Attention as a Leverage for Deep Learning

Erik Perillo
Advisor: Profa. Dra. Esther Colombini

May 30, 2019

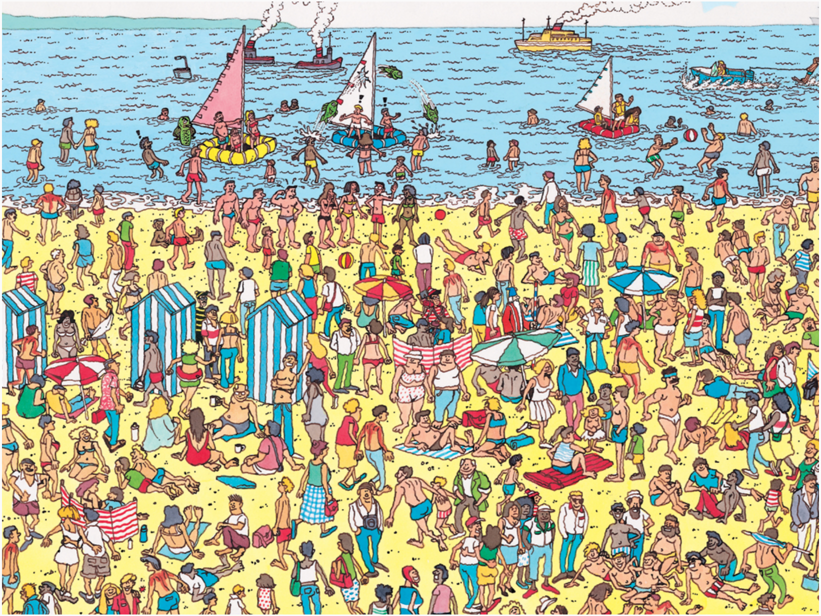
Institute of Computing - Unicamp - Brazil

