AI Platform Why Skymind? Solutions Case Studies About Resources

A.I. Wiki

Do you like this content? We'll send you more. Email SUBMIT

Artificial Intelligence Wiki

Search articles...

AI vs. ML vs. DL

Apache Spark & Deep Learning

Attention Mechanisms & Memory Networks

Automated Machine Learning & AI

AI & Autonomous Vehicles

Backpropagation

Bag of Words & TF-IDF

Clojure AI

Comparison of AI Frameworks

Convolutional Neural Network (CNN)

Data for Deep Learning

Datasets and Machine Learning

Decision Tree

Deep Autoencoders

Deep-Belief Networks

Deep Reinforcement Learning

Deep Learning Resources

Deeplearning4j

Denoising Autoencoders

Machine Learning DevOps

Differentiable Programming

Eigenvectors, Eigenvalues, PCA, Covariance and Entropy

Evolutionary & Genetic Algorithms

Fraud and Anomaly Detection

Generative Adversarial Network (GAN)

A Beginner's Guide to Attention Mechanisms and Memory Networks

> “ I cannot walk through the suburbs in the solitude of the night without thinking that the night pleases us because it suppresses idle details, much like our memory.- *Jorge Luis Borges* [1]

Attention mechanisms in neural networks serve to orient perception as well as memory access (you might even say perception is just a very short-term subset of all memory). Attention filters the perceptions that can be stored in memory, and filters them again on a second pass when they are to be retrieved from memory. Attention can be aimed at the present and the past.

Attention matters because it has been shown to produce state-of-the-art results in machine translation and other natural language processing tasks, when combined with neural word embeddings, and is one component of breakthrough algorithms such as Transformer and BERT, which is setting new records in accuracy in NLP. So attention is part of our best effort to date to create real natural-language understanding in machines. If that succeeds, it will have an enormous impact on society and almost every form of business.

Why “Attention”? Why Memory?

“Attention” is defined as the “active direction of the mind to an object.” Attention is about choice – it *is* choice. You make choices about how to direct your attention among the many social networks, YouTube videos, apps and websites that are trying to colonize your mind.

Glossary

Gluon

Graph Analytics

Hopfield Networks

Hyperparameter

Wiki Home

Java AI

Jumpy

Logistic Regression

LSTMs & RNNs

Machine Learning Algorithms

Machine Learning Demos

Machine Learning Software

Machine Learning Operations (MLOps)

Machine Learning Research Groups & Labs

Machine Learning Workflows

Machine Learning

Markov Chain Monte Carlo

Multilayer Perceptron

Natural Language Processing (NLP)

ND4J

Neural Network Tuning

Neural Networks

Open Datasets

Python AI

Questions When Applying Deep Learning

Radial Basis Function Networks

Random Forest

Recurrent Network (RNN)

Recursive Neural Tensor Network

Restricted Boltzmann Machine (RBM)

Robotic Process Automation (RPA) & AI

Scala AI

Single-layer Network

Skynet, or How to Regulate AI

Spiking Neural Networks

Neural networks make choices about which features they pay attention to. Attention is both the currency of Silicon Valley, and the currency of AI.

The word describes the mind's ability to allocate consideration unevenly across a field of sensation, thought and proprioception, to focus and bring certain inputs to the fore, while ignoring or diminishing the importance of others. So for neural networks, we're basically talking about credit assignment. And the two challenges with credit assignment are long-range dependencies (i.e. things that impact your predictions, but which happened a long time ago in a galaxy far, far way); and dealing with massive instances of data, like very large images.

One of the natural language processing problems that researchers have struggled with is how to link pronouns to antecedents, which resulted in this old joke, mimicking the call and response of a protest march:

> WHAT DO WE WANT?

> Natural language processing!

> WHEN DO WE WANT IT?

> Sorry, when do we want what?

A neural network armed with an attention mechanism can actually understand what "it" is referring to. That is, it knows how to disregard the noise and focus on what's relevant.

