Article Classification

Nir Tal

DEEP LEARNING with Python

SECOND EDITION

François Chollet

MANNING

Introduction

Frontiers for Young Minds

- Online & open-access science journal for kids (8 to 15 years old)
- Kid-friendly articles on core concepts and recent discoveries
- Articles written by academic scientists and reviewed by kids and their science mentors

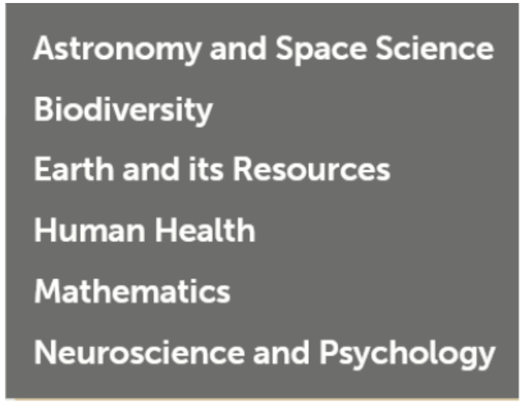
Science for kids, edited by kids

The Problem

Article Classification (Topic Analysis)

- Articles already organized in 6 sections:

frontiers FOR YOUNG MINDS



Astronomy and Space Science

Biodiversity

Earth and its Resources

Human Health

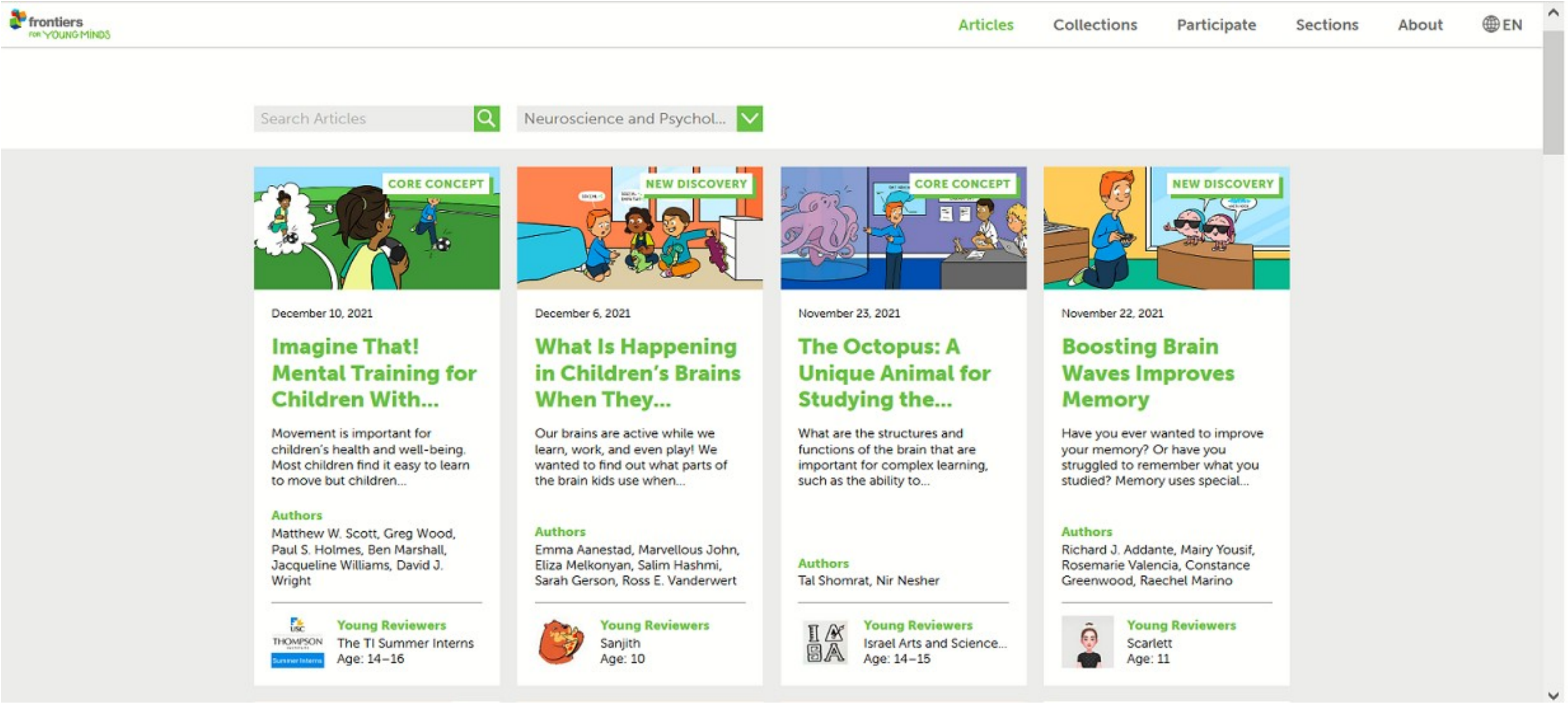
Mathematics

Neuroscience and Psychology

- How to classify new articles into sections?

The Data

Section Webpage



December 10, 2021

Imagine That! Mental Training for Children With...

Movement is important for children's health and well-being. Most children find it easy to learn to move but children...

