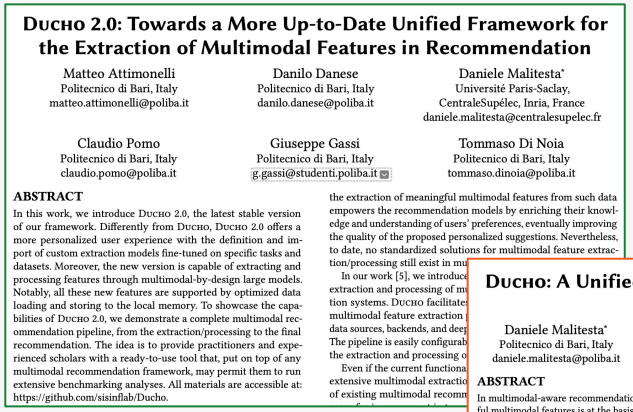
07

Introduction to Ducho

What is Ducho?

Ducho is a unified framework that facilitates the extraction of multimodal features for recommendation.



Ducho 2.0: Towards a More Up-to-Date Unified Framework for the Extraction of Multimodal Features in Recommendation

Matteo Attimonelli
Politecnico di Bari, Italy
matteo.attimonelli@poliba.it

Danilo Danese
Politecnico di Bari, Italy
danilo.danese@poliba.it

Daniele Malitesta*
Université Paris-Saclay, CentraleSupélec, Inria, France
daniele.malitesta@centralesupelec.fr

Claudio Pomo
Politecnico di Bari, Italy
claudio.pomo@poliba.it

Giuseppe Gassi
Politecnico di Bari, Italy
g.gassi@studenti.poliba.it

Tommaso Di Noia
Politecnico di Bari, Italy
tommaso.dinoia@poliba.it

ABSTRACT

In this work, we introduce Ducho 2.0, the latest stable version of our framework. Differently from Ducho, Ducho 2.0 offers a more personalized user experience with the definition and import of custom extraction models fine-tuned on specific tasks and datasets. Moreover, the new version is capable of extracting and processing features through multimodal-by-design large models. Notably, all these new features are supported by optimized data loading and storing to the local memory. To showcase the capabilities of Ducho 2.0, we demonstrate a complete multimodal recommendation pipeline, from the extraction/processing to the final recommendation. The idea is to provide practitioners and experienced scholars with a ready-to-use tool that, put on top of any multimodal recommendation framework, may permit them to run extensive benchmarking analyses. All materials are accessible at: https://github.com/sisinflab/Ducho.

Ducho: A Unified Framework for the Extraction of Multimodal Features in Recommendation

Daniele Malitesta*
Politecnico di Bari, Italy
daniele.malitesta@poliba.it

Giuseppe Gassi*
Politecnico di Bari, Italy
g.gassi@studenti.poliba.it

Claudio Pomo
Politecnico di Bari, Italy
claudio.pomo@poliba.it

Tommaso Di Noia
Politecnico di Bari, Italy
tommaso.dinoia@poliba.it

ABSTRACT

In multimodal-aware recommendation, the extraction of meaningful multimodal features is at the basis of high-quality recommendations. Generally, each recommendation framework implements its multimodal extraction procedures with specific strategies and tools. This is limiting for two reasons: (i) different extraction strategies do not ease the interdependence among multimodal recommendation frameworks; thus, they cannot be efficiently and fairly compared; (ii) given the large plethora of pre-trained deep learning models made available by different open source tools, model designers do not have access to shared interfaces to extract features. Motivated by the outlined aspects, we propose Ducho, a unified framework for the extraction of multimodal features in recommendation. By integrating three widely-adopted deep learning libraries as backends, namely, TensorFlow, PyTorch, and Transformers, we provide a shared interface to extract and process features where each backend's specific methods are abstracted to the end user. Noteworthy, the extraction pipeline is easily configurable with a YAML-based file where the user can specify, for each modality, the list of models (and their specific backends/parameters) to perform the extraction. Finally, to make Ducho accessible to the community, we build a

1 INTRODUCTION AND MOTIVATION

With the advent of the digital era and the Internet, numerous online services have emerged, including platforms for e-commerce, media streaming, and social networks. The vast majority of such websites rely on recommendation algorithms to provide users with a personalized surfing experience. In specific domains such as fashion [3], music [6], food [5], and micro-video [8] recommendation, recommender systems have demonstrated to be effectively supported in their decision-making process by all types of multimodal data sources the users usually interact with (e.g., product images and descriptions, users' reviews, audio tracks).

The literature refers to multimodal-aware recommender systems (MRSs) as the family of recommendation algorithms leveraging multimodal (i.e., audio, visual, textual) content data to augment the representation of items, thus tackling issues in the field such as the sparsity of the user-item matrix and the inexplicable nature of users' actions (e.g., clicks, views) on online platforms which may not always be easy to profile for the recommendation algorithms.

Despite being the initial stage of any multimodal recommendation pipeline, the extraction of meaningful multimodal features is paramount in delivering high-quality recommendations [2]. How-

[44] Attimonelli et al., Ducho 2.0: Towards a More Up-to-Date Unified Framework for the Extraction of Multimodal Features in Recommendation. WWW (Companion Volume) 2024: 1075-1078

[45] Malitesta et al., Ducho: A Unified Framework for the Extraction of Multimodal Features in Recommendation. ACM Multimedia 2023: 9668-9671

Why Ducho?