At any given moment, your mind concentrates on a subset of the total information available to it. For example, you are reading these words as a larger world flows around you: maybe you're in a room with traffic coursing outside, or in a plane and the pilot is making another annoying announcement about turbulence, but your focus is **HERE**.

Learn to build AI apps now »

This is important, because the field of sensation is wide (the totality of the state of the world at any moment is so wide as to be unknowable), the mind's bandwidth to process information is narrow, and some inputs are indeed more important than others, with regard to any given goal. Just as a student of Buddhism channels their own attention to attain enlightenment, or an artist channels the attention of others to evoke emotion, a neural network can channel its attention to maximize the accuracy of its predictions. Just as physical mechanisms, like two rotating gears interlocking at their teeth, serve in the transfer of motion, an attention mechanism serves in the transfer of information.

Stacked Denoising Autoencoder (SDA)

Strong AI vs. Weak AI

Supervised Learning

Symbolic Reasoning

Text Analysis

Thought Vectors

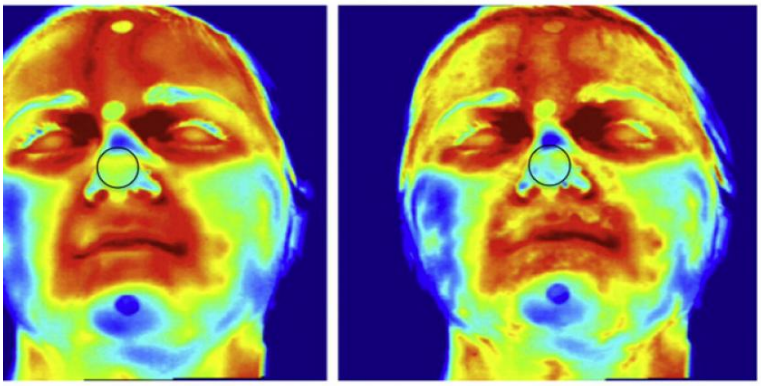
Unsupervised Learning

Deep Learning Use Cases

Variational Autoencoder (VAE)

Word2Vec, Doc2Vec and Neural Word Embeddings

In sum, algorithms can allocate attention, and they can learn how to do so, by adjusting the weights they assign to various inputs. Attention is used for machine translation, speech recognition, reasoning, image captioning, summarization, and the visual identification of objects. Imagine a heat map over a photo. The heat is attention.



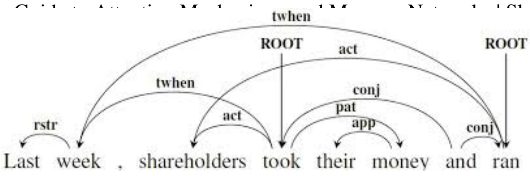
Beyond Word Vectors

One of the limits of traditional word vectors is that they presume that a word's meaning is relatively stable across sentences. This is not so. Polysemy abounds, and we must beware of massive differences in meaning for a single word: e.g. lit (an adjective that describes something burning) and lit (an abbreviation for literature); or get (a verb for obtaining) and get (an animal's offspring).

While projects like WordNet and algorithms such as sense2vec represent admirable attempts to disentangle the meanings of all words in English, there are many shades of meaning for a given word that only emerge due to its situation in a passage and its inter-relations with other words. Words are social creatures. Like humans, they derive much of their meaning from relationships.

So we need to expand the domain of inputs that an algorithm considers and include the context of a given word. The charm of linear algebra is that you can calculate many relationships at once; in this case, we are calculating the relationships of each word in a sentence to every other word, and expressing those variable relationships that suggest a word's meaning as a vector. We have to vectorize all the things. And we can do that with the attention mechanism.

In self-attention, or intra-attention, you might talk about the attention that words pay to each other within a sentence. For any given word, we seek to quantify the context that the sentence supplies, and identify which other words supply the *most* context with regard to the word in question. The directed arcs of a semantic dependency graph may give you an intuition of how words connect with one another across a crowded sentence, some rising above others to help define certain of their fellows.