Authors
Matthew W. Scott, Greg Wood, Paul S. Holmes, Ben Marshall, Jacqueline Williams, David J. Wright

Young Reviewers
The TI Summer Interns
Age: 14–16

December 6, 2021

What Is Happening in Children's Brains When They...

Our brains are active while we learn, work, and even play! We wanted to find out what parts of the brain kids use when...

Authors
Emma Aanestad, Marvellous John, Eliza Melkonyan, Salim Hashmi, Sarah Gerson, Ross E. Vanderwert

Young Reviewers
Sanjith
Age: 10

November 23, 2021

The Octopus: A Unique Animal for Studying the...

What are the structures and functions of the brain that are important for complex learning, such as the ability to...

Authors
Tal Shomrat, Nir Nesher

Young Reviewers
Israel Arts and Science...
Age: 14–15

November 22, 2021

Boosting Brain Waves Improves Memory

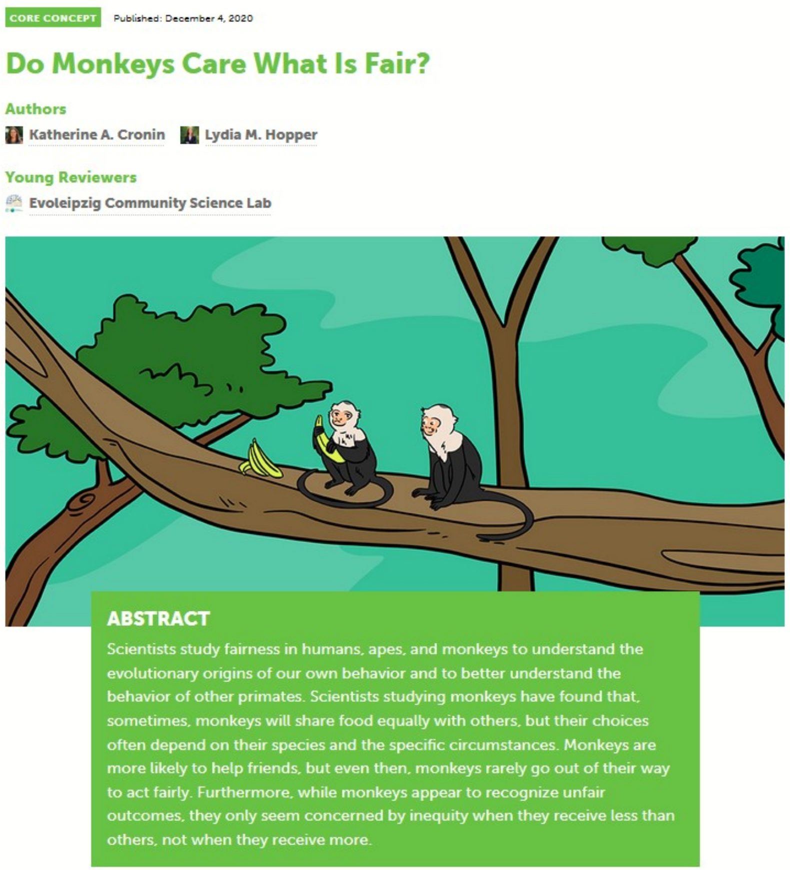
Have you ever wanted to improve your memory? Or have you struggled to remember what you studied? Memory uses special...

