MultiSChuBERT: Effective Multimodal Fusion for Scholarly Document Quality Prediction

Gideon Maillette de Buy Wenniger

Thomas van Dongen
Lambert Schomaker

university of groningen

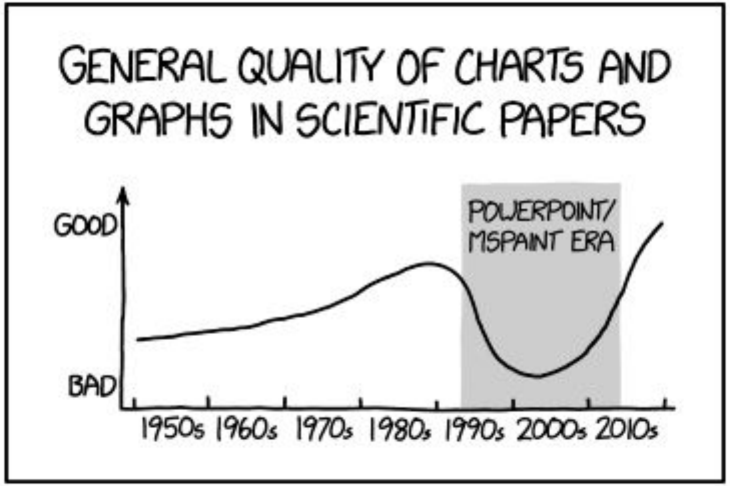
Open Universiteit
www.ou.nl

Date: 7-11-2023

Scholarly Document Quality Prediction

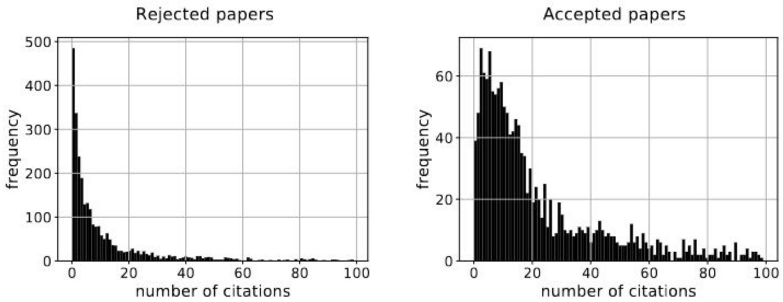
- Predict quality from the document alone: Textual vs visual clues on quality

- What indicators of quality to predict?
  - Accept/Reject
    - Simple and well understood
    - Scarce data
  - Number of Citations
    - Large data availability
    - We predict: log(#citations +1)

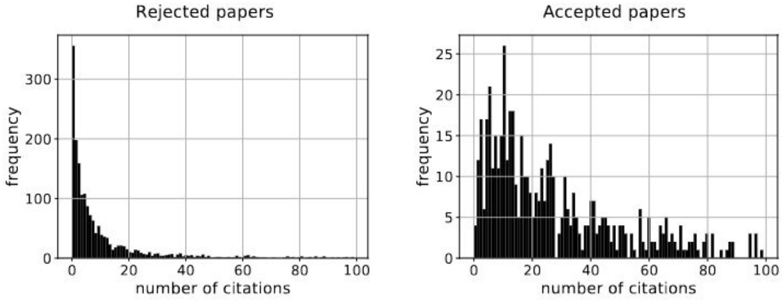


Source:https://m.xkcd.com/1945/

Is it reasonable to look at #citations?



(a) Machine Learning domain.

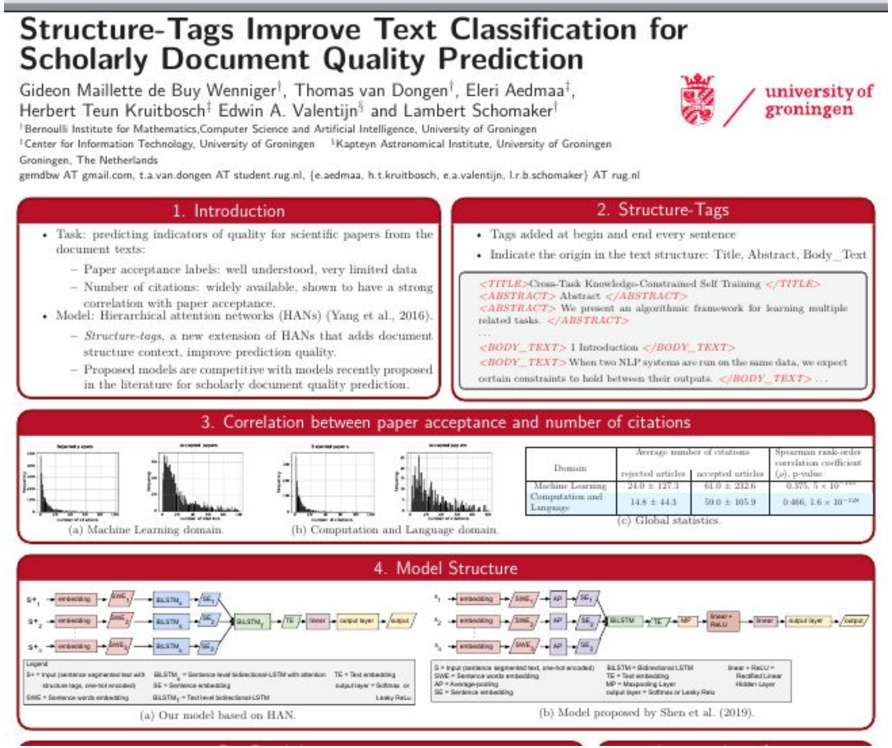


(b) Computation and Language domain.

Domain	Average number of citations		Spearman rank-order correlation coefficient (ρ), p-value
	rejected articles	accepted articles	
Machine Learning	24.0 ± 127.3	61.0 ± 232.6	$0.375,\ 5 \times 10^{-153}$
Computation and Language	14.8 ± 44.3	59.0 ± 105.9	$0.466,\ 1.6 \times 10^{-128}$

(c) Global statistics.

Background earlier work



Structure-Tags Improve Text Classification for Scholarly Document Quality Prediction

Gideon Maillette de Buy Wenniger[†], Thomas van Dongen[†], Eleri Aedmaa[‡], Herbert Teun Kruitbosch[‡] Edwin A. Valentijn[§] and Lambert Schomaker[†]
[†]Bernoulli Institute for Mathematics,Computer Science and Artificial Intelligence, University of Groningen
[‡]Center for Information Technology, University of Groningen [§]Kapteyn Astronomical Institute, University of Groningen
Groningen, The Netherlands
gemdbw AT gmail.com, t.a.van.dongen AT student.rug.nl, {e.aedmaa, h.t.kruitbosch, e.a.valentijn, l.r.b.schomaker} AT rug.nl

university of groningen

1. Introduction

- Task: predicting indicators of quality for scientific papers from the document texts:
  - Paper acceptance labels: well understood, very limited data
  - Number of citations: widely available, shown to have a strong correlation with paper acceptance.
- Model: Hierarchical attention networks (HANs) (Yang et al., 2016).
  - *Structure-tags*, a new extension of HANs that adds document structure context, improve prediction quality.
  - Proposed models are competitive with models recently proposed in the literature for scholarly document quality prediction.