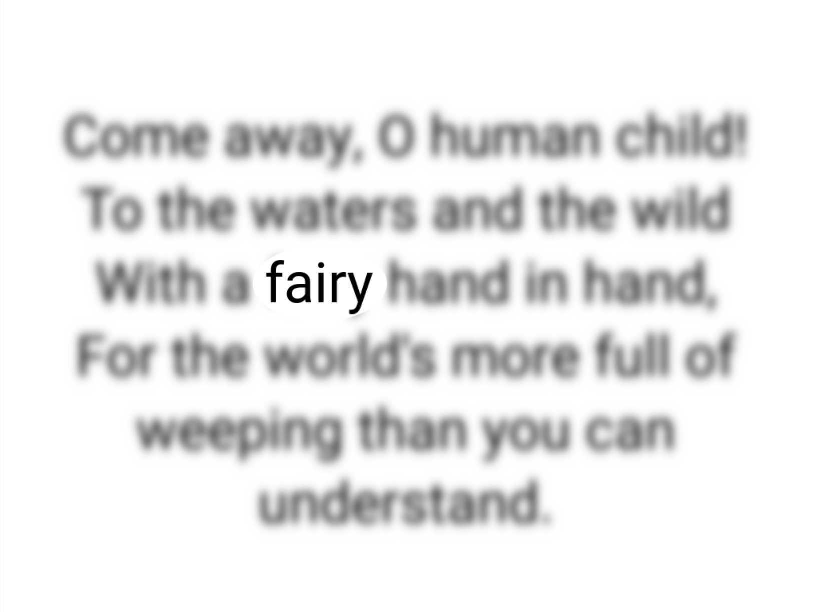
Outline

- Introduction
- Background
- Methodology
- Work so far

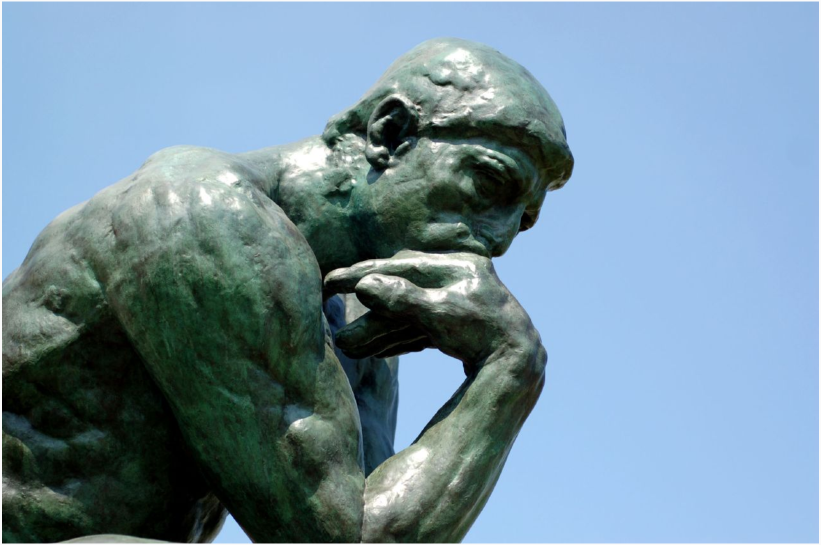
Introduction

Introduction




Come away, O human child!
To the waters and the wild
With a fairy hand in hand,
For the world's more full of
weeping than you can
understand.

Introduction

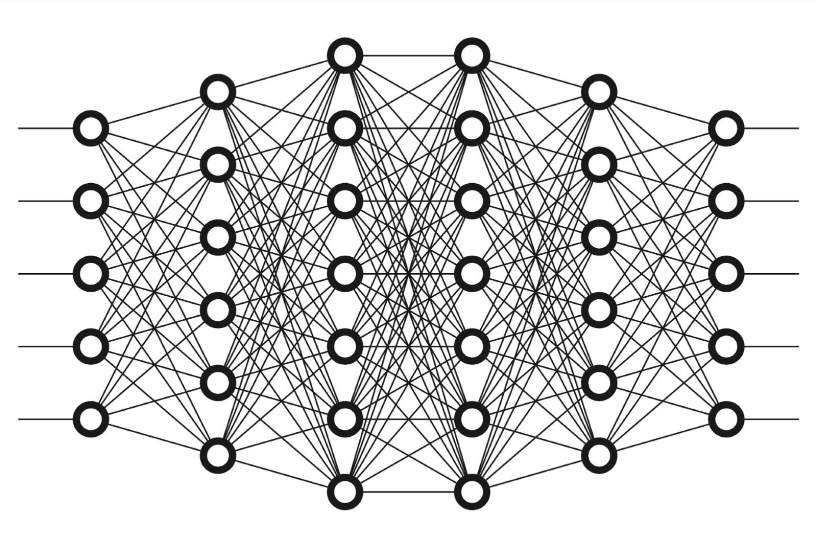


Attention: the ability to **filter and select** *relevant stimuli, to* **keep focus** *on a task for an adequate amount of time. To appropriately* **direct mental resources**