Authors
Richard J. Addante, Mairy Yousif, Rosemarie Valencia, Constance Greenwood, Raechel Marino

Young Reviewers
Scarlett
Age: 11

The Data

Article Structure



CORE CONCEPT Published: December 4, 2020

Do Monkeys Care What Is Fair?

Authors

Katherine A. Cronin Lydia M. Hopper

Young Reviewers

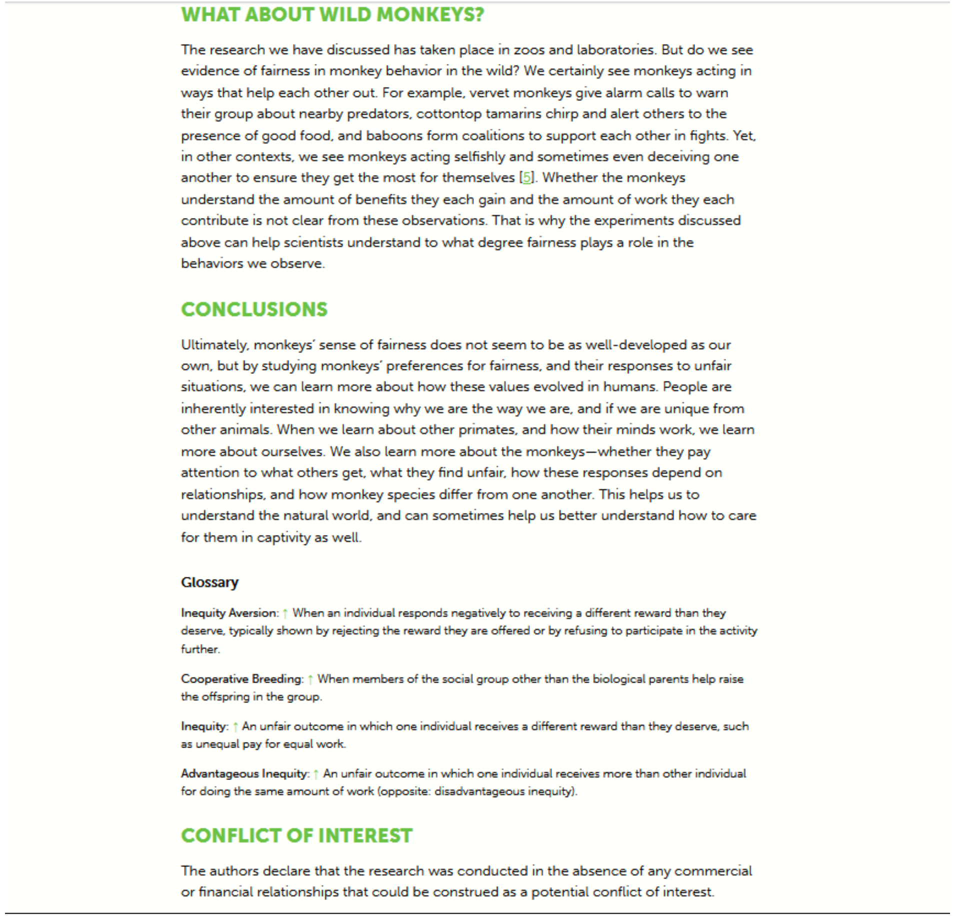
Evoleipzig Community Science Lab

ABSTRACT

Scientists study fairness in humans, apes, and monkeys to understand the evolutionary origins of our own behavior and to better understand the behavior of other primates. Scientists studying monkeys have found that, sometimes, monkeys will share food equally with others, but their choices often depend on their species and the specific circumstances. Monkeys are more likely to help friends, but even then, monkeys rarely go out of their way to act fairly. Furthermore, while monkeys appear to recognize unfair outcomes, they only seem concerned by inequity when they receive less than others, not when they receive more.

The Data

Article Structure - Continued



WHAT ABOUT WILD MONKEYS?