2. Structure-Tags

- Tags added at begin and end every sentence
- Indicate the origin in the text structure: Title, Abstract, Body_Text

<TITLE>Cross-Task Knowledge-Constrained Self Training </TITLE>
...
<BODY_TEXT> 1 Introduction </BODY_TEXT>
<BODY_TEXT> When two NLP systems are run on the same data, we expect certain constraints to hold between their outputs. </BODY_TEXT> ...

3. Correlation between paper acceptance and number of citations

(a) Machine Learning domain.

(b) Computation and Language domain.

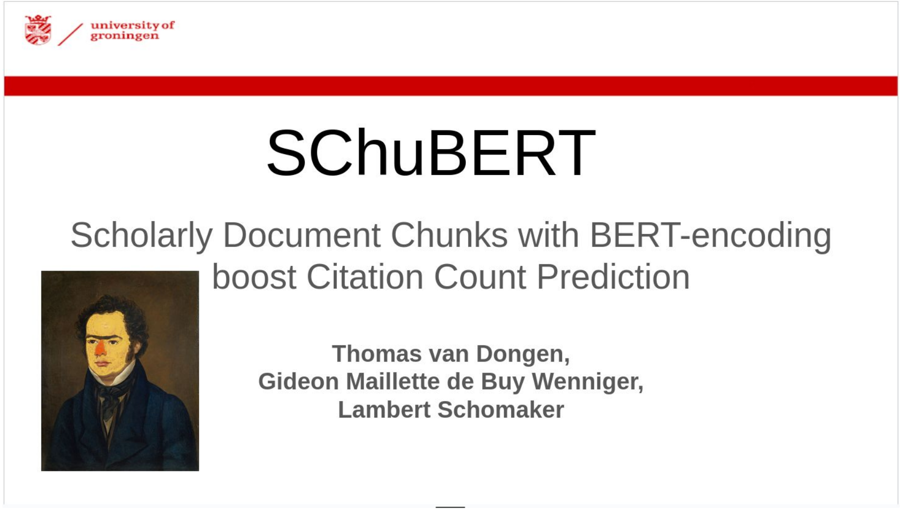
Domain	Average number of citations		Spearman rank-order correlation coefficient (ρ), p-value
	rejected articles	accepted articles	
Machine Learning	24.0 ± 127.3	61.0 ± 232.6	0.375, 5 × 10[illegible]
Computation and Language	14.8 ± 44.3	59.0 ± 105.9	0.466, 1.6 × 10[illegible]

(c) Global statistics.

4. Model Structure

(a) Our model based on HAN.

(b) Model proposed by Shen et al. (2019).



This work: multimodality

university of groningen / faculty of science and engineering

MASTER'S THESIS

UNIVERSITY OF GRONINGEN

DEPARTMENT OF ARTIFICIAL INTELLIGENCE

Quality Prediction of Scientific Documents Using Textual and Visual Content

Author:
Thomas Anton van Dongen

First Supervisor:
Prof. dr. L.R.B. Schomaker

Second Supervisor:
Dr. G.E. Maillette de Buy Wenniger

March 22, 2021

Cornell University

We gratefully acknowledge supp

arXiv > cs > arXiv:2308.07971

Search...

Help | Advanced

Computer Science > Computation and Language

[Submitted on 15 Aug 2023]

MultiSChuBERT: Effective Multimodal Fusion for Scholarly Document Quality Prediction

Gideon Maillette de Buy Wenniger, Thomas van Dongen, Lambert Schomaker

Automatic assessment of the quality of scholarly documents is a difficult task with high potential impact. Multimodality, in particular the addition of visual information next to text, has been shown to improve the performance on scholarly document quality prediction (SDQP) tasks. We propose the multimodal predictive model MultiSChuBERT. It combines a textual model based on chunking full paper text and aggregating computed BERT chunk-encodings (SChuBERT), with a visual model based on Inception V3.Our work contributes to the current state-of-the-art in SDQP in three ways. First, we show that the method of combining visual and textual embeddings can substantially influence the results. Second, we demonstrate that gradual-unfreezing of the weights of the visual sub-model, reduces its tendency to ovefit the data, improving results. Third, we show the retained benefit of multimodality when replacing standard $\text{BERT}_{\text{BASE}}$ embeddings with more recent state-of-the-art text embedding models.
Using $\text{BERT}_{\text{BASE}}$ embeddings, on the (log) number of citations prediction task with the ACL-BiblioMetry dataset, our MultiSChuBERT (text+visual) model obtains an R^2 score of 0.454 compared to 0.432 for the SChuBERT (text only) model. Similar improvements are obtained on the PeerRead accept/reject prediction task. In our experiments using SciBERT, scincl, SPECTER and SPECTER2.0 embeddings, we show that each of these tailored embeddings adds further improvements over the standard $\text{BERT}_{\text{BASE}}$ embeddings, with the SPECTER2.0 embeddings performing best.

Subjects: **Computation and Language (cs.CL)**; Machine Learning (cs.LG)

https://arxiv.org/abs/2308.07971

Selected Related work

Journal of Artificial Intelligence Research 68 (2020) 607-632 Submitted 09/2019; published 07/2020

A Multimodal Approach to Assessing Document Quality

Aili Shen AILIS@STUDENT.UNIMELB.EDU.AU
Bahar Salehi BAHAR.SALEHI@GMAIL.COM
Jianzhong Qi JIANZHONG.QI@UNIMELB.EDU.AU
Timothy Baldwin TB@LDWIN.NET
School of Computing and Information Systems
The University of Melbourne
Victoria 3010, Australia

Abstract

The perceived quality of a document is affected by various factors, including grammaticality, readability, stylistics, and expertise depth, making the task of document quality assessment a complex one. In this paper, we explore this task in the context of assessing the quality of Wikipedia articles and academic papers. Observing that the visual rendering of a document can capture implicit quality indicators that are not present in the document text — such as images, font choices, and visual layout — we propose a joint model that combines the text content with a visual rendering of the document for document quality assessment. Our joint model achieves state-of-the-art results over five datasets in two domains (Wikipedia and academic papers), which demonstrates the complementarity of textual and visual features, and the general applicability of our model. To examine what kinds of features our model has learned, we further train our model in a multi-task learning setting, where document quality assessment is the primary task and feature learning is an auxiliary task. Experimental results show that visual embeddings are better at learning structural features while textual embeddings are better at learning readability scores, which further verifies the complementarity of visual and textual features.

Main Question

How can we (still) get benefit from multimodality in combination with stronger textual encoders, and while using domain-specialized text embedding?