Attention for intelligence and AI

- Fundamental for intelligence
- Fundamental for AI

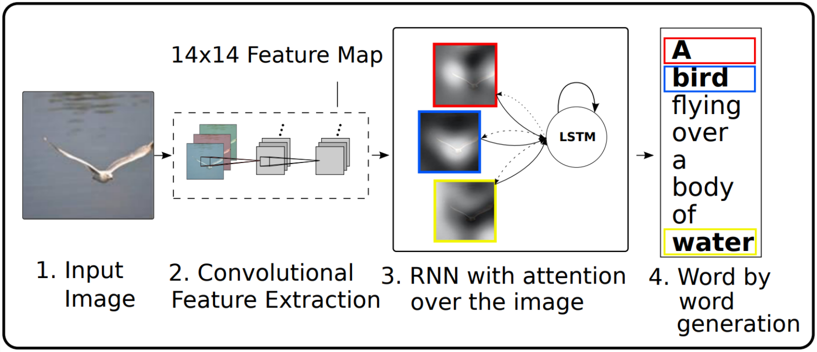
The rise of Deep Learning



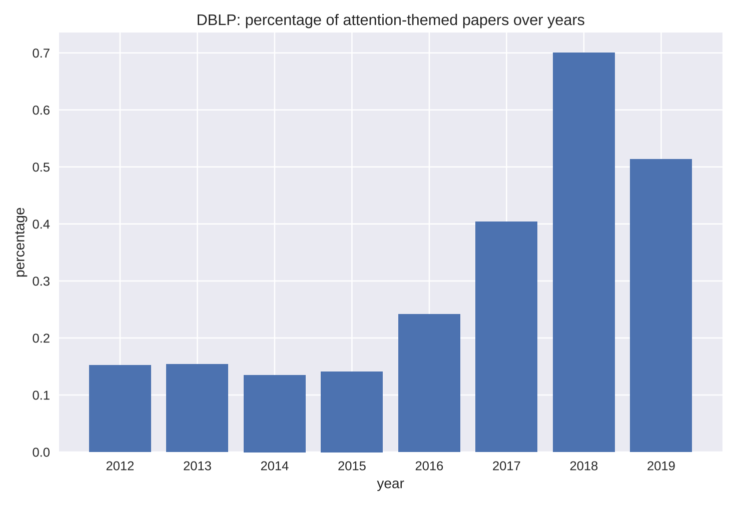
Deep Learning and Attention

- Increasingly more common!
- Constantly sets a new SOTA for the tasks attacked

Deep Learning and Attention: Image Captioning



Deep Learning and Attention: rise in interest



[1] source: DBLP (https://dblp.uni-trier.de/)

Motivation

- Many tasks are approached with Deep Learning yet still do not use Attention
- There are aspects of Attention still to be explored
- We believe it's possible to further generalize Attention for the benefit of Deep Learning

The main contribution

To establish a framework for applicability of Attention to Deep Learning to help guide future development in the area

Objectives

- To perform an extensive **literature review** on the use of Attention in modern Deep Learning
- To identify **general elements of Attention** to be applied to Deep Learning
- To identify **specific problems** in different classes (robotics, vision, NLP...) with improvement potential through the use of Attention;
- To **propose and implement** one or more solutions based on the findings of the work to validate the ideas and evaluate them in an application

Background

Main concepts of Attention: Functionalities

- To **select stimuli** that is relevant
- To **sustain focus** on a specific semantic element for a certain period
- To **guide processing** in a sequential manner that is relevant for a task
- To **orient resources** to new important stimuli

Main concepts of Attention: Bottom-up vs Top-down

- **Bottom-up** Attention: involuntarily started and guided by external and conspicuous stimuli
- **Top-down** Attention: cognition and goals voluntarily guide the focus

Main concepts of Attention: Soft vs Hard

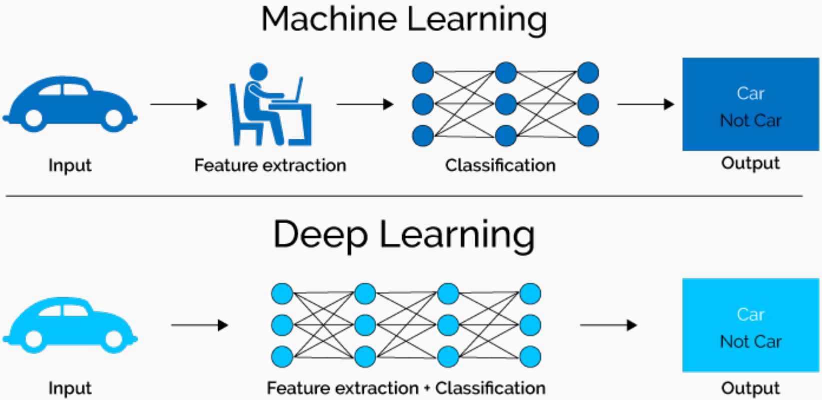
- **Hard** Attention: **choice** of items in a possibly non-deterministic manner