The research we have discussed has taken place in zoos and laboratories. But do we see evidence of fairness in monkey behavior in the wild? We certainly see monkeys acting in ways that help each other out. For example, vervet monkeys give alarm calls to warn their group about nearby predators, cottontop tamarins chirp and alert others to the presence of good food, and baboons form coalitions to support each other in fights. Yet, in other contexts, we see monkeys acting selfishly and sometimes even deceiving one another to ensure they get the most for themselves [5]. Whether the monkeys understand the amount of benefits they each gain and the amount of work they each contribute is not clear from these observations. That is why the experiments discussed above can help scientists understand to what degree fairness plays a role in the behaviors we observe.

CONCLUSIONS

Ultimately, monkeys' sense of fairness does not seem to be as well-developed as our own, but by studying monkeys' preferences for fairness, and their responses to unfair situations, we can learn more about how these values evolved in humans. People are inherently interested in knowing why we are the way we are, and if we are unique from other animals. When we learn about other primates, and how their minds work, we learn more about ourselves. We also learn more about the monkeys—whether they pay attention to what others get, what they find unfair, how these responses depend on relationships, and how monkey species differ from one another. This helps us to understand the natural world, and can sometimes help us better understand how to care for them in captivity as well.

Glossary

Inequity Aversion: ↑ When an individual responds negatively to receiving a different reward than they deserve, typically shown by rejecting the reward they are offered or by refusing to participate in the activity further.

Cooperative Breeding: ↑ When members of the social group other than the biological parents help raise the offspring in the group.

Inequity: ↑ An unfair outcome in which one individual receives a different reward than they deserve, such as unequal pay for equal work.

Advantageous Inequity: ↑ An unfair outcome in which one individual receives more than other individual for doing the same amount of work (opposite: disadvantageous inequity).

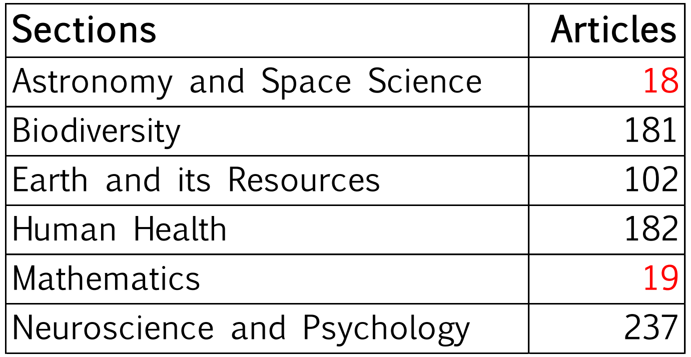
CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The Data

Web Scraping

- Retrieving article links from each section
- Opening links and finding html elements: title, abstract, paragraphs
- Extracting text and writing to txt files: one paragraph per line



Sections	Articles
Astronomy and Space Science	18
Biodiversity	181
Earth and its Resources	102
Human Health	182
Mathematics	19
Neuroscience and Psychology	237

The Data

Text Standardization

- Parsing paragraphs to sentences
- Splitting sentences to words (tokenization)
- Removing stop words (NLTK)
- Recombining tokens to a single string
- Writing strings to txt files: one file per article

The Data

Text Dataset Preparation

- Data splitting into 3 directories: train, validation, and test
- Eeach directory contains 4 sub-directories, i.e., 4 classes:
  - Biodiversity
  - Earth-and-its-resources
  - Human-health
  - Neuroscience-and-psychology
- TensorFlow/Keras text_dataset_from_directory():
  - Samples are string tensors
  - Targets are float tensors, multi-hot encoding 4 classes
  - For example, the Biodiversity class is encoded by [1, 0, 0, 0]
  - Batch size is 32
  - Train/Val/Test dataset samples: 489/139/74 (70%/20%/10%)

The Models

Bag-of-Words Models: *Unigram* Model

- Dataset processing with Keras TextVectorization layer:
  - Single words (unigrams) with multi-hot encoding
  - Max tokens are 20,000
  - Building the vocabulary using the *Train* dataset
- Building the model:
  - Inputs layer: 20,000 dimensions
  - Dense layer: 24 dimensions, ReLU activation
  - Outputs (dense) layer: 4 dimensions, SoftMax activation
  - Loss: categorical crossentropy
  - Optimizer: rmsprop
  - Metrics: accuracy

