redes neurais convencionais. Seu algoritmo cortical é robusto e capaz de resolver problemas como classificação, predição e detecção de anomalias.

A maioria das críticas ao modelo HTM são na verdade críticas ao seu autor Jeff Hawkins, devido a ele não ter um *background* acadêmico na área e ter como objetivo primário o uso da tecnologia para fins comerciais.

Sobre a teoria HTM e sua implementação no software NuPIC, pode-se dizer que constantes aprimoramentos tem sido realizados. Várias iniciativas foram feitas durante o ano de 2014 pela Numenta para popularizar o uso da tecnologia.

Jeff Hawkins publicou uma lista ideias que visam melhorar o algoritmo cortical inicialmente proposto. Estas ideias devem ser implementadas no NuPIC nos próximos anos e visam deixar a predição mais poderosa [11].

O *Engine* Cortical da empresa Cortical.IO, ao atribuir características semânticas para as palavras, parece tentar resolver o problema do *Grounding* para palavras isoladas. Porém este significado dado a elas estão lastreados em outras palavras e documentos. Não há um significado semântico baseado em experiências sensoriais como visão e audição.

O *engine* cortical pode encontrar similaridade semântica entre palavras de alguns idiomas, no entanto, ainda não há suporte ao idioma português.

As tentativas de se modelar o neocórtex cerebral são válidas e ajudam a entender como emergir inteligência em uma máquina. Porém, o neocórtex não pode ser visto como a única forma de cognição de alto nível. Em 2012 o laboratório de Ragsdale em Chicago confirmou uma antiga teoria de que as aves possuem uma estrutura com a mesma funcionalidade do neocórtex chamada de crista ventricular dorsal (DVR), ou seja, apesar das aves e mamíferos terem seguido ramos diferentes durante a evolução, parece que a seleção natural modelou diferentes arquiteturas cognitivas que se convergiram na solução do mesmo problema [12] [13].

REFERÊNCIAS

[1] D. E. Rumelhart, G. E. Hinton, R. J. Williams. *Learning representations by back-propagating errors*. Nature 323 (6088): 1986, 533–536.

[2] Dorland's. *Dorland's Illustrated Medical Dictionary* (32nd ed.). 2012, Elsevier Saunders. p. 1238

[3] Lui, J. H.; Hansen, D. V.; Kriegstein, A. R. *Development and Evolution of the Human Neocortex*. 2011. Cell 146 (1): 18–36.

[4] Numenta. (2011). *MEMÓRIA TEMPORAL HIERÁRQUICA Incluindo Algoritmos de Aprendizagem Cortical HTM* . [Online]. Disponível: http://numenta.com/assets/pdf/whitepapers/hierarchical-temporal-memory-cortical-learning-algorithm-0.2.1-pt.pdf

[5] G. Stuart, N. Spruston, M. Häusser, *Dendrites*, second edition.New York: Oxford University Press, 2008.

[6] Hebb, D.O. *The Organization of Behavior*. 1949, New York: Wiley & Sons.

[7] Edelman, Gerald Neural Darwinism. *The Theory of Neuronal Group Selection* (Basic Books, New York 1987)

[8] Cortical.IO (2014). *Technology* [Online] Disponível: http://www.cortical.io/technology.html

[9] Cortical.IO (2014). *Semantic Fingerprint*. [Online] Disponível: http://www.cortical.io/contexts_semantic.html

[10] S. Ahmad (2013). *Natural Language Processing*. [Online] Disponível: http://numenta.org/resources/hackathon/2013-10/nupic_nlp.pdf

[11] Numenta (2014) *New Ideas About Temporal Pooling*. [Online]. Disponível: http://github.com/numenta/nupic/wiki/New-Ideas-About-Temporal-Pooling

[12] Jennifer Dugas-Ford, Joanna J. Rowell, and Clifton W. Ragsdale. *Cell-type homologies and the origins of the neocortex*. Proceedings of the National Academy of Sciences, 2012

[13] HJ Karten, W Hodos. *A Stereotaxic Atlas of the Brain of the Pigeon:(Columba Livia)*. Johns Hopkins Press, 1967

[14] J. Hawkins, S. Blakeslee. *On Intelligence*. St. Martin's Griffin, 2005.

[15] Field, D. J. *Relations between the statistics of natural images and the response properties of cortical cells*. 1987. J. Opt. Soc. Am. A 4, 2379-2394.

[16] P. Tabacof . *O Modelo Cortical HTM e sua Aplicação na Classificação de Gêneros Musicais*. SICA 2011. Unicamp.