Statistics of the used datasets

(a) Data sizes and label types

Dataset	#Documents (train + validation + test)	Labels
AI	4092 (3682 + 205 + 205)	Accept/reject
CL	2638 (2374 + 132 + 132)	Accept/reject
LG	5048 (4543 + 252 + 253)	Accept/reject
ACL-BiblioMetry	30950 (27853 + 1548 + 1549)	Citations

(b) PeerRead accept/reject distribution

Dataset	Train Accept : Reject	Validation Accept : Reject	Test Accept : Reject
AI	10.5% : 89.5%	8.3% : 91.7%	7.8% : 92.2%
CL	24.3% : 75.7%	22.0% : 78.0%	31.1% : 68.9%
LG	36.4% : 63.6%	36.5% : 63.5%	32.0% : 68.0%

Data and scope this presentation

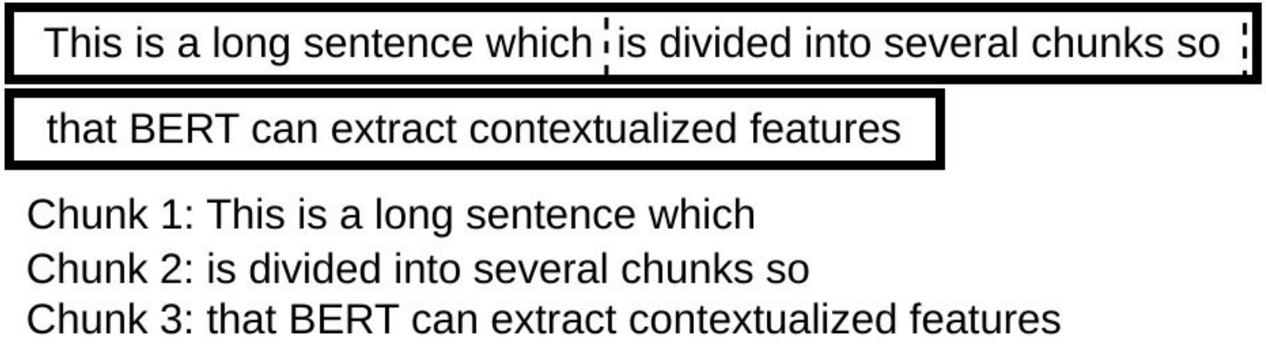
- We experiment with the ACL Bibliometry (number of citations prediction) and PeerRead (accept reject prediction) datasets

- ACL Bibliometry is much larger (30950 examples) than even the largest PeerRead (LG) subset (5048 examples)
  – Stronger models that use more context (multimodality) and full text input have more chance to thrive with larger training data

- We will focus on the ACL Bibliometry results in this presentation, but more results available in: https://arxiv.org/abs/2308.07971
  – Main findings on Peer Read are similar to those on

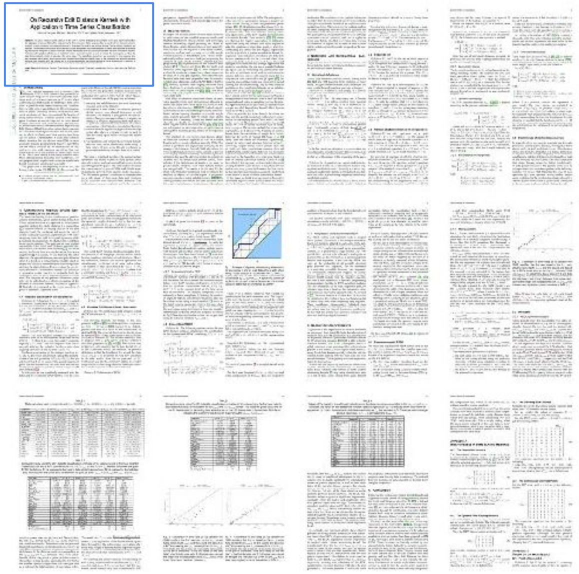
    ACL Bibliometry

Textual model input



Example of the chunking method. In this example, a sequence length of 6 used with no overlap.

Visual model input: *Overall* appearance and layout

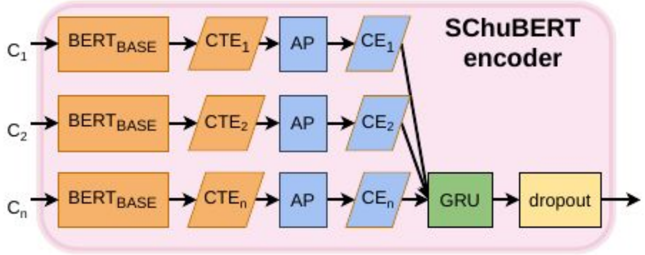


(a) Document grid – overview.

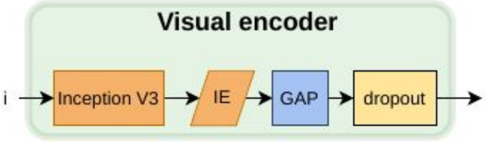
(b) Top-left part of the document grid, containing title and abstract.

Example of created document grid. The grid contains 12 pages and is of size 512x512.

Overview of the used Text,Visual and Multimodal model



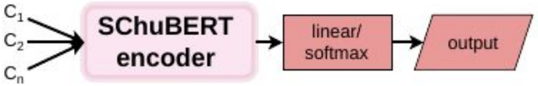
(a) SChuBERT encoder.



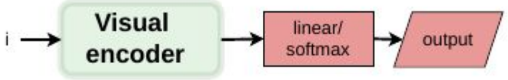
(b) Visual encoder.

Legend
C = Input (512 length text chunk, tokenized for BERT)
CTE = Chunk token embedding
AP = Average-pooling
CE = Chunk embedding
GRU = Gated recurrent unit
i = Input (512x512 Grid of paper pages)
IE = image embedding
GAP = Global Average Pooling
Conc = Concatenation Layer

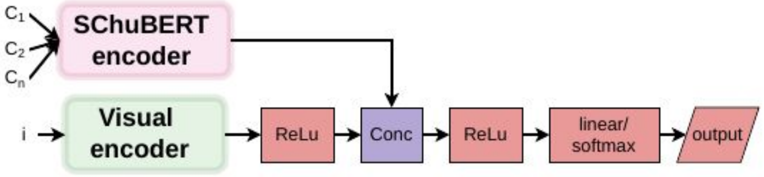
(c) Legend used symbols.



(d) The SChuBERT model (van Dongen, Maillette de Buy Wenniger, and Schomaker 2020).

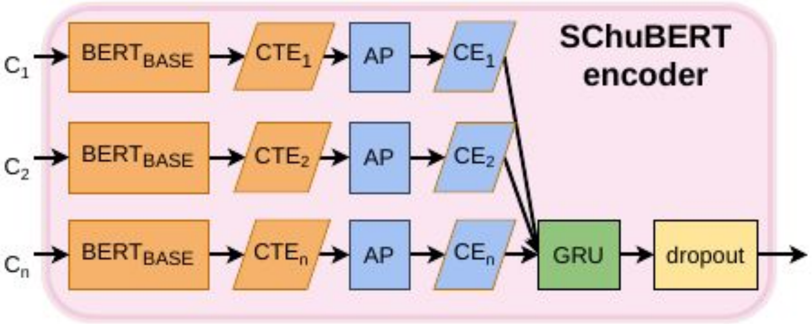


(e) The $INCEPTION_{GU}$ model proposed in this work, based of the INCEPTION model from (Shen et al. 2019).