The Models

Bag-of-Words Models: *Unigram* Model

- Training the model (4 epochs):

```
Train_loss: 0.0153
Train_accuracy: 1.0000
Val_loss: 0.4315
Val_accuracy: 0.8417
```

- Testing the model:

```
Test_loss: 0.3449
Test_accuracy: 0.8243
```

The Models

Bag-of-Words Models: *Bigram* Model

- Dataset processing with Keras TextVectorization layer:
  - Bigrams with multi-hot encoding
  - Max tokens are 20,000
  - Building the vocabulary using the *Train* dataset
- Building the model:
  - Inputs layer: 20,000 dimensions
  - Dense layer: 24 dimensions, ReLU activation
  - Outputs (dense) layer: 4 dimensions, SoftMax activation
  - Loss: categorical crossentropy
  - Optimizer: rmsprop
  - Metrics: accuracy

The Models

Bag-of-Words Models: *Bigram* Model

- Training the model (4 epochs):

  Train_loss: 0.0093

  Train_accuracy: 1.0000

  Val_loss: 0.3859

  Val_accuracy: 0.8633

- Testing the model:

  Test_loss: 0.3359

  Test_accuracy: 0.8514

The Models

Bag-of-Words Models: *TF-IDF* Model

› Dataset processing with Keras TextVectorization layer:

- Bigrams with TF-IDF encoding
- Max tokens are 20,000
- Building the vocabulary using the **Train** dataset

› Building the model:

- Inputs layer: 20,000 dimensions
- Dense layer: 24 dimensions, ReLU activation
- Outputs (dense) layer: 4 dimensions, SoftMax activation
- Loss: categorical crossentropy
- Optimizer: rmsprop
- Metrics: accuracy

The Models

Bag-of-Words Models: *TF-IDF* Model

- Training the model (4 epochs):

  Train_loss: 0.0182

  Train_accuracy: 1.0000

  Val_loss: 0.5636

  Val_accuracy: 0.7842

- Testing the model:

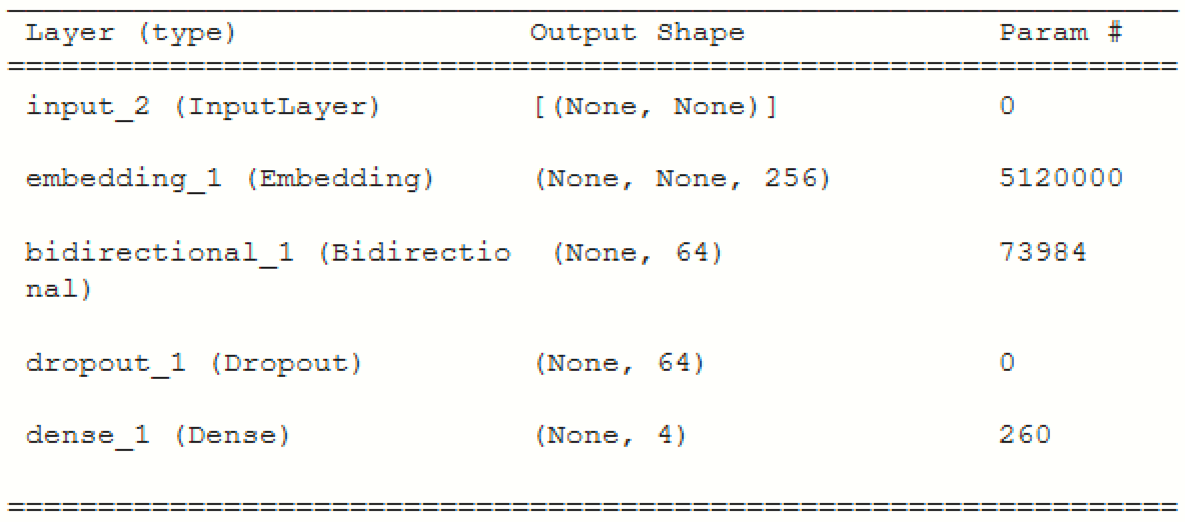
  Test_loss: 0.5021

  Test_accuracy: 0.8510

The Models

Sequence Models: Bidirectional LSTM Model #1

- Dataset processing with Keras TextVectorization layer:
  - Words with **integer indexing:** each sample (article) is represented as a sequence of integer indices
  - Max tokens are 20,000
  - Max article length is 1,000 tokens
- Building the model: learning word embeddings



```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_2 (InputLayer)        [(None, None)]            0

 embedding_1 (Embedding)     (None, None, 256)         5120000

 bidirectional_1 (Bidirectio  (None, 64)               73984
 nal)

 dropout_1 (Dropout)         (None, 64)                0

 dense_1 (Dense)             (None, 4)                 260

=================================================================
```

