## Introduction

- Knowledge is being produced at an unprecedented level*.
- About 3 million scholarly journal articles each year, with an annual growth rate of 5%.
- How to better understanding and organizing the scientific literature?
- According to cognitive and social science, a key management strategy for such information is to organize them into a hierarchy of categories.


## Approach overview

- We follow a concept-representation based approach:
  - We propose to represent the categories as distributions over concepts, which allows for more flexible combinations of the semantics of neighboring nodes in the hierarchy.
  - We propose a novel, adaptive concept-level document representation model based on the hierarchical neural attention mechanism, which models the validity and importance of the concept recognition via a natural hierarchical process.
  - We propose a principled approach for aggregating all possible concept interactions between the documents and each of the possible categories, to comparatively evaluate document-category relevance and perform categorization.

## Concept representation for taxonomy nodes

- Taxonomy nodes are represented by a weight distribution over concepts mined from corpus.
- Hierarchical structure can be encoded by enriching the concept representation of each node with aggregated semantics based from its descendants and path semantics based on all its ancestors.
- Naturally enables assignment to intermediate node and top K category prediction.

## Concept representation for documents

- Concepts can be mined comprehensively using state-of-the-art chunking and text mining approaches.
- The task now becomes, to select concepts that are 1) most plausible among different candidates, and 2) most important to the document's main theme.
- We propose a model hierarchical neural attention mechanism to capture the plausibility attention and importance attention in an end-to-end fashion.

## Assocating concept in semantic embedding space

- Assigning documents to the correct taxonomy nodes based on similarity aggregation over concept representations.
- Mapping from similarities to relevance score is learned end-to-end using dynamic hie-pooling with monotonicity enforcement and gradient path saving.

## Evaluation

- We extensively evaluate our approach for Computer Science + Physics + Math + Medicine with >10 hierarchical categories and a maximum height of 5.
- Our approach significantly outperform the state of baseline methods including Wikipedia2vec, Prem-trained Bert, UNEC, Datasesss.
- Document's relevance to taxonomy nodes can also be visualized as a combination of the individual concepts' relevance weighted by attention.

--

**HierCon: Hierarchical Organization of Technical Documents Based on Concepts**

Keqian Li, Shiyang Li, Semih Yavuz, Hanwen Zha, Yu Su, and Xifeng Yan
University of California, Santa Barbara & Google Brain & The Ohio State University

---