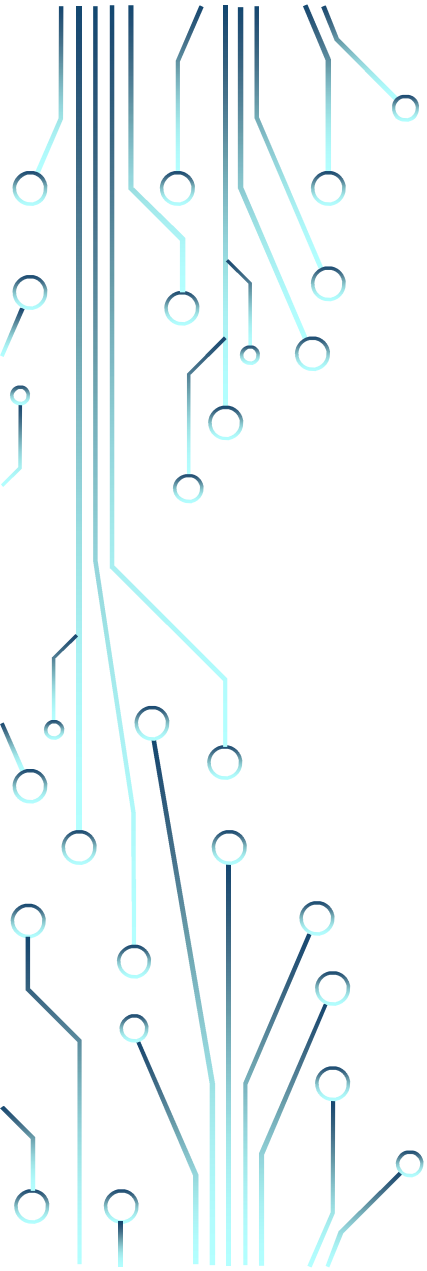


PERSONALIZED SEARCH: CONTRIBUTIONS TO NEURAL APPROACHES AND TO EVALUATION

GABRIELLA PASI

UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

OUTLINE OF THE LECTURE

- Language and Personalization
- Search and Personalization
  - User modeling
  - Search results personalization (query expansion and result re-ranking)
- Some contributions to personalization in neural search settings:
  - User Modeling with Multiple-representation
  - Query-Aware User Modeling with Denoising Attention
  - Personalized LLMs through parameter efficient fine-tuning techniques

The IKR3 Research Lab

Ikr3 Lab — information and knowledge representation, retrieval, and reasoning LAB

Focus on:

Information Retrieval, User Modeling and Personalization, Social Computing, Knowledge Representation and Reasoning, Hybrid AI

Lab members:

- **Head**: Gabriella Pasi; **Associate Professors**: Rafael Penaloza Nyssen and Marco Viviani; **Tenure track researcher**: Alessandro Raganato; **Post Doc**: Sandip Modha, Georgios Peikos, Anima Pramanik; **PhD students**: Renzo Alva Principe, Marco Braga, Pranav Kasela, Gian Carlo Milanese, Effrosyni Sokli, Paolo Tenti

Active Projects:

- *HORIZON-MSCA-2021-DN-01 (Marie Skłodowska-Curie Innovative Training Networks)* **Learning with Multiple Representations (LEMUR)**, *2022-2025.*
- *H2020-MSCA-ITN-2019 (Marie Skłodowska-Curie Innovative Training Networks)* **Domain Specific Systems for Information Extraction and Retrieval (DOSSIER)**, *2019-2023.*
- *PRIN (Progetti di Ricerca di Interesse Nazionale) Project* **PerLIR: Personal Linguistic resources in Information Retrieval**, *2019-2022.*
- *PRIN (Progetti di Ricerca di Interesse Nazionale)* **Project MoT – The Measure of Truth: An Evaluation-Centered Machine-Human Hybrid Framework for Assessing Information Truthfulness**. 2024-2025.

Communication, Information, and Language

In a communication process, the generation of information is potential.
It may be comprised by:

- Imprecision or false statements from the source
- Prejudice of the recipient;
- Accessibility problems, i.e. capability of the recipient to access the content
  - e.g., lack of understanding of the language used by the source to convey the information.

Search Engine

Language and Personalization

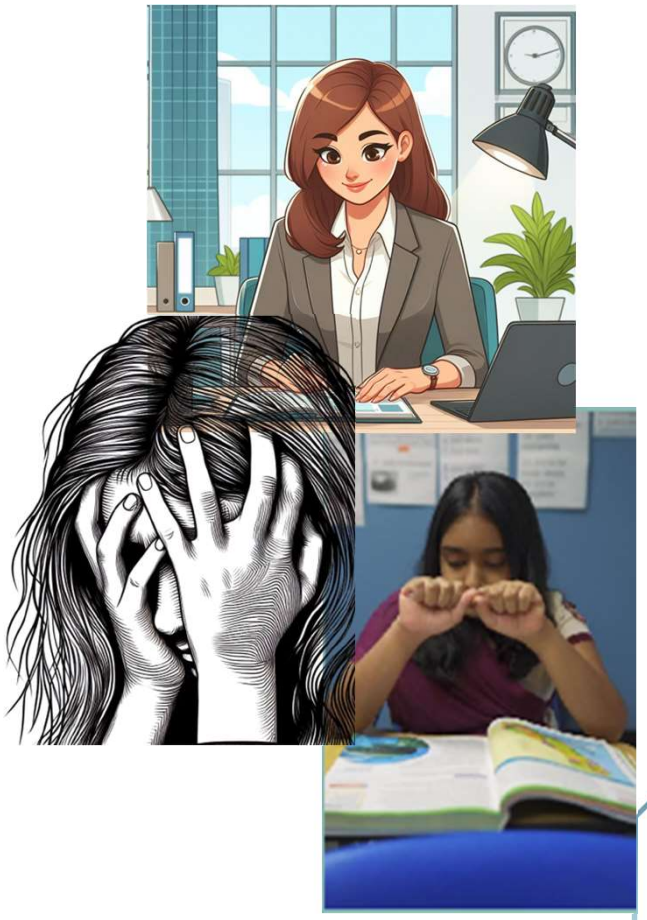
Examples of “personal” (use of) languages:

- professional
- related to specific mental states (e.g. depression)
- biased by sentiment

Example of “group” (use of) languages:

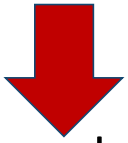
- Thematic groups in social media
- Professional categories of users (medical doctors, lawyers, …)
- Social phenomena (bullyism, harassment)

The lexicon and the structure of the employed language change depending on the context.