$$z = choice\left(\{x_1, x_2, \ldots, x_n\}\right)$$

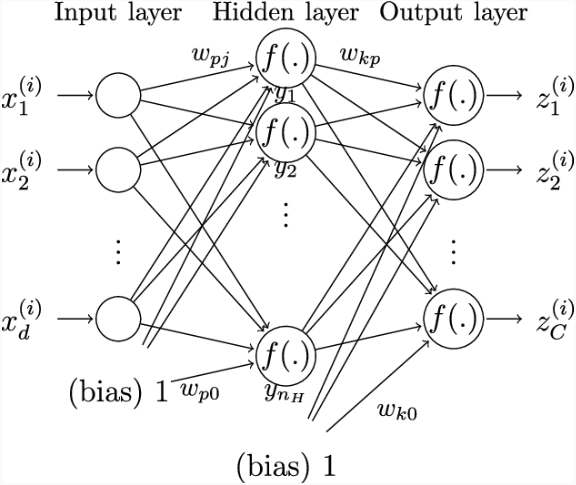
- **Soft** Attention: **weighting** of items in a deterministic manner

$$z = \sum_{i=1}^{n} x_i \alpha_i, \quad 0 \leq \sum_{i=1}^{n} \alpha_i \leq 1$$

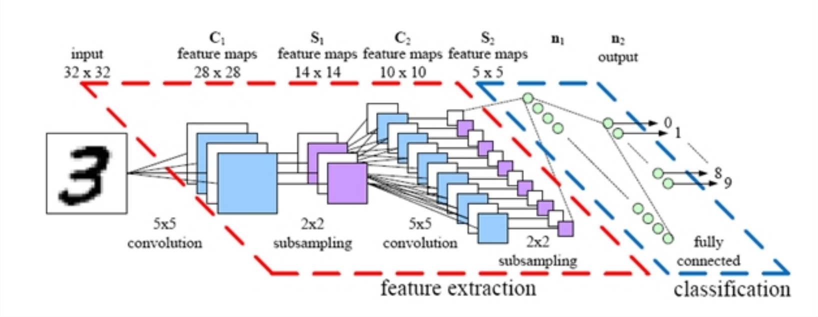
Deep Learning



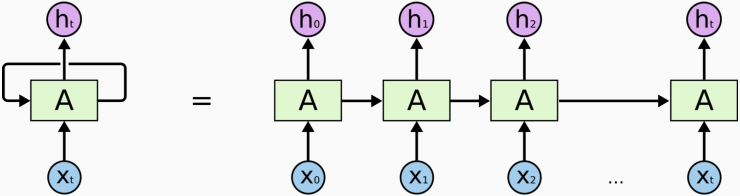
Deep Learning: Deep MLPs



Deep Learning: ConvNets (CNNs)



Deep Learning: Recurrent Neural Networks (RNNs)