(f) The MultiSChuBERT model proposed in this work.

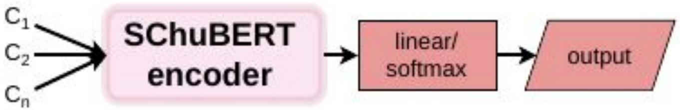
The SChuBERT model (text)



(a) SChuBERT encoder.

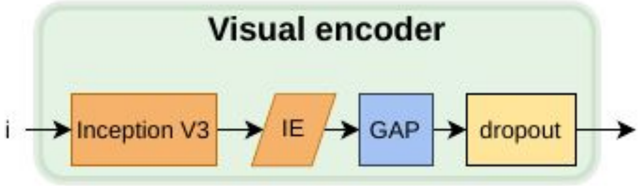
Legend
C = Input (512 length text chunk, tokenized for BERT)
CTE = Chunk token embedding
AP = Average-pooling
CE = Chunk embedding
GRU = Gated recurrent unit
i = Input (512x512 Grid of paper pages)
IE = image embedding
GAP = Global Average Pooling
Conc = Concatenation Layer

(c) Legend used symbols.



(d) The SChuBERT model (van Dongen, Maillette de Buy Wenniger, and Schomaker 2020).

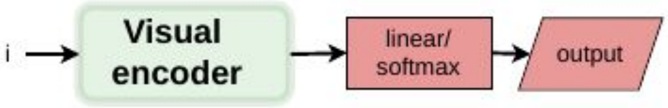
The INCEPTION model (visual)



(b) Visual encoder.

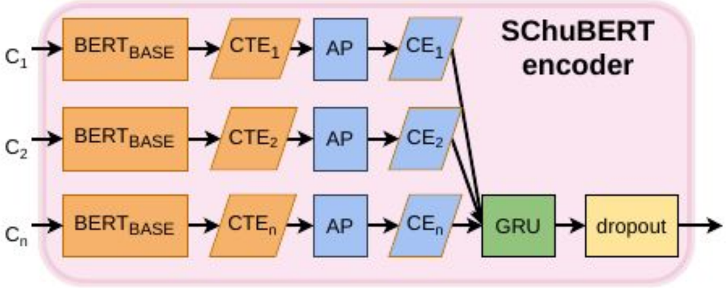
Legend
C = Input (512 length text chunk, tokenized for BERT)
CTE = Chunk token embedding
AP = Average-pooling
CE = Chunk embedding
GRU = Gated recurrent unit
i = Input (512x512 Grid of paper pages)
IE = image embedding
GAP = Global Average Pooling
Conc = Concatenation Layer

(c) Legend used symbols.

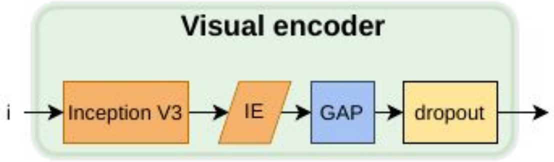


(e) The $\text{INCEPTION}_{\text{GU}}$ model proposed in this work, based of the INCEPTION model from (Shen et al. 2019).

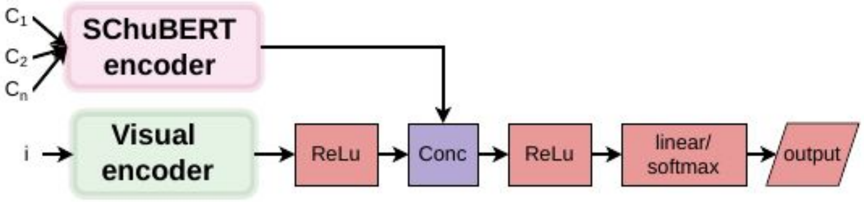
The MultiSChuBERT model



(a) SChuBERT encoder.



(b) Visual encoder.



(f) The MultiSChuBERT model proposed in this work.

Legend
C = Input (512 length text chunk, tokenized for BERT)
CTE = Chunk token embedding
AP = Average-pooling
CE = Chunk embedding
GRU = Gated recurrent unit
i = Input (512x512 Grid of paper pages)
IE = image embedding
GAP = Global Average Pooling
Conc = Concatenation Layer

(c) Legend used symbols.

Three ways to facilitate effective multimodal fusion and further improve results:

1. **Gradual unfreezing**: gradually unfix the weights of the visual submodel during training

2. **Concatenation method:** improve the manner in which textual and visual embeddings are combined

3. Use of **science-domain-specialized** text embeddings in place of $\text{BERT}_{\text{BASE}}$: **SPECTER2.0**

Gradual unfreezing: what?

- Fix all parameters of the visual (sub-)model, except the linear output layer
- Unfreeze one (of ten) inception blocks every two epochs: 22 epochs total
- Train for 18 more epochs with all inception blocks unfrozen
- Learning rate gradually lowered, with set minimum

Gradual unfreezing: Why necessary?

The numbers of total and trainable parameters for the different base models.

Model	#total params	#trainable params	
		frozen	unfrozen
SChuBERT	0.8M	0.8M	
INCEPTION	24.3M	24.3M	
INCEPTION$_{GU}$	24.3M	4K	21.6<
MultiSChuBERT	25.9M	25.9M	
MultiSChuBERT$_{GU}$	25.9M	1.3M	22.9M

- Notice: the INCEPTION submodel has much more trainable parameters than SChuBERT
- Gradual unfreezing avoids INCEPTION from immediately overfitting the data, before even properly fitting the SChuBERT submodel

Concatenation methods: How to create a multimodal embedding?

- Textual (u) and visual (v) embeddings can be combined in different ways.

- Chosen concatenation method impacts results, as shown in the literature for other applications.

- Overview concatenation methods:
  - (u, v): concatenation by taking u and v in one vector.
  - $(|u - v|)$: concatenation by taking the absolute element-wise difference between u and v.
  - $(u * v)$: concatenation by taking the element-wise product of u and v.
  - $(u, v, |u - v|)$: concatenation of u, v and their absolute element-wise difference.
  - $(u, v, u * v)$: concatenation of u, v and their element-wise product.
  - $(u, v, |u - v|, u * v)$: concatenation of u, v, their absolute element-wise difference, and their element-wise product.