What’s slightly more interesting is how the relationships embodied in a context vector can change our understanding of sentences.

Because sentences are deceptive. Soldiering along under the tyrannies of time and paper, sentences have lulled us into thinking that their meaning is linear, that it unfurls like a ribbon of print across the page. This is not true. In any sentence, some words have strong relationships with other words that are not bang-up next to them. In fact, the strongest relationships binding a given word to the rest of the sentence may be with words quite distant from it.

So it’s probably more fitting to think of sentences folding like proteins in three-dimensional space, with one part of a phrase curling around to touch another part with which it has a particularly strong affiliation, molecules pearling on a polymer.

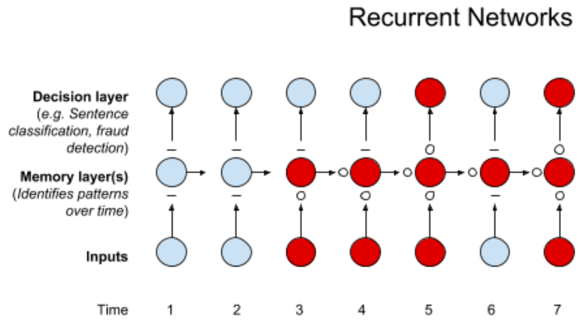
When we consider sentences as multi-dimensional objects twisting back on themselves, suddenly their ability to map to the obscure and stubborn syntax of real objects in space-time (whose causal relationships, correlations and influences can also be remote, or invisible) seems more intuitive, and it also aligns with certain approaches to neuroscience and the geometry of thought. In fact, attention can help us understand objects’ inter-relations in an image just as well as it aids us with natural-language processing.

Credit Assignment Among Available Features

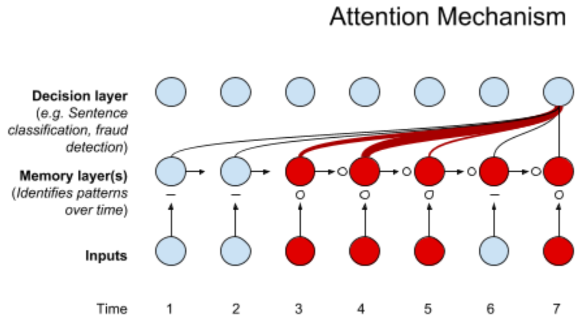
The fundamental task of all neural networks is *credit assignment*. Credit assignment is allocating importance to input features through the weights of the neural network’s model. Learning is the process by which neural networks figure out which input features correlate highly with the outcomes the net tries to predict, and their learnings are embodied in the adjusted quantities of the weights that result in accurate decisions about the data they’re exposed to.

But there are different ways to structure and channel the relationship of input features to outcomes. Feed-forward networks are a way of establishing a relationship between all input features (e.g. the pixels in a picture) and the predictions you want to make about the input (e.g. this photo represents a dog or a cat), and doing so all at the same time.

When we try to predict temporal sequences of things, like words in a sentence or measurements in a time series (e.g. temperatures or stock prices), we channel inputs in other ways. For example, a recurrent neural network like an LSTM is often used, since it takes account of information in the present time step as well as the context of past time steps. Below is one way to think about how a recurrent network operates: at each time step, it combines input from the present moment, as well as input from the memory layer, to make a decision about the data.

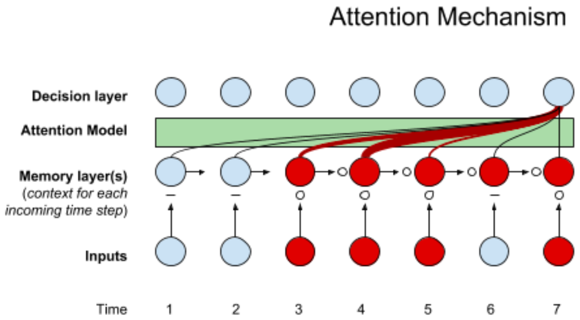

Recurrent Networks
Decision layer
(e.g. Sentence classification, fraud detection)
Memory layer(s)
(Identifies patterns over time)
Inputs
Time
1
2
3
4
5
6
7