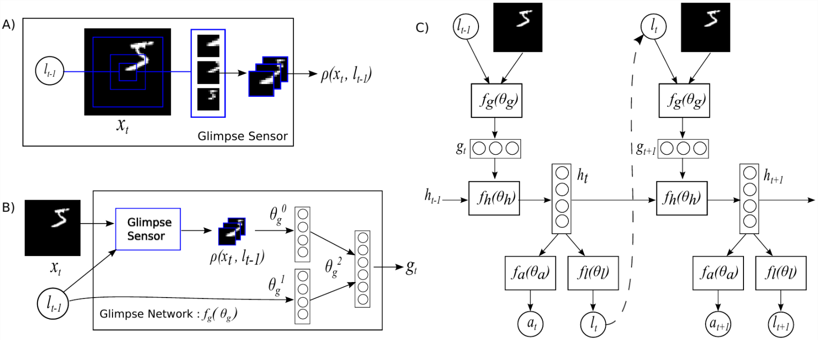


Deep Learning: Learning Process

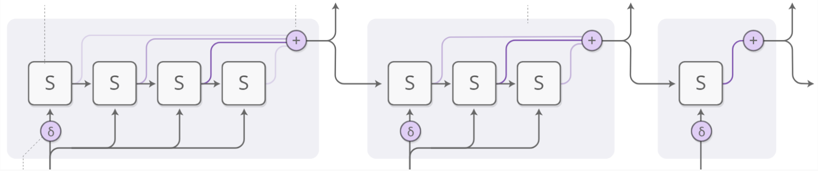
- The act of learning the appropriate weights of a given model
- Usually via *supervised learning*
- Usually obtained by the minimization of a differentiable loss function $L(y, \hat{y})$, the error between y and $\hat{y}$
- Backpropagation plays an essential role in Deep Learning:
    - forward-propagation step, which calculates the loss
    - backpropagation step which adjusts the weights:

$$\theta_{i+1} = \theta_i - \alpha \frac{\partial J}{\partial \theta}$$

Related work: Recurrent Attention Model (RAM)



Related work: Adaptive Computation Time (ACT)



Methodology

Activities

1. **A1**: Literature Review
    - **A1.1**: Theoretical framework for Attention
    - **A1.2**: Elaboration of survey
    - **A1.2**: Survey article writing
2. **A2**: Proposal of an Attention framework for Deep Learning
    - **A2.1**: Establishment of Attention components for specific Deep Learning domains
3. **A3**: Validation of framework
    - **A3.1**: Arrangement of experiments
    - **A3.2**: Execution of experiments
    - **A3.3**: Evaluation of experimental results
    - **A3.4**: Experiments article writing
4. **A0**: Masters activities
    - **A0.1**: Course's requirement fulfillment
    - **A0.2**: Qualification Exam
    - **A0.3**: Masters dissertation
    - **A0.4**: Defense of masters dissertation

Schedule

Table 1: Project schedule.

Activity	**2019**											
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
A1.2	*	*	*	*	*	*	*					
A0.2					*							
A1.3							*					
A2.1							*					
A3.1								*				
A3.2								*	*	*	*	
A3.3											*	
A3.4											*	
A0.3										*	*	*
A0.4												*

Work so far

Theoretical Framework for Attention

Two main parts:

1. A **definition** of Attention (*what* is Attention?)
2. A **model** of Attention (*how* does attention emerge?)

Theoretical Framework for Attention: why?

- We need a *precisely defined* basis to be work upon for:
    - The analysis of papers in the literature review
    - The solutions and models proposed in the future
    - ...

A definition of “Attention”

- **Goal**: define a set of *entities of interest* and the phenomenon of Attention in terms of its *functionalities* and how it relates to the entities.
- **Why this goal?**
    - There are multiple (conflicting) definitions of what “Attention” is
    - We need to postulate a *precise definition* in which all of our work will be based upon

A definition of “Attention”: entities

- **Data:** information, stimuli.
- **Program:** algorithm, sequence of computer (or mental) operations.
- **Process:** the execution of a program on a specific data instance.
- **Computer:** the executor of processes, the brain.
- **Resource:** when not specified, we mean computational resources, e.g., CPU time.
- **Time:** the flow of time.
- **World:** the external environment.
- **Agent:** the actor in the world.
- **Actions:** the interaction of the agent with the world.
- **Goals:** the ends, objectives to be met.

A definition of “Attention”

Data, **programs** and **processes** are virtually **infinite**. Computational **resources** and **actions** are finite.

Attention is **the system for allocating resources to processes**.

In other words, **attention** is the entity in **agent** that, given **context** and a set of **processes**, **allocates resources** to execute each of them in order to **produce outputs** in form of **data** and **actions** in a **correct sequential manner** and in **sensible time** in order to reach **goals**.