Experiments:

- Data & Experimental Setup

Statistics of the used datasets

(a) Data sizes and label types

Dataset	#Documents (train + validation + test)	Labels
AI	4092 (3682 + 205 + 205)	Accept/reject
CL	2638 (2374 + 132 + 132)	Accept/reject
LG	5048 (4543 + 252 + 253)	Accept/reject
ACL-BiblioMetry	30950 (27853 + 1548 + 1549)	Citations

(b) PeerRead accept/reject distribution

Dataset	Train Accept : Reject	Validation Accept : Reject	Test Accept : Reject
AI	10.5% : 89.5%	8.3% : 91.7%	7.8% : 92.2%
CL	24.3% : 75.7%	22.0% : 78.0%	31.1% : 68.9%
LG	36.4% : 63.6%	36.5% : 63.5%	32.0% : 68.0%

Used Hyperparameters

Table 2: Hyperparameters of the proposed models. 'AR, CIT' refers to the accept/reject prediction and citation prediction tasks. 'textual, visual' refers to the textual and visual portions of the joint model.

	SChuBERT	**INCEPTION$_{GU}$**	**MultiSChuBERT**
Vocabulary size	30000	N/A	30000
Optimizer	Adam	Adam	Adam
Learning rate (AR, CIT)	0.0001, 0.001	0.0001, 0.001	0.0001, 0.001
Epochs	40	40	40
Loss function (AR, CIT)	CE, MAE	CE, MAE	CE, MAE
Weight initialization	Xavier normal	Xavier normal	Xavier normal
Dropout rate (textual, visual)	0.3, N/A	N/A, 0.5	0.3, 0.5
GRU hidden size	256	N/A	256
Joint hidden size	N/A	N/A	128
Concatenation method (AR, CIT)	N/A	N/A	(u*v), (u,v,|u-v|)
Train batch size (AI, CL, LG, ACL)	18, 17, 17, 17	18, 17, 17, 17	18, 17, 17, 17
Val batch size (AI, CL, LG, ACL)	14, 16, 13, 15	14, 16, 13, 15	14, 16, 13, 15
Test batch size (AI, CL, LG, ACL)	15, 13, 10, 18	15, 13, 10, 18	15, 13, 10, 18
Word embedding size	768	N/A	768
Image embedding size	N/A	2048	2048

Experiments:

- Results #Citation Prediction

ACLBibliometry main results

(a) System performance metrics and system statistics.

Model	test scores R2↑	MSE↓	MAE↓	validation scores & statistics R2↑	model epoch
Avg Training Label	-0.005 ± 0.000	1.643 ± 0.000	1.028 ± 0.000	-0.001 ± 0.000	–
BiLSTM	0.319 ± 0.013	1.110 ± 0.021	0.824 ± 0.009	–	–
HAN	0.339 ± 0.013	1.080 ± 0.021	0.820 ± 0.009	–	–
SChuBERT*	0.398 ± 0.006	0.985 ± 0.010	0.789 ± 0.005	–	–
CNN	0.118 ± 0.009	1.444 ± 0.013	0.952 ± 0.003	–	–
INCEPTION	0.275 ± 0.029	1.186 ± 0.048	0.852 ± 0.018	0.265 ± 0.016	8.700 ± 3.302
INCEPTION_{GU}	0.332 ± 0.014	1.092 ± 0.023	0.786 ± 0.009	0.329 ± 0.011	38.400 ± 2.413
SChuBERT	0.432 ± 0.010	0.929 ± 0.017	0.765 ± 0.009	0.394 ± 0.005	23.300 ± 8.512
MultiSChuBERT	0.427 ± 0.016	0.937 ± 0.026	0.760 ± 0.009	0.409 ± 0.010	13.700 ± 6.499
$\text{MultiSChuBERT}_{GU}$	**0.454** ± **0.006**	**0.893** ± **0.010**	**0.717** ± **0.006**	**0.436** ± **0.012**	37.600 ± 2.221

ACLBibliometry main results – statistical significance

(b) Statistical significance pairwise system score differences.

	R2						MSE						MAE					
	baseline system						baseline system						baseline system					
System	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Avg Training Label (1)	–	▼	▼	▼	▼	▼	–	▼	▼	▼	▼	▼	–	▼	▼	▼	▼	▼
INCEPTION (2)	▲	–	▼	▼	▼	▼	▲	–	▼	▼	▼	▼	▲	–	▼	▼	▼	▼
INCEPTION$_{GU}$ (3)	▲	▲	–	▼	▼	▼	▲	▲	–	▼	▼	▼	▲	▲	–	▼	▼	▼
SChuBERT (4)	▲	▲	▲	–		▼	▲	▲	▲	–		▼	▲	▲	▲	–		▼
MultiSChuBERT (5)	▲	▲	▲		–	▼	▲	▲	▲		–	▼	▲	▲	▲		–	▼
MultiSChuBERT$_{GU}$ (6)	▲	▲	▲	▲	▲	–	▲	▲	▲	▲	▲	–	▲	▲	▲	▲	▲	–

Statistical significance computed using an in-house adaptation of Multeval, multi-run resampling testing to support classification and regression metrics. ▲ triangle pointing up='*better than other*' with $p < 0.001$
MultiSchubert_GU (15x) > MultiSchubert (9x) = Schubert (9x) ⇒ **GU is needed**

ACLBibliometry concatenation method comparison

(a) System performance metrics and system statistics.

concatenation method	test scores R2↑	MSE↓	MAE↓	validation scores & statistics R2↑	model epoch
(u, v)	0.446 ± 0.010	0.905 ± 0.016	0.723 ± 0.006	0.431 ± 0.005	37.700 ± 1.160
$(\|u - v\|)$	0.449 ± 0.007	0.901 ± 0.012	0.722 ± 0.006	0.429 ± 0.008	38.000 ± 3.266
$(u * v)$	0.443 ± 0.013	0.910 ± 0.021	0.731 ± 0.016	0.431 ± 0.006	35.400 ± 8.708
$(\|u - v\|, u * v)$	0.442 ± 0.011	0.912 ± 0.019	0.726 ± 0.008	0.424 ± 0.006	38.200 ± 2.440
$(u, v, u * v)$	0.445 ± 0.010	0.908 ± 0.017	0.725 ± 0.008	0.433 ± 0.007	37.900 ± 2.424
$(u, v, \|u - v\|)$	0.450 ± 0.005	0.900 ± 0.009	0.721 ± 0.006	**0.436** ± **0.009**	38.100 ± 2.998
$(u, v, \|u - v\|, u * v)$	**0.454** ± **0.006**	**0.893** ± **0.010**	**0.717** ± **0.006**	**0.436** ± **0.012**	37.600 ± 2.221

ACLBibliometry concatenation method comparison – statistical significance

(b) Statistical significance pairwise system score differences.