The Models

Sequence Models: Bidirectional LSTM Model #1

- Training the model (8 epochs):

  Train_loss: 0. 3876

  Train_accuracy: 0.9530

  Val_loss: 0.9109

  Val_accuracy: 0.6835

- Testing the model:

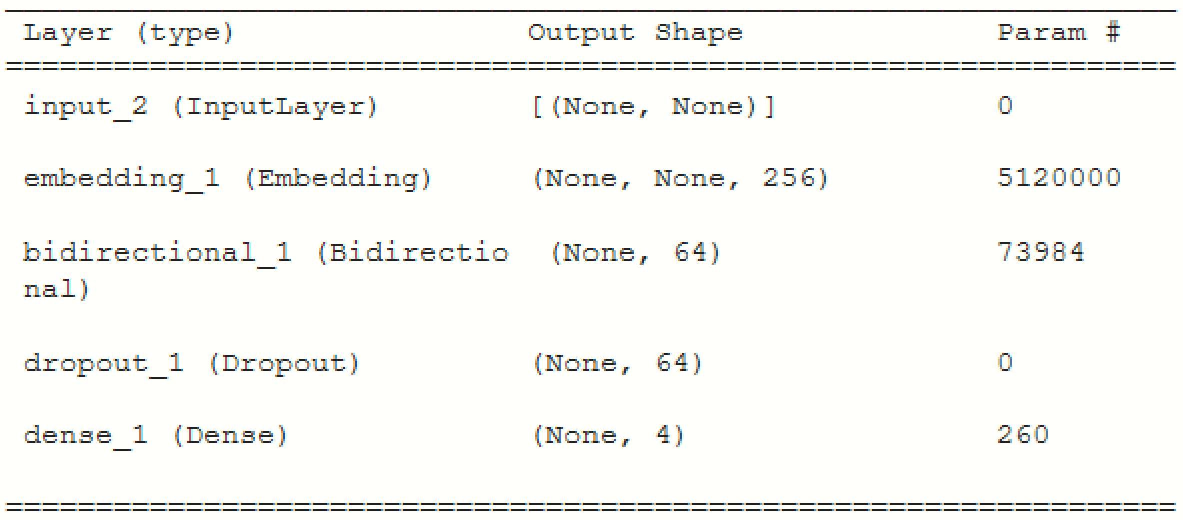
  Test_loss: 0.8738

  Test_accuracy: 0.6892

The Models

Sequence Models: Bidirectional LSTM Model #2

- Dataset processing with Keras TextVectorization layer:
  - Words with **integer indexing:** each sample (article) is represented as a sequence of integer indices
  - Max tokens are 20,000
  - Max article length is 1,000 tokens
- Building the model: learning word embeddings & *masked zeros*



```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_2 (InputLayer)        [(None, None)]            0

 embedding_1 (Embedding)     (None, None, 256)         5120000

 bidirectional_1 (Bidirectio  (None, 64)               73984
 nal)

 dropout_1 (Dropout)         (None, 64)                0

 dense_1 (Dense)             (None, 4)                 260

=================================================================
```

The Models

Sequence Models: Bidirectional LSTM Model #2

- Training the model (8 epochs):

  Train_loss: 0.1972

  Train_accuracy: 0.9775

  Val_loss: 0.7821

  Val_accuracy: 0.7770

- Testing the model:

  Test_loss: 0.6210

  Test_accuracy: 0.7838

Summary

Conclusions

- For this small dataset, bag-of-words models perform better than sequence models, and in particular the *bigram* model
- Because the dataset is small, overfitting is very fast. Hence we needed to reduce the vocabulary size and dense layer dimensions

frontiers

FOR YOUNG MINDS

Thank You

24Slides