A model for Attention

- Proposal: to model Attention as a phenomenon that emerges from the use of **attention modules** in a system

A model for Attention

A model for Attention

At each time step t, the module receives as *input*:

- Current *outer state* $o_t \in O$, where O is the *outer state set*
- Group of *focus targets* $\tau_t = \{\tau_{t1}, \ldots, \tau_{tk}\}, \tau_{ti} \in T$, where T is the *focus target set*
- Past *inner state* $\iota_{t-1} \in I$, where I is the *inner state set*

The module produces as *output* (as a function of both inputs):

- Current *inner state* $\iota_t \in I$
- Current *focus output* $\alpha_t = \{\alpha_{t1}, \ldots, \alpha_{tk}\}, \alpha_{ti} \in A$, where A is the *focus output set*

A model for Attention: Focus output

- The main element of the module
- Can be used to allocate *finite resources* to a set of candidate targets by giving them an importance score
- Each element α_{tk} is respective to a target element τ_{tk}.
- Target elements ($\tau \in T$) may effectively be *programs* (tasks) or *data*.

A model for Attention: Soft and Hard Attention

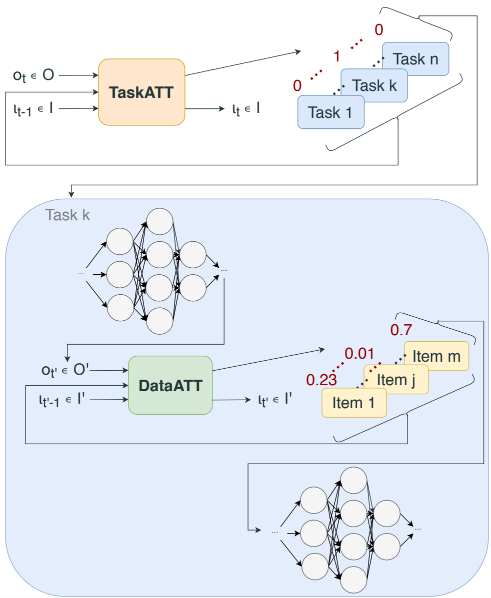
- **Soft Attention:** $A = [0, 1]$, with $0 \leq \sum_{i=1}^{k} \alpha_{ti} \leq 1$
- **Hard Attention:** $A = \{0, 1\}$, with $0 \leq \sum_{i=1}^{k} \alpha_{ti} \leq M$ and $0 \leq M \leq |\tau_t|$

A model for Attention: Bottom-up and Top-down Attention

Depends on the *location* of the module:

- **Bottom-up**: module connected to external stimulus features (e.g. images)
- **Top-down**: module connected to internal/context information

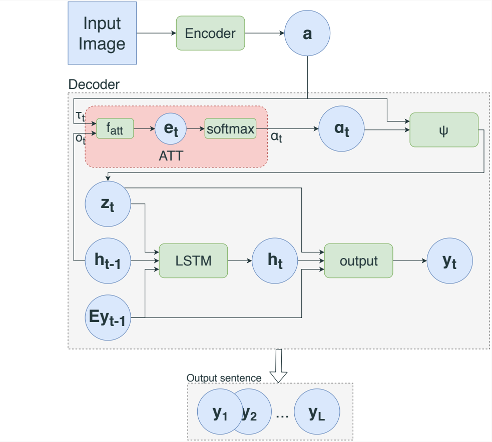
A model for Attention: Example



Validating the model for Attention: Image Captioning

- Work is among the first to propose using attention to image caption generation
- Encoding of the input image is represented as a set of vectors - each respective to a certain spatial region of the image -
- The attentional component gives weights to each vector at each step to produce another vector to be used in further computations

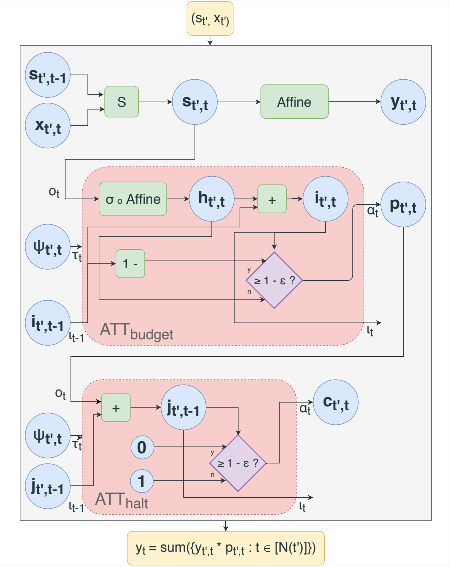
Validating the model for Attention: Image Captioning