	R2 baseline system							MSE baseline system							MAE baseline system						
System	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7
(u, v) (1)	–						▽	–						▽	–		▲				▽
$(\|u - v\|)$ (2)		–		△					–		△					–	▲	△			
$(u * v)$ (3)			–				▼			–				▼	▼	▼	–		▽	▼	▼
$(\|u - v\|, u * v)$ (4)		▽		–		▽	▼		▽		–		▽	▼		▽		–		▽	▼
$(u, v, u * v)$ (5)					–		▼					–		▼			△		–		▼
$(u, v, \|u - v\|)$ (6)				△		–					△		–				▲	△		–	
$(u, v, \|u - v\|, u * v)$ (7)	△		▲	▲	▲		–	△		▲	▲	▲		–	△		▲	▲	▲		–

Experiments:

- Results Accept/Reject Prediction

Main Results PeerRead cs.AI

(a) System performance metrics and system statistics.

Model	test scores Accuracy↑	ROC AUC↑	F_1-score↑	validation scores & statistics Accuracy↑	model epoch
Maj Training Label	92.2 ± 0.00%	0.500 ± 0.000	0.000 ± 0.00	91.7 ± 0.00%	–
CNN	92.2 ± 0.00%	–	–	–	–
INCEPTION	92.3 ± 1.36%	0.834 ± 0.045	0.392 ± 0.069	92.6 ± 1.20%	1.800 ± 0.789
INCEPTION$_{GU}$	93.0 ± 0.87%	0.826 ± 0.031	0.441 ± 0.092	92.5 ± 0.95%	31.500 ± 0.707
SChuBERT	93.5 ± 0.52%	0.912 ± 0.012	0.461 ± 0.080	91.9 ± 0.35%	19.200 ± 5.534
MultiSChuBERT	92.7 ± 0.43%	0.830 ± 0.027	0.363 ± 0.160	92.7 ± 1.15%	1.900 ± 0.876
MultiSChuBERT$_{GU}$	**93.6** ± **1.02%**	**0.913** ± **0.020**	**0.551** ± **0.087**	**93.1** ± **0.94%**	26.000 ± 6.342

(b) Statistical significance pairwise system score differences.

System	Accuracy baseline system 1	2	3	4	5	6	ROC AUC baseline system 1	2	3	4	5	6	F_1-Score baseline system 1	2	3	4	5	6
Maj Training Label (1)	–			▼		▽	–	▼	▼	▼	▼	▼	–	▼	▼	▼	▼	▼
INCEPTION (2)		–		▽		▼	▲	–		▼		▼	▲	–				▼
INCEPTION$_{GU}$ (3)			–				▲		–	▼		▼	▲		–			▼
SChuBERT (4)	▲	△		–			▲	▲	▲	–	▲		▲			–		▼
MultiSChuBERT (5)					–		▲			▼	–	▼	▲				–	▼
MultiSChuBERT$_{GU}$ (6)	△	▲				–	▲	▲	▲		▲	–	▲	▲	▲	▲	▲	–

Main Results PeerRead cs.CL

(a) System performance metrics and system statistics.

Model	test scores			validation scores & statistics	
	Accuracy↑	ROC AUC↑	F_1-score↑	Accuracy↑	model epoch
Maj Training Label	68.9 ± 0.00%	0.500 ± 0.000	0.000 ± 0.00	78.0 ± 0.00%	–
CNN	68.9 ± 0.00%	–	–	–	–
INCEPTION	80.8 ± 1.93%	0.871 ± 0.020	0.667 ± 0.072	82.3 ± 2.08%	1.200 ± 0.422
INCEPTION$_{GU}$	80.2 ± 3.38%	0.869 ± 0.020	0.661 ± 0.096	**83.9 ± 2.16%**	32.000 ± 1.633
SChuBERT	82.4 ± 2.14%	**0.920 ± 0.004**	0.640 ± 0.070	78.6 ± 1.14%	9.800 ± 2.860
MultiSChuBERT	83.3 ± 3.04%	0.893 ± 0.023	0.708 ± 0.099	**83.9 ± 1.56%**	2.300 ± 0.823
MultiSChuBERT$_{GU}$	**85.2 ± 1.20%**	**0.920 ± 0.015**	**0.740 ± 0.032**	82.8 ± 2.76%	24.100 ± 11.220

(b) Statistical significance pairwise system score differences.

System	Accuracy baseline system						ROC AUC baseline system						F_1-score baseline system					
	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Maj Training Label (1)	–	▼	▼	▼	▼	▼	–	▼	▼	▼	▼	▼	–	▼	▼	▼	▼	▼
INCEPTION(2)	▲	–			▼	▼	▲	–		▼	▼	▼	▲	–			▼	▼
INCEPTION$_{GU}$ (3)	▲		–		▼	▼	▲		–	▼	▼	▼	▲		–		▼	▼
SChuBERT(4)	▲			–		▼	▲	▲	▲	–	▲		▲			–	▼	▼
MultiSChuBERT (5)	▲	▲	▲		–	▽	▲	▲	▲	▼	–	▼	▲	▲	▲	▲	–	
MultiSChuBERT$_{GU}$ (6)	▲	▲	▲	▲	△	–	▲	▲	▲		▲	–	▲	▲	▲	▲		–

Main Results PeerRead cs.LG

(a) System performance metrics and system statistics.

Model	test scores			validation scores & statistics	
	Accuracy↑	ROC AUC↑	F_1-score↑	Accuracy↑	model epoch
Maj Training Label	68.0 ± 0.00%	0.500 ± 0.000	0.000 ± 0.00	63.5 ± 0.00%	–
CNN	65.7 ± 2.79%	–	–	–	–
INCEPTION	82.2 ± 1.42%	0.904 ± 0.011	0.729 ± 0.026	83.3 ± 2.52%	2.500 ± 2.121
INCEPTION$_{GU}$	83.6 ± 1.86%	0.904 ± 0.013	0.752 ± 0.023	84.1 ± 1.45%	31.600 ± 0.516
SChuBERT	80.3 ± 1.37%	0.880 ± 0.006	0.723 ± 0.014	76.9 ± 0.56%	13.000 ± 3.091
MultiSChuBERT	83.4 ± 1.65%	0.921 ± 0.012	0.750 ± 0.017	**84.9** ± **1.80%**	1.900 ± 0.876
MultiSChuBERT$_{GU}$	**84.9** ± **1.40%**	**0.931** ± **0.007**	**0.781** ± **0.016**	83.5 ± 1.56%	32.300 ± 1.947

(b) Statistical significance pairwise system score differences.