RNNs cram everything they know about a sequence of data elements into the final hidden state of the network. An attention mechanism takes into account the input from several time steps, say, to make one prediction. It distributes attention over the hidden states of several time steps. And just as importantly, it accords different weights, or degrees of importance, to those inputs, reflected below in the lines of different thicknesses and color. In neural networks, **attention primarily serves as a memory-access mechanism.**


Attention Mechanism
Decision layer
(e.g. Sentence classification, fraud detection)
Memory layer(s)
(Identifies patterns over time)
Inputs
Time
1
2
3
4
5
6
7

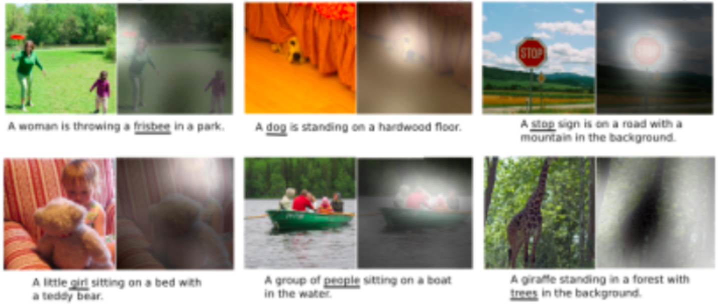
The original work on a basic attention mechanism represented a leap forward for machine translation. That advance, like many increases in accuracy, came at the cost of increased computational demands. With attention, you didn't have to fit the meaning of an entire English phrase into a single hidden state that you would translate to French.

Another way to think about attention models is like this:

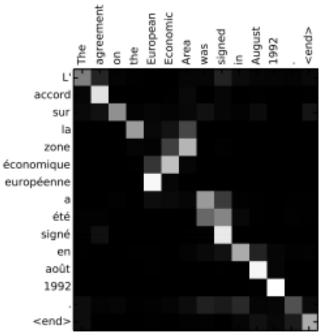


Let's say you are trying to generate a caption from an image. Each input could be part of an image fed into the attention model. The memory layer would feed in the words already generated, the context for future word predictions. The attention model would help the algorithm decide which parts of the image to focus on as it generated each new word (it would decide on the thickness of the lines), and those assignments of importance would be fed into a final decision layer that would generate a new word.

. Examples of attending to the correct object (*white* indicates the attended regions, *underlines* indicated the corresponding word)


A woman is throwing a frisbee in a park.
A dog is standing on a hardwood floor.
A stop sign is on a road with a mountain in the background.
A little girl sitting on a bed with a teddy bear.
A group of people sitting on a boat in the water.
A giraffe standing in a forest with trees in the background.

Above, a model highlights which pixels it is focusing on as it predicts the underlined word in the respective captions. Below, a language model highlights the words from one language, French, that were relevant as it produced the English words in the translation. As you can see, attention provides us with a route to interpretability. We can render attention as a heat map over input data such as words and pixels, and thus communicate to human operators how a neural network made a decision. (This could be the basis of a feedback mechanism whereby those humans tell the network to pay attention to certain features and not others.)



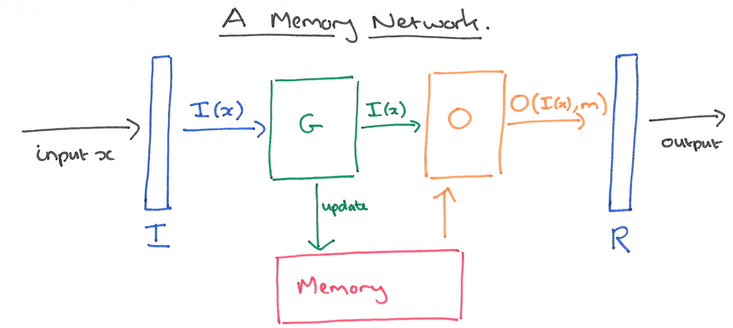
In autumn 2017, Google separated the attention mechanism from recurrent networks and showed that it could outperform RNNs alone, with an architecture called Transformer.