Language and Personalization

- Defining a formal (semantic) representation of the language used by either a user (layering) or a group of users



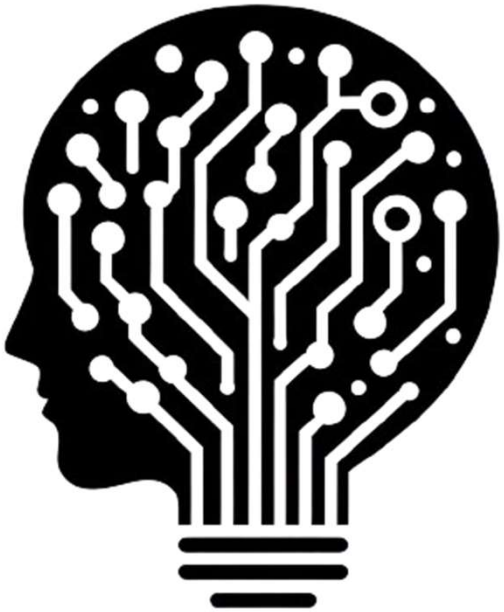
Personal or group language model

- Defining processes to use personal or group linguistic models (e.g. in relation to specific tasks)

The social identity of speakers and listeners is intrinsically linked to the use of language (linguistic variations due to social factors, e.g., age etc.)

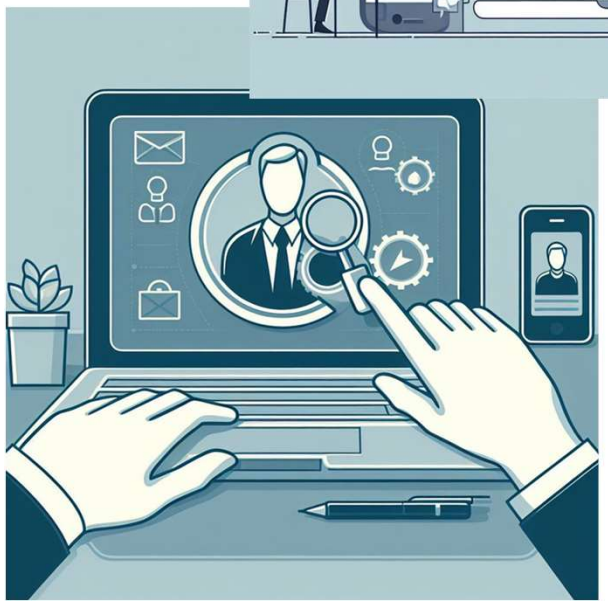
Possible Applications

- Identification of individuals at risk on social media (e.g. depression, or other social phenomena)
- Prevention of social phenomena, for example, social bullying, harassment of various kinds
- Filtering of unsuitable content for minor

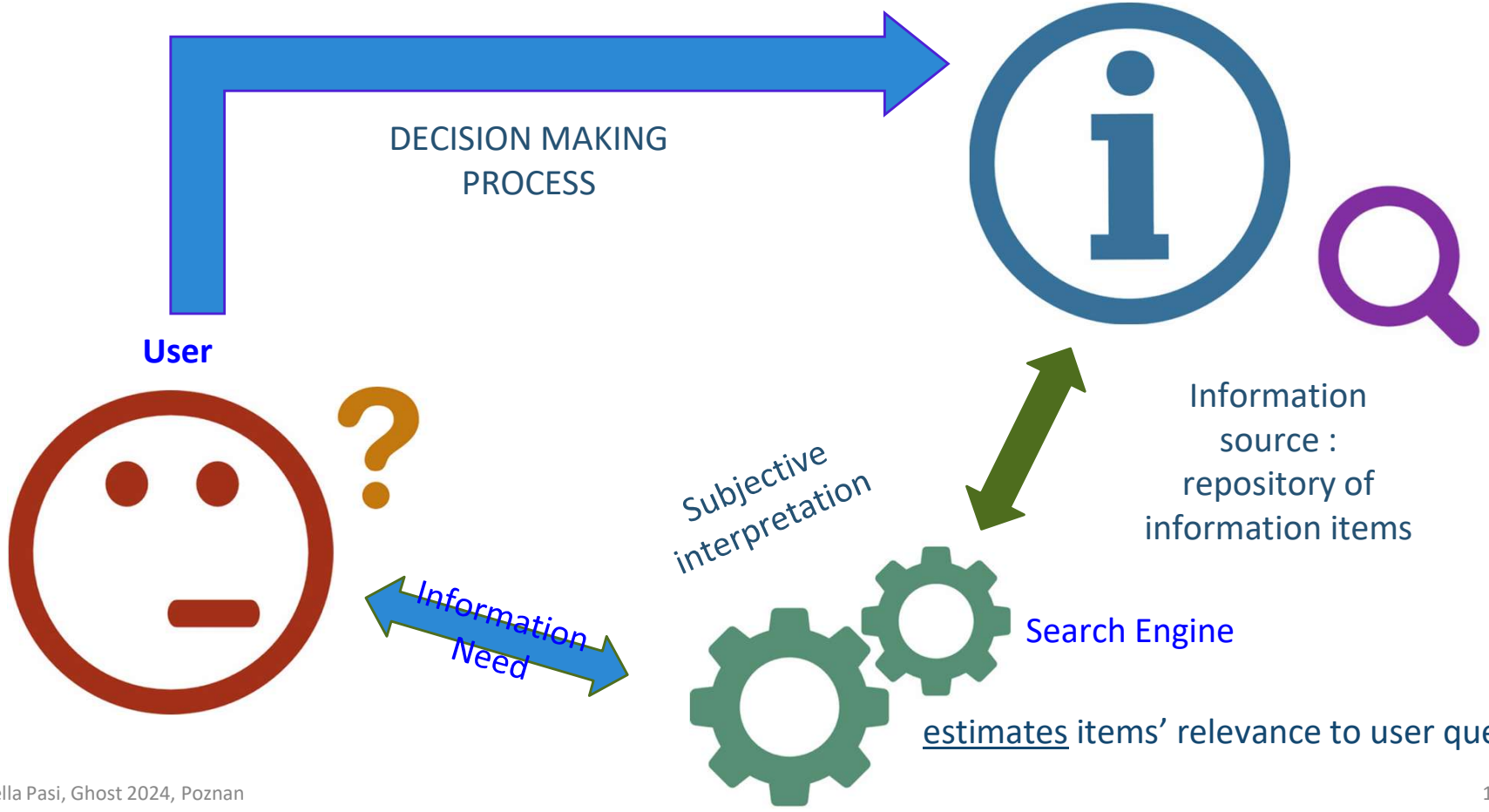


Possible Applications

- Assessment of the veracity of the contents
- Service for professional categories:
  - Search engine - Professional Search
  - Content recommendation
- Job Placement