System	Accuracy baseline system						ROC AUC baseline system						F_1-score baseline system					
	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Maj Training Label (1)	–	▼	▼	▼	▼	▼	–	▼	▼	▼	▼	▼	–	▼	▼	▼	▼	▼
INCEPTION (2)	▲	–	▽	△	▽	▼	▲	–		△	▼	▼	▲	–	▼		▽	▼
INCEPTION$_{GU}$ (3)	▲	△	–	▲		▽	▲		–	▲	▼	▼	▲	▲	–	△		▼
SChuBERT (4)	▲	▽	▼	–	▼	▼	▲	▽	▼	–	▼	▼	▲		▽	–	▽	▼
MultiSChuBERT (5)	▲	△		▲	–	▽	▲	▲	▲	▲	–	▼	▲	△		△	–	▼
MultiSChuBERT$_{GU}$ (6)	▲	▲	△	▲	△	–	▲	▲	▲	▲	▲	–	▲	▲	▲	▲	▲	–

Concatenation methods comparison PeerRead cs.AI

(a) System performance metrics and system statistics.

concatenation method	test scores Accuracy	ROC↑ AUC↑	F_1-score↑	validation scores & statistics Accuracy↑	model epoch
(u, v)	93.9 ± 0.70%	**0.922 ± 0.012**	**0.578 ± 0.055**	92.9 ± 0.70%	26.800 ± 2.741
$(\|u - v\|)$	93.7 ± 0.61%	0.912 ± 0.009	0.506 ± 0.076	92.4 ± 0.78%	17.500 ± 6.078
$(u * v)$	93.5 ± 0.69%	0.894 ± 0.008	0.481 ± 0.055	92.2 ± 0.75%	16.300 ± 4.473
$(\|u - v\|, u * v)$	**94.0 ± 0.52%**	0.907 ± 0.015	0.533 ± 0.086	92.4 ± 0.66%	18.000 ± 7.916
$(u, v, u * v)$	93.7 ± 0.78%	0.908 ± 0.012	0.484 ± 0.122	92.1 ± 0.71%	16.500 ± 5.759
$(u, v, \|u - v\|)$	93.6 ± 1.02%	0.913 ± 0.020	0.551 ± 0.087	**93.1 ± 0.94%**	26.000 ± 6.342
$(u, v, \|u - v\|, u * v)$	93.8 ± 1.08%	0.909 ± 0.016	0.488 ± 0.165	92.5 ± 0.870%	16.200 ± 7.671

(b) Statistical significance pairwise system score differences.

System	Accuracy baseline system 1	2	3	4	5	6	7	ROC AUC baseline system 1	2	3	4	5	6	7	F_1-score baseline system 1	2	3	4	5	6	7
(u, v) (1)	–							–		▲	▲	△		△	–	△	▲		▲		▲
$(\|u - v\|)$ (2)		–							–	△					▽	–					
$(u * v)$ (3)			–					▼	▽	–					▼		–			▽	
$(\|u - v\|, u * v)$ (4)				–				▼			–							–			
$(u, v, u * v)$ (5)					–			▽				–			▼				–	▽	
$(u, v, \|u - v\|)$ (6)						–							–				△		△	–	
$(u, v, \|u - v\|, u * v)$ (7)							–	▽						–	▼						–

Concatenation methods comparison PeerRead cs.CL

(a) System performance metrics and system statistics.

concatenation method	test scores Accuracy↑	ROC AUC↑	F_1-score↑	validation scores & statistics Accuracy↑	model epoch
(u, v)	84.9 ± 2.03%	0.917 ± 0.005	0.733 ± 0.053	81.8 ± 2.74%	22.400 ± 11.918
$(\|u - v\|)$	85.2 ± 1.20%	0.920 ± 0.015	0.740 ± 0.032	**82.8 ± 2.76%**	24.100 ± 11.220
$(u * v)$	85.5 ± 1.37%	**0.921 ± 0.007**	0.742 ± 0.037	78.9 ± 0.54%	8.000 ± 1.247
$(\|u - v\|, u * v)$	**85.8 ± 1.24%**	0.918 ± 0.014	**0.758 ± 0.023**	80.2 ± 1.85%	17.900 ± 11.949
$(u, v, u * v)$	85.4 ± 1.96%	0.918 ± 0.008	0.749 ± 0.048	79.8 ± 2.97%	12.700 ± 10.166
$(u, v, \|u - v\|)$	**85.8 ± 2.40%**	0.919 ± 0.010	0.755 ± 0.052	81.8 ± 2.97%	23.200 ± 11.708
$(u, v, \|u - v\|, u * v)$	**85.8 ± 1.88%**	**0.921 ± 0.006**	0.747 ± 0.050	80.5 ± 2.81%	16.000 ± 11.235

(b) Statistical significance pairwise system score differences.

System	Accuracy baseline system 1	2	3	4	5	6	7	ROC AUC baseline system 1	2	3	4	5	6	7	F_1-score baseline system 1	2	3	4	5	6	7
(u, v) (1)	–							–		$\triangledown$					–						
$(\|u - v\|)$ (2)		–							–							–					
$(u * v)$ (3)			–					$\triangle$		–	$\triangle$	$\triangle$					–				
$(\|u - v\|, u * v)$ (4)				–						$\triangledown$	–							–			
$(u, v, u * v)$ (5)					–					$\triangledown$		–							–		
$(u, v, \|u - v\|)$ (6)						–							–							–	
$(u, v, \|u - v\|, u * v)$ (7)							–							–							–

Concatenation methods comparison PeerRead cs.LG

(a) System performance metrics and system statistics.

concatenation method	test scores Accuracy↑	ROC AUC↑	F_1-score↑	validation scores & statistics Accuracy↑	model epoch
(u, v)	84.2 ± 2.02%	0.924 ± 0.009	0.762 ± 0.028	83.2 ± 1.91%	32.500 ± 0.972
$(\|u - v\|)$	**84.9** ± **1.40%**	**0.931** ± **0.007**	**0.781** ± **0.016**	**83.5** ± **1.56%**	32.300 ± 1.947
$(u * v)$	81.8 ± 1.87%	0.908 ± 0.007	0.725 ± 0.033	81.4 ± 2.36%	26.800 ± 6.070
$(\|u - v\|, u * v)$	82.6 ± 1.68%	0.912 ± 0.010	0.750 ± 0.020	81.2 ± 3.11%	26.900 ± 6.008
$(u, v, u * v)$	83.6 ± 1.88%	0.918 ± 0.009	0.760 ± 0.020	82.0 ± 3.36%	27.700 ± 7.931
$(u, v, \|u - v\|)$	84.2 ± 1.59%	0.921 ± 0.013	0.767 ± 0.027	82.7 ± 2.73%	30.400 ± 4.624
$(u, v, \|u - v\|, u * v)$	82.5 ± 1.15%	0.912 ± 0.009	0.750 ± 0.016	81.7 ± 1.96%	28.300 ± 6.550

(b) Statistical significance pairwise system score differences.

