CS60092 Information Retrieval Project Options

Instructor: Somak Aditya Spring 2023

†TAs: Sachin Vashishtha, Bishal Santra, Vivek Karde, Deepak Chaudhary

Project 1: E-commerce Clothing Product Categorization with limited data (Potential TA: Bishal)

Domain: E-commerce/English/Text and images(Multimodal)

Description: The task contains matching images with categories (that belong to a taxonomy). For example, Clothing > Pants > Jeans is a hierarchy and you may have images associated with it. The dataset is available along with well performing models. However, the task is to train models with limited amounts of data.

Suggestions - (1) Checking out contrastive learning-based self-supervised methods that potentially reduce the amount of data required (2) Utilizing well-known taxonomies concepts to enhance the models (ConceptNet/YAGO).

The task can be of two forms. One can input an image and retrieve a serialized categories sequence; or otherwise, one can query a serialized sequence of categories and retrieve a list of images.

Library: Transformers, pytorch

Reference:

1. https://sigir-ecom.github.io/ecom2020/ecom20Papers/paper9.pdf
2. https://github.com/vumaasha/Atlas

Project 2: Content-based Image Retrieval System (Potential TA: Deepak)

Domain: Image search using image

Description: This task gives us a flavor of the reverse image search, where one can use certain images to find related images. Reverse image search has many applications such as in fashion industries. The dataset and baseline methods can be found in https://arxiv.org/pdf/2002.07877v1.pdf .

Library: Transformers, Pytorch

Reference:

1. https://github.com/abhinav23dixit/Text-and-Content-Based-Image-Retrieval

Project 3: Cross-lingual Information Retrieval (CLIR) (Potential TA: Vivek)

Domain: Queries in English and 7 European languages

Description: A CLIR system usually includes two steps, the first step is the translation step, which includes translating either the queries into the language of the document collection, or translating document collection into the query language. After translation is done, the task can be reduced into a monolingual IR task.

However, current multilingual models (such as mBERT, XLM-R) embeds many languages into the same vector space and allows for cross-lingual semantic similarity. There are multiple possibilities. Comparing the methods using translation vs., using semantic similarity matching using mBERT is a start. The question here would be how to improve models and make it more resource-efficient, while maintaining the state-of-the-art accuracy (or F1 scores) for retrieval.

Libraries: transformers, bert, multilabel-classification

References:

1. https://aclanthology.org/2020.acl-main.613.pdf
2. https://github.com/suamin/multilabel-classification-bert-icd10

Project 4: Evidence Retrieval for Fact Verification (Potential TA: Deepak)

Domain: Wikipedia/En

Description: Fact or claim verification is a two-step process. First, you retrieve supporting or refuting evidence related to a claim. Then based on the set of evidence snippets, the task is to determine whether the claim is true or false. In this project, we are interested in the first step, i.e., evidence retrieval.

Dataset: FEVER Dataset

References:

1. https://arxiv.org/pdf/1908.01843.pdf
2. Baseline: https://github.com/thunlp/GEAR

Project 5: Offensive query detection (on reddit/Twitter dataset) and generalization to multi-lingual setting (Potential TA: Bishal)

Domain: Social Media/En (and other languages)

Description: Offensive queries have become an important avenue for search engine companies. Often new socio-political events trigger new searches, which if shown as suggestions, can be deemed offensive to the users. Offensive text detection and generalizing to multilingual settings, hence, is of high relevance to many companies.

In the below datasets, the task is to propose an approach to automatically classify the tweets into 3 classes : hateful, offensive and clean. Test and compare various models for Hate-Speech detection on basis of Precision, Recall and F1 score. Most importantly, show how similar methods can generalize to multilingual settings. Here, we expect you to propose methods that do not use multilingual models (or ULMs such as XLM-R and mBERT).

Dataset: https://github.com/sayarghoshroy/Hate-Speech-Detection ,
https://github.com/mohit19014/Hindi-Hostility-Detection-CONSTRAINT-2021 ,
https://github.com/renuka-fernando/sinhalese_language_racism_detection

References:

1. https://arxiv.org/pdf/2010.12472.pdf

Project 6: Efficient and Fast Image Retrieval (Potential TA: Deepak)

Domain: Images and Text

Description: Regarding content-based image retrieval, it is often claimed that visually similar images are clustered in this feature space. However, there are two major problems with this approach: 1) Visual similarity does not always correspond to semantic similarity. 2) The classification objective does not enforce a high distance between different classes, so that the nearest neighbors of some images may belong to completely different classes.

Hierarchy-based semantic embeddings overcome these issues. The task is to learn semantic embeddings(more details are provided in paper) using the CIFAR-100 dataset.

Variation: An interesting **extension/variation** to above is utilizing the semantic embeddings; how do we create an indexing strategy which makes query processing fast and efficient. You may look at

https://towardsdatascience.com/billion-scale-semantic-similarity-search-with-faiss-sbert-c845614962e2 and https://blog.vespa.ai/billion-scale-knn/ for reference.