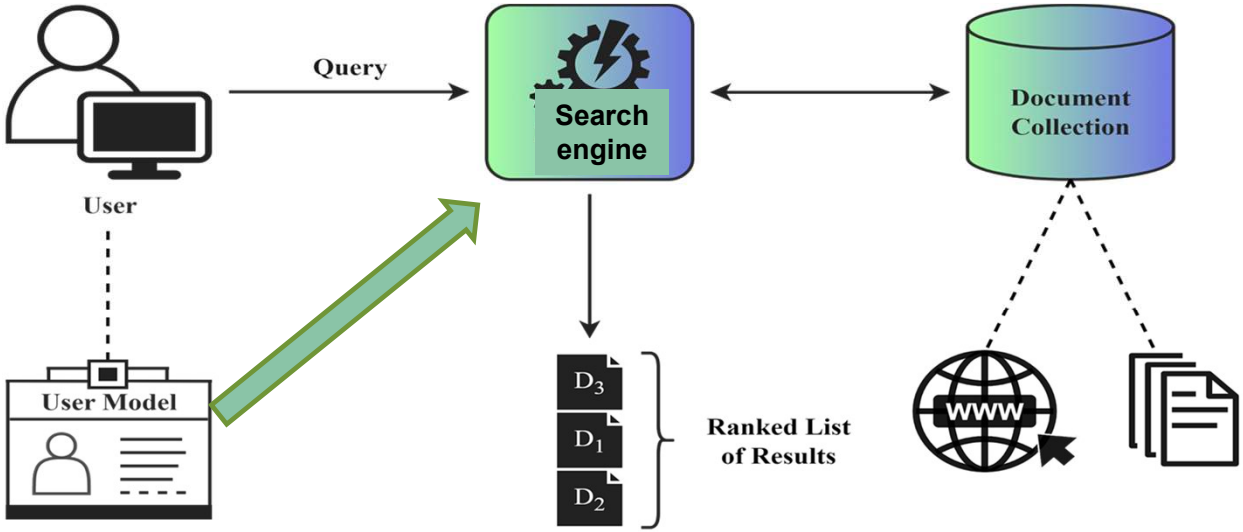


IR SYSTEMS (aka SEARCH ENGINES) Implement a Decision Making Process



Personalized Search

Personalized Search (and contextual search) is one of the main developments of IR, finalised at overcoming the "*one size fits all*" search paradigm, and at providing search results that better suit user's needs. It relies on *user models.*



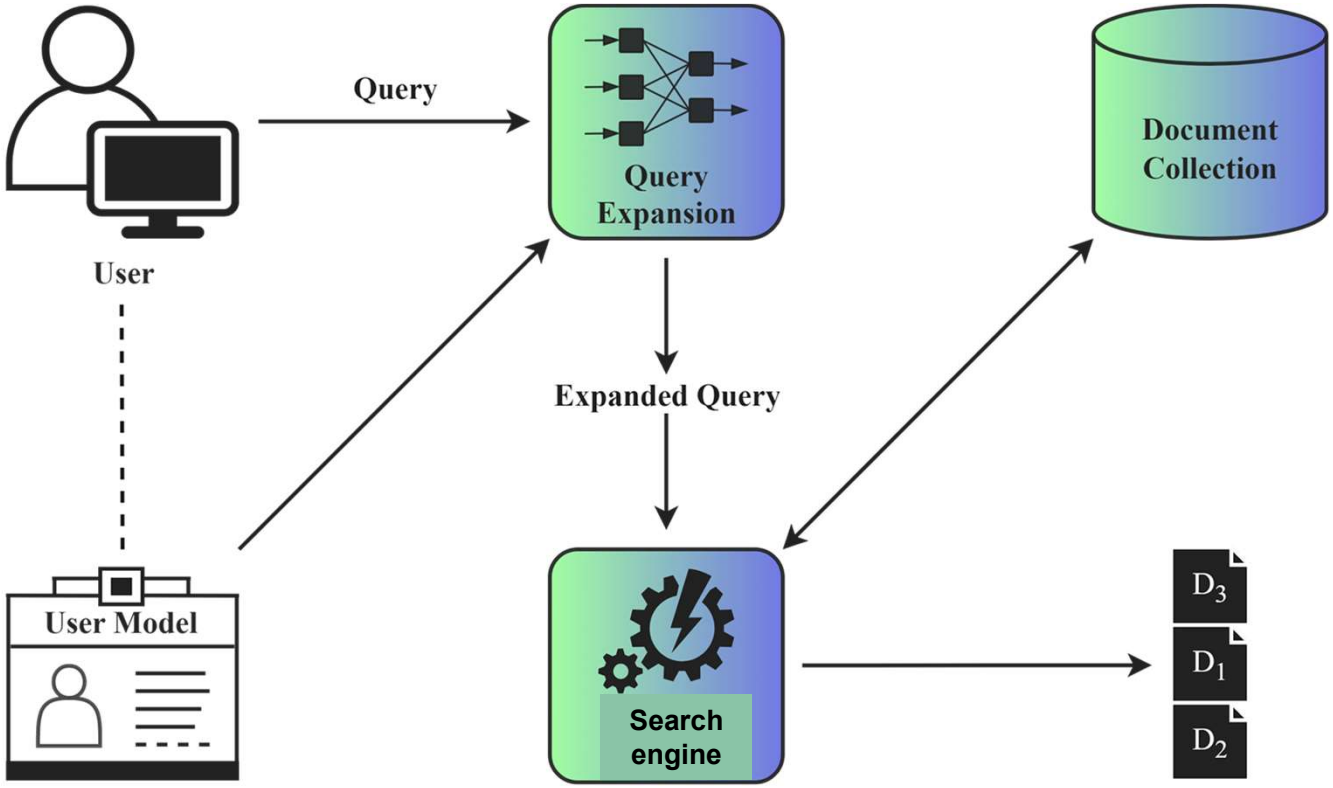
PERSONALIZATION: User Model

1. **User-Related Information Gathering**
   - Explicit
   - Implicit

2. **Representation of the User-Related Information**
   - Long-term user modeling
   - Short-term user modeling

3. **Exploitation of the User-Related Information**
   - Personalized Query Expansion (pre-processing)
   - Personalized Results Re-Ranking (post-processing)