Integration

Ducho provides a standardized framework

Simplicity

Works via a simple yaml-based interface

Flexibility

Provides different processing and fusion techniques

Reproducibility

Ensures reproducibility for feature extraction and processing

Features

Customization and Optimization

- Faster data loading/storing
- Custom processing strategies
- Allows the extraction of features using custom models

Backends and Large Multimodal Models (LMMs)

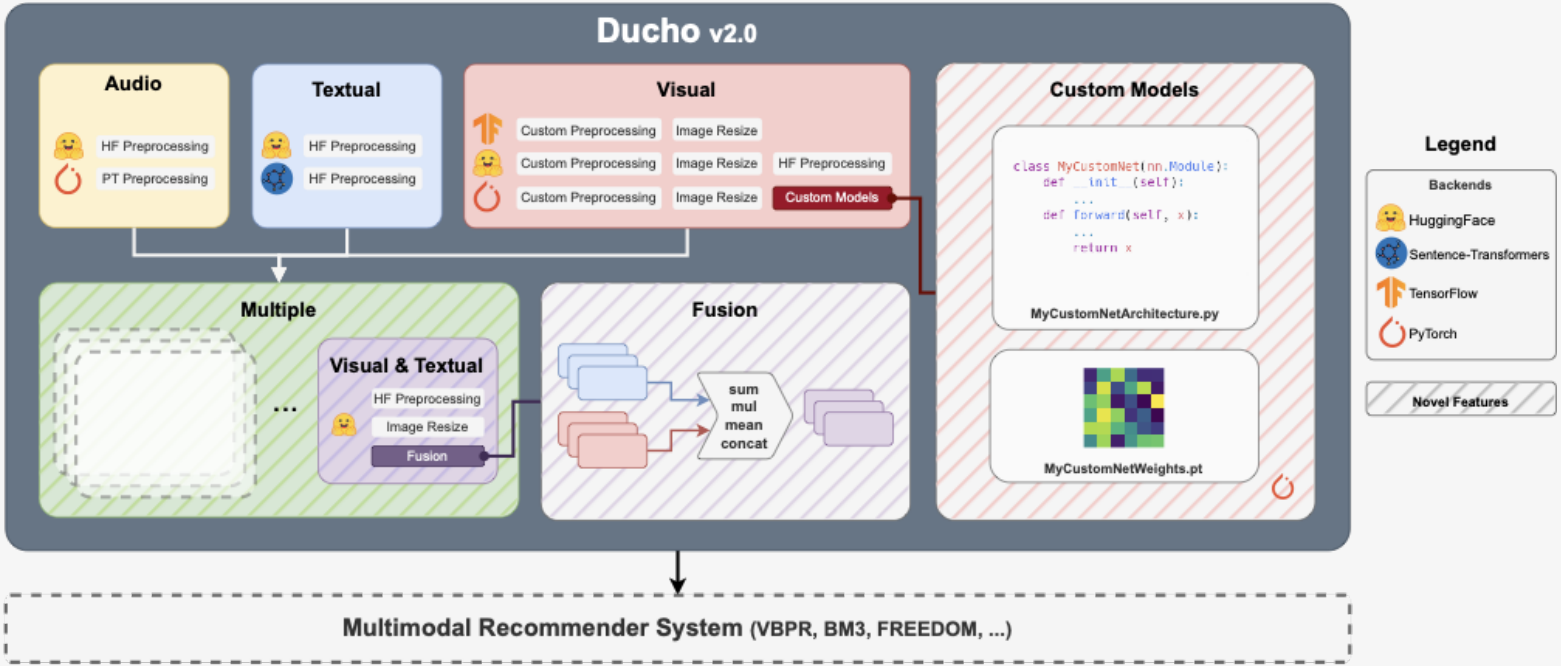
- TensorFlow, PyTorch, Transformers, Sentence-Transformers
- Extraction of Multimodal-by-design features via LMMs
- Facilitate modality fusion

Configuration File

- Specify different modalities
- Set source and output paths/files
- Define (custom) extractors
- Fuse extracted features

```
dataset_path: ./my/dataset/path
gpu list: 0
visual:
 items:
  input_path:  images
  output_path: visual_embeddings
  model: [
    { model_name: ResNet18, output_layers: avgpool,
      reshape: [224, 224], backend: torch, preprocessing: zscore,
      mean: [0.485, 0.456, 0.406], std: [0.229, 0.224, 0.225] },
    { model_name: ./MyCustomNetWeights.pt, backend: torch,
      output_layers: pooler_output, preprocessing: minmax },
    { model_name: ./MyCustomHFModel, backend: transformers,
      output_layers: [MyCustomOutputLayer, avgpool],
      image_processor: ./MyCustomImageProcessor } ]
textual:
 items:
  input_path:  descriptions.tsv
  output_path: textual_embeddings
  item_column: asin
  text_column: description
  model: [
   { model_name: ./MyCustomHFModel, clear_text: False,
     output_layers: MyCustomOutputLayer, backend: transformers,
     tokenizer_name: ./MyCustomTokenizer } ]
visual_textual:
 items:
  input_path:  { visual: images, textual: meta.tsv }
  output_path: { visual: vis_embeddings, textual: text_embeddings }
  item_column: asin
  text_column: description
  model: [
   { model_name: openai/clip-vit-base-patch16, fusion: concat,
     output_layers: 1, backend: transformers  } ]
```

Architecture Overview



[46] Attimonelli et al., Ducho meets Elliot: Large-scale Benchmarks for Multimodal Recommendation. CoRR abs/2409.15857 (2024)