Validating the model for Attention: Adaptive Computation Time

- Work proposes a RNN with dynamically variable number of computation steps
- Uses attention to allocate processing "budget" and selection of data

Validating the model for Attention: Adaptive Computation Time



Validating the model for Attention: Adaptive Computation Time

Proposed model can be thought of as having two attention modules:

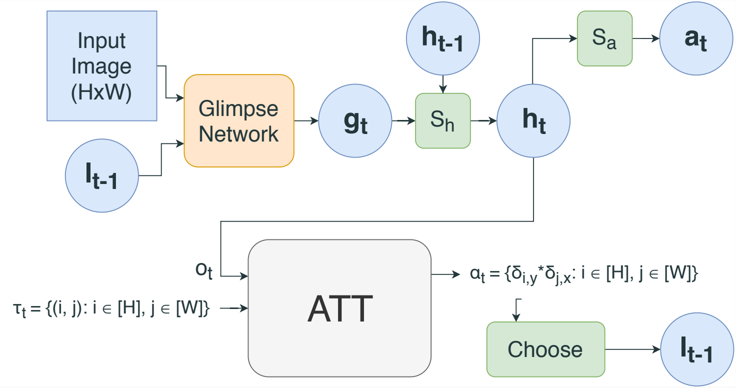
- ATT_{budget}:
    - Computes the value $0 \leq p_{t',t} \leq 1$ to be spent at a given sub-step
    - *Focus output* $p_{t',t}$, represents values to be consumed from the budget and an importance weight for the final output y_t.
- ATT_{halt}:
    - Computes the *continue* value $c_{t',t} \in \{0, 1\}$

The emergent effect: the model can *allocate resources to processes* both by *choosing the data* and *amount of computation time to use*

Validating the model for Attention: Recurrent Visual Attention

- The work proposes a general recurrent model that uses visual attention at each step
- Model selects a retina-like representation of a portion of the input image
- An arbitrary action a_t can be executed to possibly alter the environment

Validating the model for Attention: Recurrent Visual Attention



Survey

- **Main goal**: to perform a **broad analysis of recent works** that propose attention-based solutions **under the perspective of our theoretical framework**

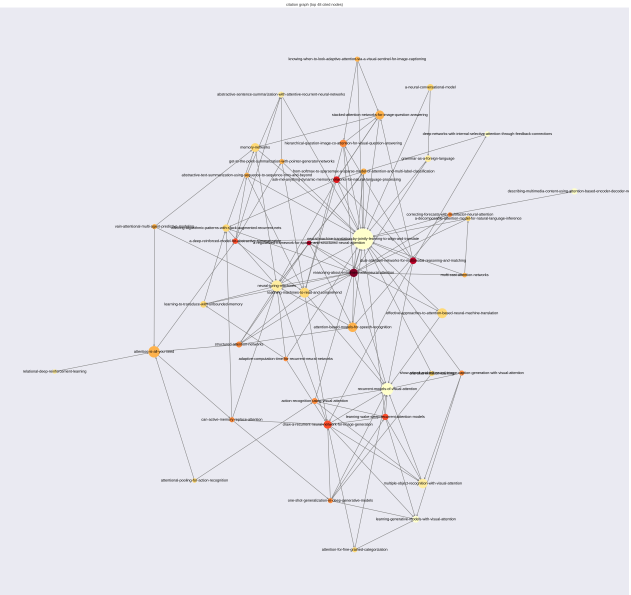
Survey: collection of relevant works

- **Publication date range**: from 2014 to 2019
- **Databases searched**:
  - **arXiv** - `https://arxiv.org/`
  - **DeepMind** - `https://deepmind.com/research/publications/`
  - **Google AI** - `https://ai.google/research/pubs/`
  - **OpenAI** - `https://openai.com/research/#publications`
  - **NIPS** - `https://nips.cc/`
  - **ICML** - `https://icml.cc/`
  - **CVPR** - `http://cvpr2018.thecvf.com/`
  - ...
- **Terms (in title or abstract)**: "attention", "attentive" or "attentional"
- The **relevance** of each work was confirmed upon the reading of the abstract
- As a result, we collected around **300 papers**
- We used *Zotero* and grouped works based on application domain, architectures...