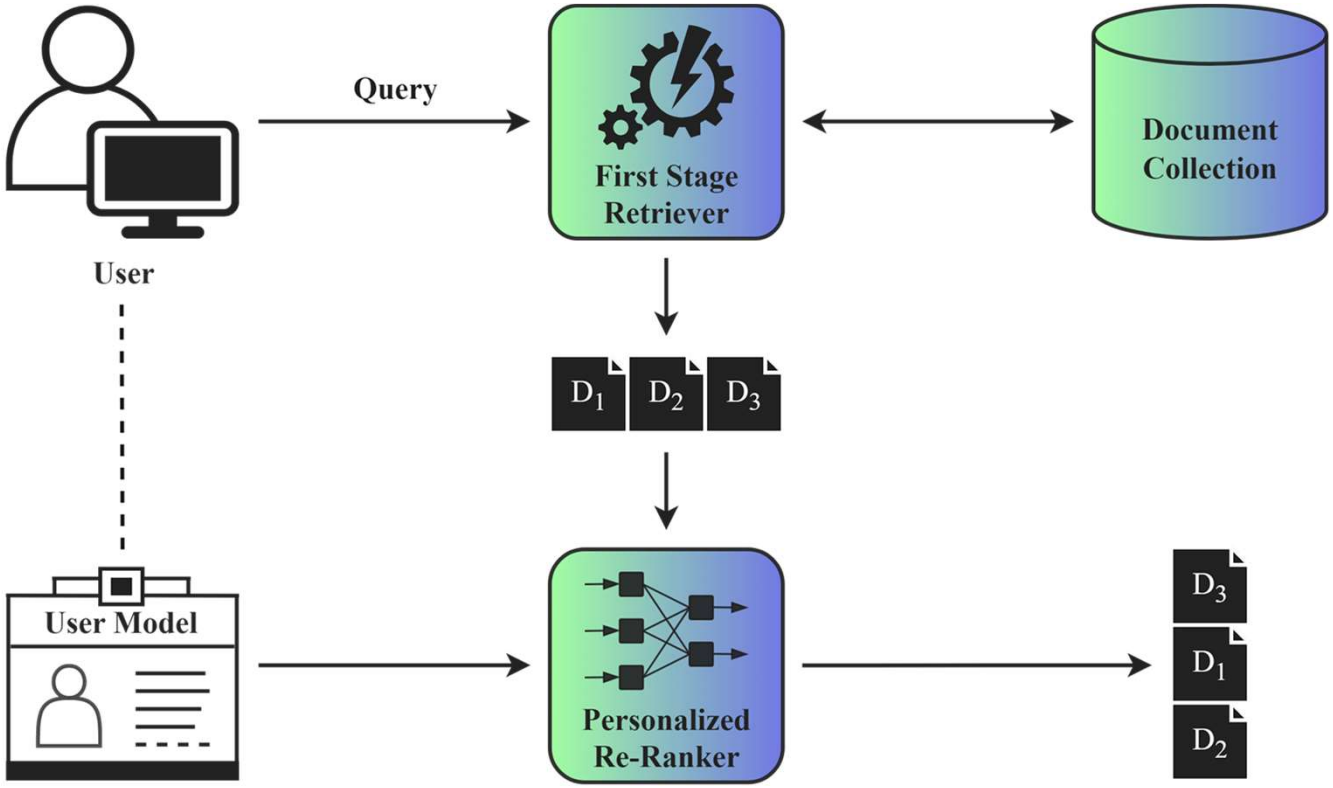
PERSONALIZATION: Query Expansion



PERSONALIZATION: Query Expansion (Example)

1. **Original query :** 70s jazz
2. **User interests :** japanese music
3. **Expanded query :** 70s jazz **japan**

PERSONALIZATION: Result Re-ranking



PERSONALIZATION: Result Re-ranking (Example)

1. **Query :** jazz music
2. **First Stage Retriever results :** 1  2  3 
3. **User interests :** Black Sabbath, piano
4. **Personalized Re-Ranking :** 1  2  3 

Recent Contributions

- How to **represent in a user model** and **leverage** in search the user preferences inferred from **heterogeneous information** sources.
  Multi-Representation User Model based on heterogeneous information sources

- How to **exploit** the **user interests** related to the current query and how to decide if personalize query processing?
  Query-Aware User Model based on a novel Neural Attention variant, the Denoising Attention

- How to **personalize Large Language Models** with user-specific information?
  Personalized Large Language Model through Parameter Efficient Fine-Tuning Technique

MULTI-REPRESENTATION USER MODELING

HOW TO **REPRESENT INTO A USER MODEL** AND **LEVERAGE** IN SEARCH THE USER PREFERENCES INFERRED FROM **HETEROGENEOUS INFORMATION** SOURCES

MULTI-REPRESENTATION USER MODELING

Context : Product Search (e-commerce)

- user preferences and diversity strongly affect the notion of **relevance**

Multi-representation User Model: multiple sources of user-related information:

- User-generated content - e.g., product reviews
- User-items interactions - e.g., view, add to cart, buy…
- Categorical information …….

Personalization process: result re-ranking

Previous works: text only, monolithic architecture

E. Bassani and G. Pasi. *A Multi-Representation Re-Ranking Model for Personalized Product Search*, Information Fusion, 2022

PROPOSED APPROACH

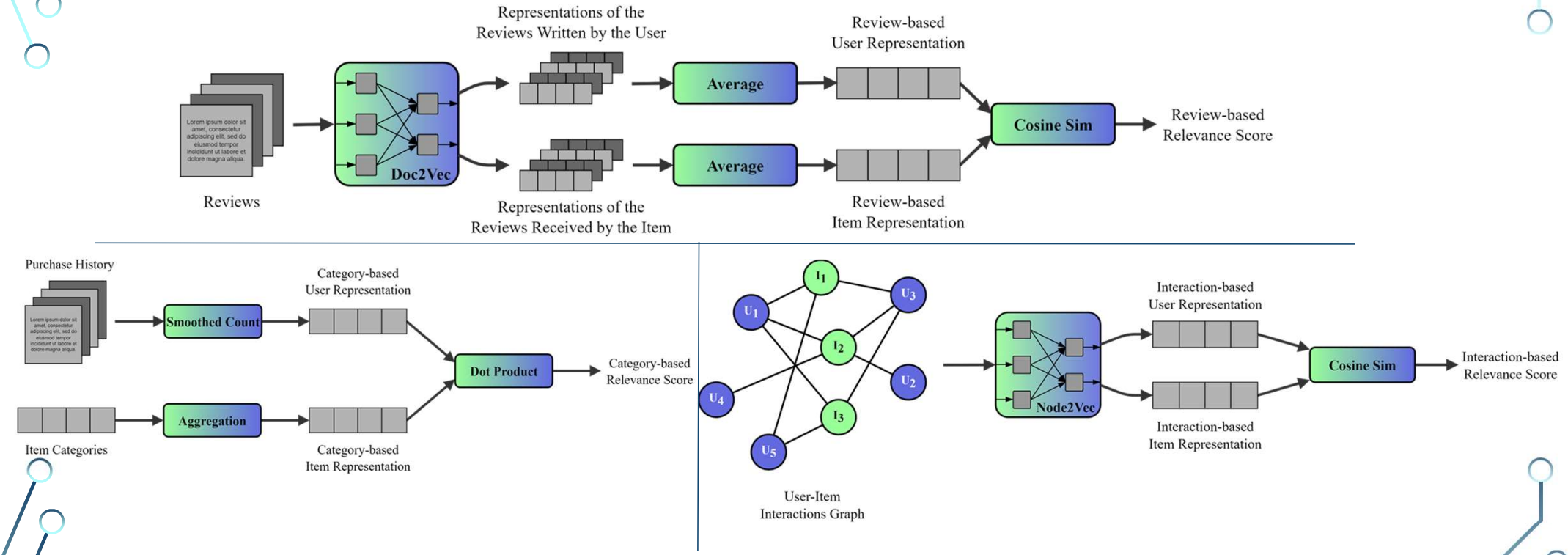
A Modular and **extendable** re-ranking approach that:

- represents separately each type of user related information

- matches each user representation separately with items representations and **combine** the obtained *compatibility* scores between into an overall assessment

- **Both the user and the item representations are built upon:**
  - **Content Information**
    - Reviews
    - Categorical Information
  - **Collaborative Information**
    - User-Item Interactions
    - Item Popularity

MULTIPLE REPRESENTATIONS



EFFECTIVENESS

- We test against
  - classic retrieval model [1]
  - personalized models [2,3]
- **Datasets** built upon **Amazon** data
- Our approach achieved strong improvements on the Electronics and the CDs & Vinyl datasets, **+11% and +7%** on the **NDCG** score, respectively, while performing similarly on the Cell Phones & Accessories

1. Robertson et al., *The probabilistic relevance framework: BM25 and beyond*, Foundations and Trends in Information Retrieval, 2009
2. Ai et al., *Learning a hierarchical embedding model for personalized product search*, SIGIR 2017
3. Ai et al., *Explainable product search with a dynamic relation embedding model*, TOIS 2019

QUERY-AWARE USER MODELING WITH DENOISING ATTENTION

How to exploit the user interests related to the current query and how to decide if personalize query processing?

QUERY-AWARE USER MODELING

Issues:

- not all user information are related to the needs expressed in a specific query
- personalization should not be applied to any query

E .Bassani, P. Kasela, and G. Pasi. *Denoising Attention for Query-aware User Modeling in Personalized Search*, NAACL 2024. Axiv preprint: 2308.15968

QUERY-AWARE USER MODELING

Objective: to *define* an embedded user model at query time, which emphasizes the user's interests aligned with the current query

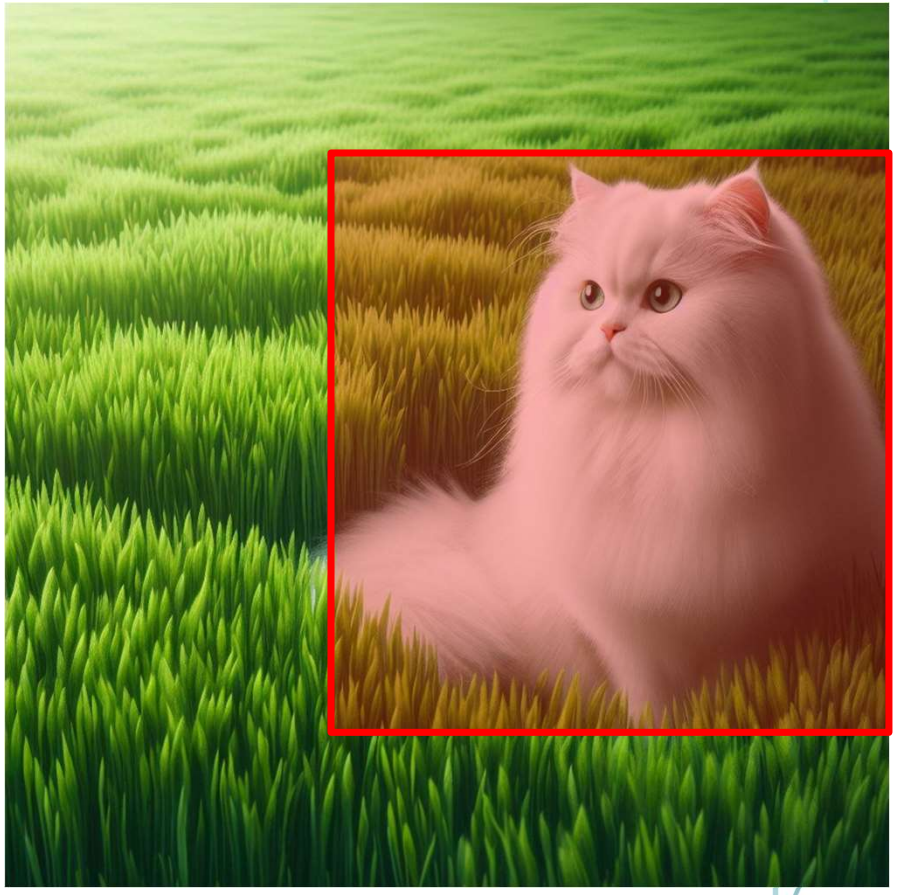
Proposed solution: weighting the user-related information by means of a modified version of the *Attention Mechanism*

E .Bassani, P. Kasela, and G. Pasi. *Denoising Attention for Query-aware User Modeling in Personalized Search*, NAACL 2024. Axiv preprint: 2308.15968

NEURAL ATTENTION MECHANISM

What is Attention mechanism?

- It mimics the selective focus of a human being
- Like in the image you would focus on the cat
- Same applies for texts



NEURAL ATTENTION MECHANISM

Same applies for the texts

The attention mechanism works like the human selective focus

This selective property is very useful for selecting user data important for the user search

NEURAL ATTENTION MECHANISM

- **Attention[1]:**
- computes a context vector by weighting the available contextual information w.r.t. a given input

user model | user-related information sources | query

1. Bahdanau et al., *Neural machine translation by jointly learning to align and translate*, ICLR, 2015

TRADITIONAL ATTENTION MECHANISM
(in our applicative context)

Scoring (alignment):

how well user related documents and the query align (match).