Memory Networks

You could say that attention networks are a kind of short-term memory that allocates attention over input features they have recently seen. Attention mechanisms are components of memory networks, which learn to access external memory storage rather than a sequence of hidden states in an RNN.

Memory networks are a little different, but not too. They work with external data storage, and they are useful for, say, mapping questions as input to answers stored in that external memory. Rather than surfacing the relevant features of an immediate experience amid the noise of perception, attention can pull a distant episode from the past, as encoded in memory.

That external data storage acts as an embedding that the attention mechanism can alter, writing to the memory what it learns, and reading from it to make a prediction. While the hidden states of a recurrent neural network are a sequence of embeddings, memory is an accumulation of those embeddings (imagine performing max pooling on all your hidden states – that would be like memory).



- Credit: A. Colyer's Blog*

To quote Hassabis et al:

> While attention is typically thought of as an orienting mechanism for perception, its "spotlight" can also be focused internally, toward the contents of memory. This idea, a recent focus in neuroscience studies (Summerfield et al., 2006), has also inspired work in AI. In some architectures, attentional mechanisms have been used to select information to be read out from the internal memory of the network. This has helped provide recent successes in machine translation (Bahdanau et al., 2014) and led to important advances on memory and reasoning tasks (Graves et al., 2016). These architectures offer a novel implementation of content-addressable retrieval, which was itself a concept originally introduced to AI from neuroscience (Hopfield, 1982).

Further Resources

- Transformer: A Novel Neural Network Architecture for Language Understanding
- Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing
- VIDEO: Successes and Challenges in Neural Models for Speech and Language - Michael Collins
- VIDEO: Comparing Attention with convolutional networks and recurrent nets
- Neuroscience-Inspired Artificial Intelligence, by Demis Hassabis et al

> ‘ Up until quite lately, most CNN models worked directly on entire images or video frames, with equal priority given to all image pixels at the earliest stage of processing. The primate visual system works differently. Rather than processing all input in parallel, visual attention shifts strategically among locations and objects, centering processing resources and representational coordinates on a series of regions in turn (Koch and Ullman, 1985, Moore and Zirnsak, 2017, Posner and Petersen, 1990). Detailed neurocomputational models have shown how this piecemeal approach benefits behavior, by prioritizing and isolating the information that is relevant at any given moment (Olshausen et al., 1993, Salinas and Abbott, 1997). As such, attentional mechanisms have been a source of inspiration for AI architectures that take “glimpses” of the input image at each step, update internal state representations, and then select the next location to sample (Larochelle and Hinton, 2010, Mnih et al., 2014) (Figure 1A). One such network was able to use this selective attentional mechanism to ignore irrelevant objects in a scene, allowing it to perform well in challenging object classification tasks in the presence of clutter (Mnih et al., 2014). Further, the attentional mechanism allowed the computational cost (e.g., number of network parameters) to scale favorably with the size of the input image. Extensions of this approach were subsequently shown to produce impressive performance at difficult multi-object recognition tasks, outperforming conventional CNNs that process the entirety of the image, both in terms of accuracy and computational efficiency (Ba et al., 2015), as well as enhancing image-to-caption generation (Xu et al., 2015).

- Hybrid computing using a neural network with dynamic external memory, by Graves et al

1) *No puedo caminar por los arrabales en la soledad de la noche, sin pensar que ésta nos agrada porque suprime los ociosos detalles, como el recuerdo.*

Interactive Demo

Learn to build AI applications using our interactive learning portal.

TRY IT NOW

Company
About
Press Kit
Contact Us
Press
Privacy

Platform
SKIL
Subscriptions
Documentation
Community Support

International
English
Japanese

Follow Us
Facebook
Twitter
Linkedin
Gitter

Subscribe to our mailing list

Keep me updated!