Libraries: CNN, clustering

Dataset: https://www.cs.toronto.edu/~kriz/cifar.html

References:

1. https://arxiv.org/abs/1809.09924
2. https://github.com/cvjena/semantic-embeddings

Project 7: Efficient API/Code Snippet Retrieval (Potential TA: Sachin)

Domain: Code-snippets and text

Description: Imagine, rather than waiting for answers in stack-overflow; you can query using natural language and the search engine comes up with a plausible code snippet. This facility can help ease many simpler tasks, which often a professional programmer needs to re-do. Such motivation has motivated the PL (programming languages) and ML community to come together and propose Deep neural-net based code search; which has seemingly become quite efficient. Here, based on the dataset, given a query task is to retrieve the most relevant code snippet.

Dataset: https://conala-corpus.github.io/ (data should be used in a retrieval setting)

References:

1. https://arxiv.org/pdf/2008.12193.pdf
2. https://people.eecs.berkeley.edu/~ksen/papers/ncs.pdf
3. https://github.com/nokia/codesearch

Project 8: Query-by-Example for Scientific Article Retrieval (Potential TA: Sachin)

Domain: Scientific Publications/En

Description: Using background/objective, method or results as queries, the task is to retrieve most similar scientific papers. The dataset CSFCube contains an annotated test set for validation.

Dataset: https://arxiv.org/pdf/2103.12906.pdf, https://github.com/iesl/CSFCube

Project 9: Information Retrieval in a Code-mixed context (Potential TA: Vivek)

Description: MSR India researchers recently published a large collection of tasks in different code-mixed languages under the umbrella of GLUECoS. Here, one can attempt to solve QA task in a limited resource scenario as a retrieval problem. Say, given a question and a paragraph (context), rank and retrieve relevant sentences. Track Recall@K, Precision@K (K=1, 5, 10). We mark a sentence correct if the annotated answer is present in the sentence

The challenge again here is to avoid large Universal Language models (ULMs), as student groups may not have such resources. How do we build plausible intuitive ML systems that can process such code-mixed data, and yet be competitive.

Library: pytorch, simpletransformers, huggingface

Reference:

1. https://aclanthology.org/2020.acl-main.329.pdf

Project 10: Breast Cancer Image Retrieval (Potential TA: Bishal)

Domain: Medical Image Retrieval

Description: Content Based Medical Image Retrieval (CBMIR) is considered as a common technique to retrieve relevant images by comparing the features contained in the query image with the features contained in the image located in the database. Currently, the study related to CBMIR on breast cancer image however remains challenging due to inadequate research in such area. Hence, the goal is to retreive highly relevant Medical Images related to Breast Cancer which can help in diagnosis.

How CBMIR help in diagnosis? Medical-image-based diagnosis is a tedious task, and small lesions in various medical images can be overlooked by medical experts due to the limited attention span of the human visual system, which can adversely affect medical treatment. In this case, we can use CBMIR.

Dataset: BreakHis Dataset

http://www.inf.ufpr.br/vri/databases/BreaKHis_v1.tar.gz

Code: 1. https://github.com/forderation/breast-cancer-retrieval

2. https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/

References: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9502205

Project 11: Adapting Pretrained Dialog Retrieval Models for Dialog System Evaluation

Domain: Social Media Chatbots / Dialog Systems

Description: The goal of dialog system evaluation is to assess the system's ability to understand and respond appropriately to user needs, and to identify areas for improvement. This can be done through a variety of methods, such as human evaluations, automated metrics, and user studies. As human evaluations are costly to conduct at scale, the design of automated metrics for response generation is a crucial area of research in the dialog generation community. However, designing these metrics is challenging as it involves measuring subjective metrics such as fluency, coherence, relevance, and consistency.

In this project, we aim to investigate the feasibility of using pretrained dialog retrieval models, such as PolyEncoder and DMI, as a metric for evaluating dialog systems. We propose to employ finetuning techniques on top of existing large scale retrieval models to adapt these pretrained models for the dialog evaluation task. The goal of this research is to develop an efficient and reliable evaluation metric for dialog systems that can be applied to various domains and contexts.

Datasets:

GRADE, Holistic-Eval, USR, DSTC9

Code:

https://github.com/exe1023/DialEvalMetrics

https://github.com/tanyuqian/ctc-gen-eval

References:

[2106.03706] A Comprehensive Assessment of Dialog Evaluation Metrics

[2109.06379] Compression, Transduction, and Creation: A Unified Framework for Evaluating Natural Language Generation

Poly-encoders: Transformer Architectures and Pre-training Strategies for Fast and Accurate Multi-sentence Scoring

On Large Language Models for Understanding Human Language Christopher Manning (Youtube Video)

Project 12: Preventing Hallucinations in Dialog Generation using Retrieval Models to Search for Explanation

Domain: Social Media Chatbots / Dialog Systems

Description:

Hallucination in language generation models refers to the model generating text that is not based on the input provided, but rather on patterns it has learned from the training data. This can lead to the model producing text that is not coherent or accurate. In simpler terms, hallucination means the model is making up or imagining things that are not real.

This research project aims to investigate a retrieval based solution for preventing hallucinations in deep learning based dialog generation models (or any language models in general). The proposed solution includes incorporating a justification search component as part of the model and allowing the model to back off when a valid justification is not available. The goal of this research is to enhance the reliability and credibility of dialog generation by providing a justification for every generation.

Datasets:

Wizard-of-Wikipedia

Code:

https://github.com/prakharguptaz/EDGE-exemplars

References:

Controlling Dialogue Generation with Semantic Exemplars - ACL Anthology

Explainable AI: Language Models