Matching scores are computed for each user-related document

Normalization:

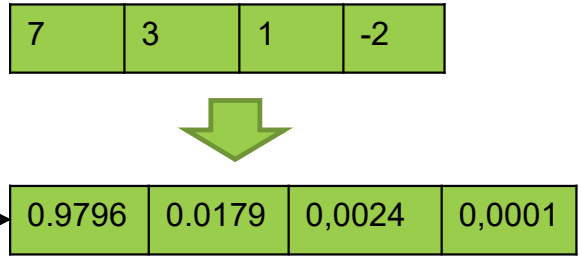
normalization of the matching scores computed by the alignment model, which produces the attention weights. This step is usually accomplished through the use of the Softmax function

Aggregation

The third step consists in the weighted (with attention weights) aggregation of the contextual information (user related documents representations) to produce the context vector, which, in our case, represents the user model.

TRADITIONAL ATTENTION SHORTCOMINGS

- **Softmax normalization:**
  - *Softened* version of Argmax
    - It selects one among *n* options
  - Probability distribution
    - Always positive
    - Sum up to one



- **Shortcomings:**
  - Skewed user representations
  - Noisy user representations
  - Personalization will always be performed (*no zero weights*)

DENOISING ATTENTION MECHANISM

Scoring (alignment):

how well user related documents and the query align (match)

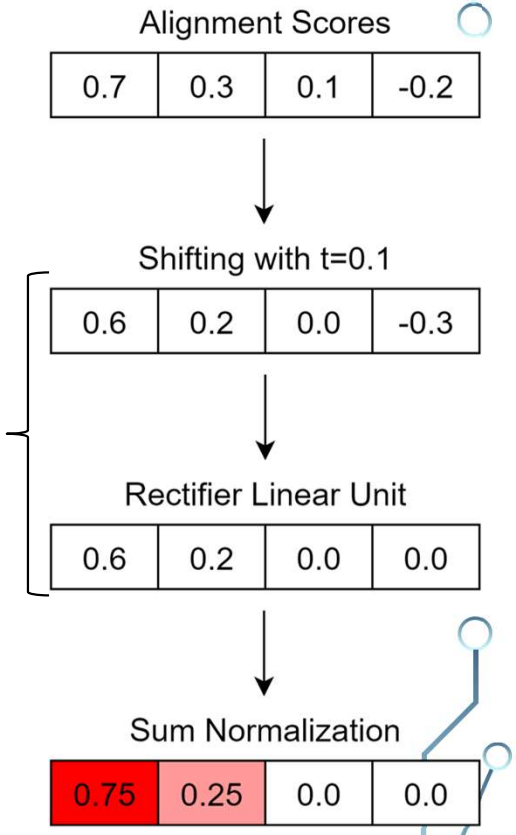
Rectifier Linear Unit (ReLU):

A *two steps filtering step* conceived to not consider user documents loosely related to the query.

Normalization:

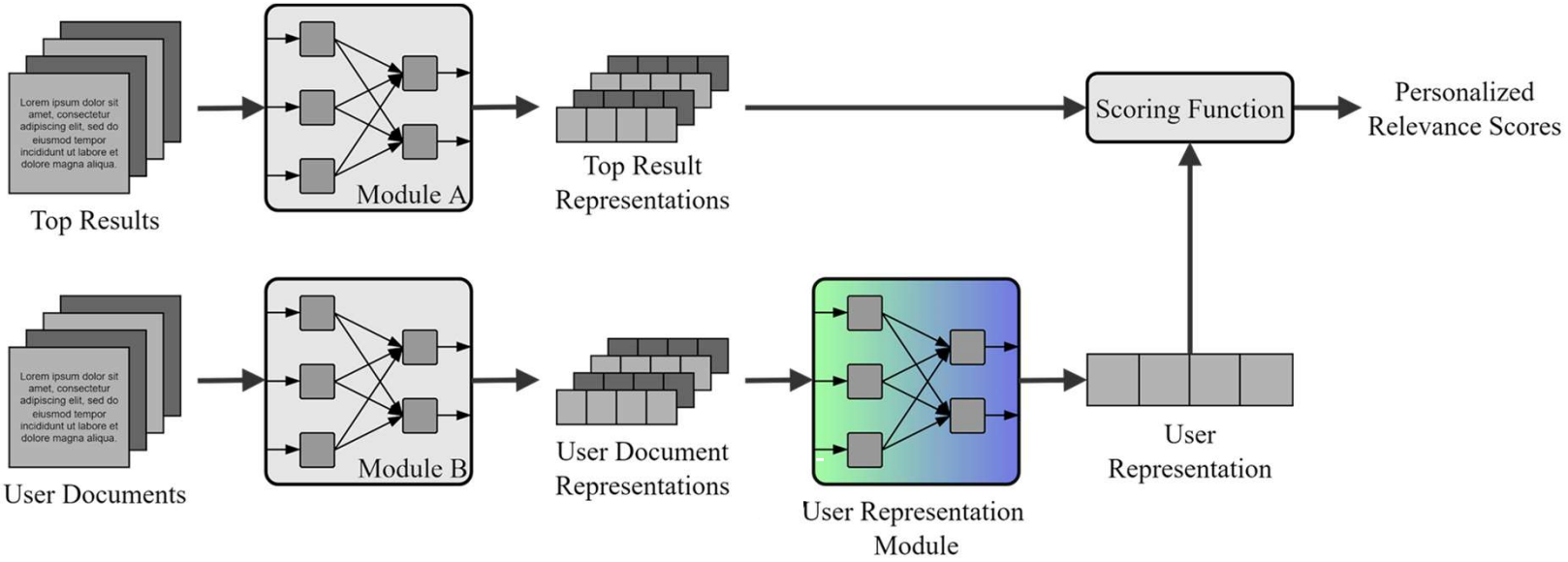
plain normalization operation (of the filtered weights) in replacement of the Softmax for the computation of the *attention weights*.

Can produce zero Attention weights
Can filter out the user model



COMPARISON WITH SOFTMAX

PERSONALIZED RESULTS RE-RANKING FRAMEWORK



EFFECTIVENESS

- We test against
  - one non-personalized model [1]
  - one long term personalized model
  - three user models that relies on attention variants [2,3,4]
- Datasets: Web Search Dataset and Academic Search Dataset
- Improvement **above 15%** over the best performing baseline in both datasets

1. Robertson et al., *The probabilistic relevance framework: BM25 and beyond*, Foundations and Trends in Information Retrieval, 2009
2. Bahdanau et al., *Neural machine translation by jointly learning to align and translate*, ICLR, 2015
3. Ai et al., *A Zero Attention Model for Personalized Product Search*, CIKM, 2019
4. Vaswani et al., *Attention is All You Need*, arXiv, 2017

PERSONALIZED LLMs THROUGH PARAMETER EFFICIENT FINE-TUNING TECHNIQUES

How to **personalize Large Language Models** with user-specific information?