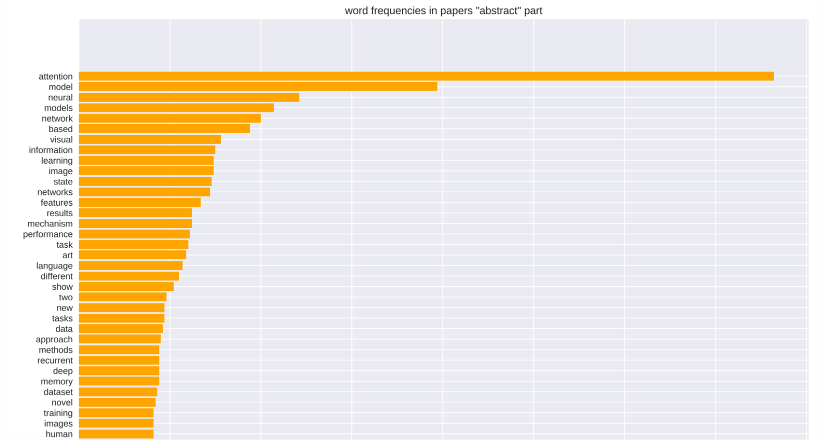
Survey: Visualization of papers data

- Some visualizations were generated for insights
- Analysis include:
  - Citations graph (authors and works)
  - Abstract/title word frequencies
  - Frequency of attention-themed papers over the years

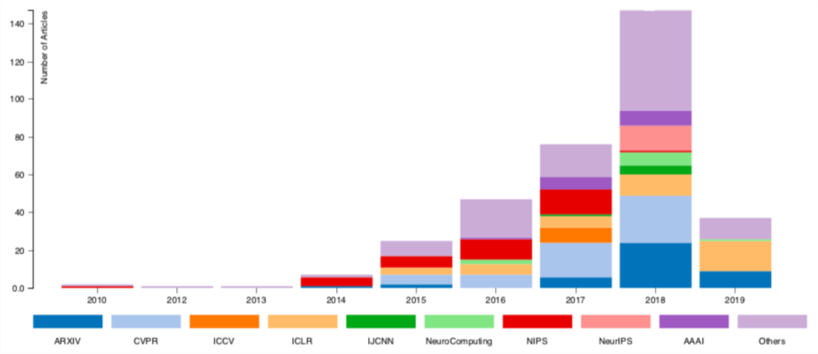
Survey: Data visualization - works citations graph



Survey: Data visualization - frequency of terms in abstract



Survey: Data visualization - frequency of papers over years



Survey: paper relevance analysis

- Goal: to assess the relevance and problem domain of each work
- To each work, we attributed the citation count, domains and an impact score ranging from 1 to 5
- Score was assessed in a quick and rough manner via the abstract of each work:
    - How innovative is the proposed model(s) of the work?
    - How general is the proposed model(s)?
    - Does the proposed model(s) archives/surpasses state-of-the-art in some task?
    - Is attention a central component to the results of the work?

Survey: Reading and summarization of works

- Goal: Obtain a summarization and deep analysis for each paper in the collection (in order of relevance, from highest to lowest)
- A summary template was formulated and summaries were generated for some works.
- The main and longest step of the survey
- We may further refine our theoretical framework and to guide the reading of future papers as we read those papers
- Survey has shown so far that the use of attention in Deep Learning has indeed provided improvements in basically all subfields of Deep Learning.

Next Steps

- Survey:
    - Finish papers analysis
    - Refine theoretical framework
    - Write and publish survey paper
- With the framework and findings of the survey, choose a problem domain and task to attack with an attention-based model. Probably a robotics problem using Reinforcement Learning

Thank you

Questions