PERSONALIZED LLMs

- **Previous works:**
  - Based on prompting engineering
  - Vanilla Personalized Prompts
  - Retrieval Augmented Generation (RAG)
  - Profile-augmented personalized Prompts

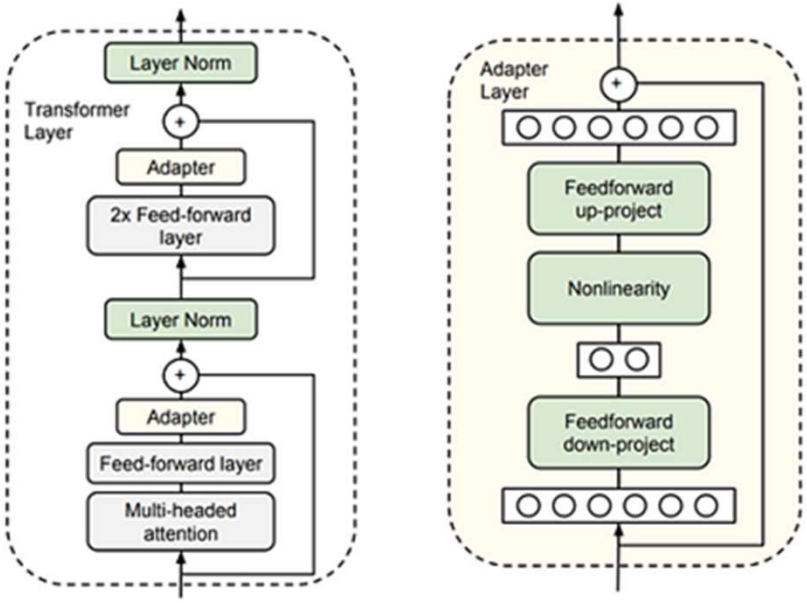
- **Issues with Prompting techniques:**
  - **Small changes** in punctuation can **drop performance** by **80%**
  - **How** to **evaluate** the personalized text and how much the model is personalizing the output

PROPOSED METHOD

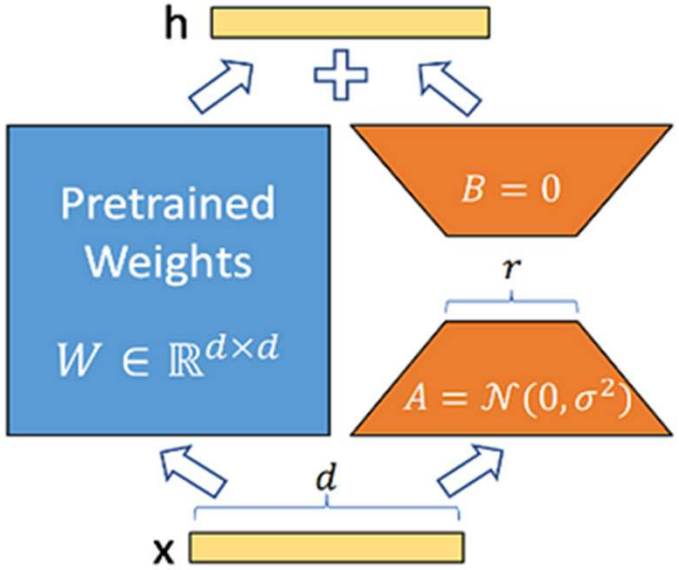
Aim: Personalize a LLM without using prompting engineering

- Fine-tuning a LLM on user data in an efficient way
- Personalize LLM through Parameter Efficient Fine-Tuning techniques (PEFT), i.e Adapters
- Definition of a user-based Mixture of Experts system

Parameter Efficient Fine-Tuning: SOTA



Houlsby Adapter

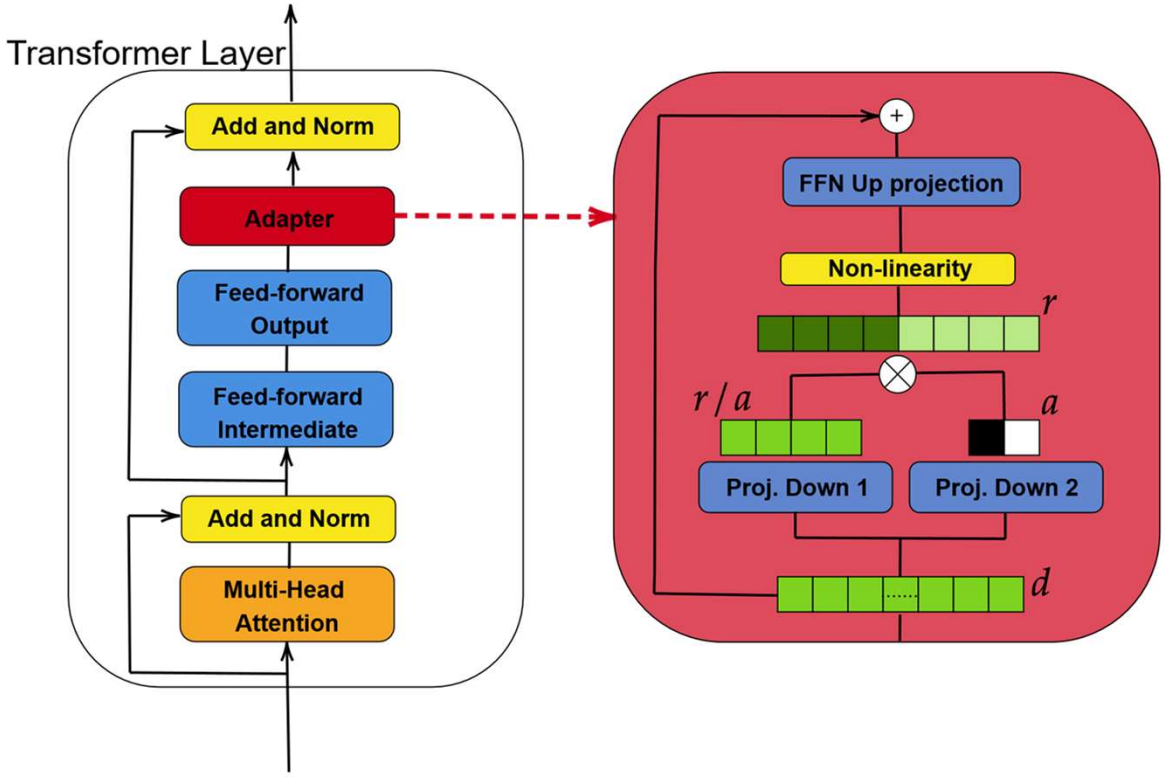


LoRA

State of the Art: Adapter[1] and LoRA[2]

1. Pfeiffer, Jonas, et al. "AdapterHub: A Framework for Adapting Transformers." EMNLP 20
2. Hu, Edward J., et al. "LoRA: Low-Rank Adaptation of Large Language Models." *ICLR 21*

Parameter Efficient Fine-Tuning: AdaKron



Our new proposed approach: **AdaKron**

AdaKron

- Employ the Kronecker product between output vectors of two feed-forward networks (FFN), which compose the down projection of the Adapter.
- The output vector of the Kronecker product has a dimension equal to the product of the dimensions of input vectors.
- Train fewer parameters in the down projection layer compared to a single FFN layer.
- Example:
  - Let be *d* the dimension of the input vector, *48* be the intermediate dimension of the adapter and *4* the dimension of the second down projection. Therefore, AdaKron requires the training of $d * \left(\frac{48}{4} + 4\right) = d * 16$, rather than $d * 48$ parameters, reducing by 33% the number of trained parameters

EFFECTIVENESS

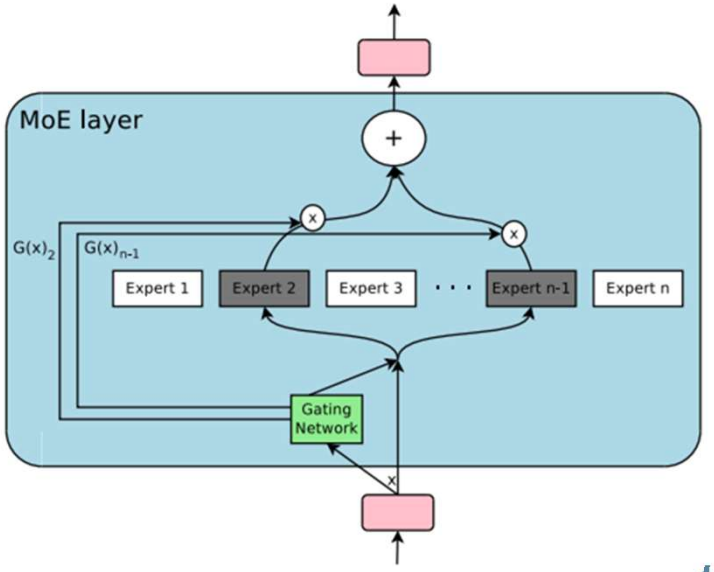
- We test against:
  - Fine-Tuning
  - Houlsby Adapters and LoRa [1,3]
  - Bit-Fit, which trains only bias parameters of the model [2]
- **Datasets:** GLUE, composed of eight different Language Inference tasks
- Better performance compared to the full Fine-Tuning and Houlsby Adapter, achieving **improvements** of **1.0**, and **0.7** average score
- **One** point **improvement** over smaller PEFT methods like BitFit and LoRA.

1. Pfeiffer, Jonas, et al. "AdapterHub: A Framework for Adapting Transformers." EMNLP 20
2. Ben Zaken et al., ”BitFit: Simple Parameter-efficient Fine-tuning for Transformer-based Masked Language-models” ACL 2022
3. Hu, Edward J., et al. "LoRA: Low-Rank Adaptation of Large Language Models." *ICLR 21*

ONGOING WORK: MIXTURE OF EXPERTS

- **Mixture of Experts [1,2,3] :** it is a new paradigm for defining and training neural Language Models
  - They are **pre-trained much faster** than dense models
  - They have **faster inference** compared to a model with the same number of parameters

- Each expert is usually defined as two feed-forward layers. Each expert can receive a group of tokens, sentences or documents, based on the definition of the Gating network



- **AdaKron and Mixture of Experts:** combine them to train in an efficient and effective way an interpretable personalized LLM

1. Shazeer, Noam, et al. "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer." *arXiv preprint arXiv:1701.06538* (2017).
2. Jiang, Albert Q., et al. "Mixtral of experts." arXiv preprint arXiv:2401.04088 (2024).
3. Kasela, Pranav, Gabriella Pasi, Raffaele Perego, and Nicola Tonellotto. "DESIRE-ME: Domain-Enhanced Supervised Information Retrieval Using Mixture-of-Experts." In European Conference on Information Retrieval, pp. 111-125. Cham: Springer Nature Switzerland, 2024.

EVALUATING PERSONALIZED SEARCH

WHAT TO EVALUATE AND HOW TO EVALUATE PERSONALIZED SEARCH?

WHAT IS ASSESSED?

Aim of a search system: to estimate the relevance of the items in a collection wrt a user query. Relevance estimate encompasses various dimensions: topical similarity, popularity, location, etc. Relevance is a multi-dimensional concept

- IRSs assess relevance of an item wrt a query
- Personalized IRSs assess relevance of an item to a query AND to a specific user (model)

EVALUATION OF AN IRS

- In IR, evaluation is a core issue that has been explored since several years
  - Cranfield paradigm: system centered evaluations (offline evaluations)
  - Interactive evaluations, user studies, user-centered
  - Evaluations of personalized search: in between, ranging from offline evaluations, to user studies.
- Measures: set-based measures, rank based measures, user related measures
- What is assessed? Relevance, novelty, diversity, user effort, …

EVALUATION OF PERSONALIZED SEARCH

What we should evaluate?

- Effectiveness of the algorithms
- User satisfaction and user experience
- Quality of the user profile / item profile

Assumption of relevance independence on users is released. From pure system-centered evaluations to user-centered evaluations.

Contributions

- A multi-domain dataset for Academic Search based on Microsoft Citation Graph [1]
- A multi domain cQA dataset by using the user generated content available on the Stackexchange website with two applications:
  - Expert Finding [2]
  - Personalized Question & Answering [3]

[1] Elias Bassani, Pranav Kasela, Alessandro Raganato, Gabriella Pasi. A Multi-Domain Benchmark for Personalized Search Evaluation. In CIKM 2022.
[2] Pranav Kasela, Gabriella Pasi, and Raffaele Perego. 2023. SE-PEF: a Resource for Personalized Expert Finding. In Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region (SIGIR-AP '23)
[3] Pranav Kasela, Marco Braga, Gabriella Pasi, and Raffaele Perego. 2024. SE-PQA : Personalized Community Question Answering. In Companion Proceedings of the ACM Web Conference 2024 (WWW '24 Companion)

CHALLENGES

- Personalization is affecting several NLP related tasks: importance of injecting personal/context knowledge into machine learning based approaches → neuro-symbolic AI
- Learning from multiple representations in NLP related tasks and in particular in search can improve effectiveness
- Personalized learning
- Evaluation of personalization is an important issue

THANK YOU FOR YOUR ATTENTION !