System	Accuracy baseline system 1	2	3	4	5	6	7	ROC AUC baseline system 1	2	3	4	5	6	7	F_1-score baseline system 1	2	3	4	5	6	7
(u, v) (1)	–		▲	▲			▲	–	▽	▲	▲	▲		▲	–	▼	▲				
$(\|u - v\|))$ (2)		–	▲	▲	▲		▲	△	–	▲	▲	▲	▲	▲	▲	–	▲	▲	▲	△	▲
$(u * v))$ (3)	▼	▼	–		▼	▼		▼	▼	–			▼		▼	▼	–	▼	▼	▼	▼
$(\|u - v\|, u * v)$ (4)	▼	▼		–		▼		▼	▼		–		▼			▼	▲	–		▽	
$(u, v, u * v)$ (5)		▼	▲		–		△	▼	▼			–	▽			▼	▲		–		
$(u, v, \|u - v\|)$ (6)			▲	▲		–	▲		▼	▲	▲	△	–	▲		▽	▲	△		–	△
$(u, v, \|u - v\|, u * v)$ (7)	▼	▼			▽	▼	–	▼	▼				▼	–		▼	▲			▽	–

Experiments:

- Adding domain-specialized embeddings

Science-domain-specialized text embedding

Model	test scores			validation scores & statistics	
	R2↑	MSE↓	MAE↓	R2↑	model epoch
Avg Training Label	-0.005 ± 0.000	1.643 ± 0.000	1.028 ± 0.000	-0.001 ± 0.000	–
SChuBERT	0.432 ± 0.010	0.929 ± 0.017	0.765 ± 0.009	0.394 ± 0.005	23.300 ± 8.512
$\textbf{MultiSChuBERT}_{\textbf{GU}}$	0.454 ± 0.006	0.893 ± 0.010	0.717 ± 0.006	0.436 ± 0.012	37.600 ± 2.221
$\textbf{SChuBERT}_{\textbf{SCIBERT}}$	0.467 ± 0.014	0.871 ± 0.022	0.743 ± 0.011	0.439 ± 0.005	15.600 ± 3.658
$\textbf{SChuBERT}_{\textbf{SCINCL}}$	0.460 ± 0.008	0.883 ± 0.013	0.751 ± 0.006	0.447 ± 0.006	33.300 ± 5.478
$\textbf{SChuBERT}_{\textbf{SR}}$	0.447 ± 0.013	0.904 ± 0.021	0.754 ± 0.010	0.440 ± 0.009	24.700 ± 10.144
$\textbf{SChuBERT}_{\textbf{SR2.0}}$	0.474 ± 0.013	0.860 ± 0.021	0.736 ± 0.009	0.460 ± 0.003	14.400 ± 6.186
$\textbf{Multi-SChuBERT}_{\textbf{GU\_SR2.0}}$	**0.503** ± **0.011**	**0.813** ± **0.018**	**0.693** ± **0.016**	**0.484** ± **0.009**	32.300 ± 11.898

Unfortunately, no control for *label leakage* in these experiments: training data of the domain-specialized embedding models expected to overlap with ACL Bibliometry data.

Fixing the label-leakage problem

- SPECTER 2.0 training data is downloadable
- Obtain a list of paper titles used in training and validation examples
- Lowercase and remove spaces to maximize recall of matching papers
- Filtered about 40% of the ACL Bibliometry data this way, because of overlap with the SPECTER2.0 training/validation data
  - Produce filtered ACL Bibliometry sets without overlap with SPECTER2.0 training/validation data

SPECTER2.0 results – filtered testset

(a) System performance metrics and system statistics.

Model	test scores R2↑	test scores MSE↓	test scores MAE↓	validation scores & statistics R2↑	validation scores & statistics model epoch
Avg Training Label	-0.130 ± 0.000	1.181 ± 0.000	0.910 ± 0.000	-0.001 ± 0.000	–
SChuBERT	0.267 ± 0.015	0.766 ± 0.015	0.693 ± 0.009	0.394 ± 0.005	23.300 ± 8.512
MultiSChuBERT$_{GU}$	0.302 ± 0.017	0.730 ± 0.018	0.652 ± 0.006	0.436 ± 0.012	37.600 ± 2.221
SChuBERT$_{SR2.0}$	0.319 ± 0.016	0.711 ± 0.017	0.675 ± 0.007	0.460 ± 0.003	14.400 ± 6.186
Multi-SChuBERT$_{GU_SR2.0}$	**0.335** ± **0.020**	**0.695** ± **0.021**	**0.643** ± **0.017**	**0.484** ± **0.009**	32.300 ± 11.898

(b) Statistical significance pairwise system score differences.

System	R2 baseline system 1	2	3	4	5	MSE baseline system 1	2	3	4	5	MAE baseline system 1	2	3	4	5
Avg Training Label (1)	–	▼	▼	▼	▼	–	▼	▼	▼	▼	–	▼	▼	▼	▼
SChuBERT (2)	▲	–	▼	▼	▼	▲	–	▼	▼	▼	▲	–	▼	▼	▼
MultiSChuBERT$_{GU}$ (3)	▲	▲	–	▽	▼	▲	▲	–	▽	▼	▲	▲	–	▲	▽
SChuBERT$_{SR2.0}$ (4)	▲	▲	△	–	▽	▲	▲	△	–	▽	▲	▲	▼	–	▼
Multi-SChuBERT$_{GU_SR2.0}$ (5)	▲	▲	▲	△	–	▲	▲	▲	△	–	▲	▲	△	▲	–

Note: Negative R2 score for Avg Training Label baseline method!

Understanding the performance drop across systems

- Label statistics coherent within datasets (ACL, filtered ACL), but different across normal and filtered ACL data.

- Mismatched label distribution between {training, validation} and {test} data explains performance drop.

Label statistics of the original and filtered ACL datasets.

(a) ACL data

subset	train	val	test
#examples	27852	1547	1548
avg label	1.729 ± 1.191	1.759 ± 1.216	1.819 ± 1.279

(b) Filtered ACL data

subset	train	val	test
#examples	16730	957	926
avg label	1.330 ± 0.978	1.350 ± 0.991	1.360 ± 1.023

Solution

- Filter all data, not just the test set
- This restores the coherence between the train, validation and test data, at the cost of smaller training data.
    - Resulting training data size ± 60% of original

SPECTER2.0 results – filtered all data

- Note: improved results despite smaller training data!

(a) System performance metrics and system statistics.

Model	test scores R2↑	test scores MSE↓	test scores MAE↓	validation scores & statistics R2↑	validation scores & statistics model epoch
Avg Training Label	-0.001 ± 0.000	1.046 ± 0.000	0.861 ± 0.000	-0.000 ± 0.000	–
SChuBERT	0.305 ± 0.008	0.726 ± 0.008	0.682 ± 0.004	0.266 ± 0.004	14.700 ± 4.270
MultiSChuBERT$_{GU}$	0.332 ± 0.024	0.698 ± 0.025	0.647 ± 0.018	0.296 ± 0.017	30.400 ± 8.733
SChuBERT$_{SR2.0}$	0.333 ± 0.011	0.697 ± 0.011	0.672 ± 0.005	0.325 ± 0.005	18.100 ± 5.238
Multi-SChuBERT$_{GU_SR2.0}$	**0.351** ± **0.026**	**0.679** ± **0.027**	**0.646** ± **0.026**	**0.336** ± **0.009**	23.200 ± 13.831

(b) Statistical significance pairwise system score differences.

System	R2 baseline system 1	2	3	4	5	MSE baseline system 1	2	3	4	5	MAE baseline system 1	2	3	4	5
Avg Training Label (1)	–	▼	▼	▼	▼	–	▼	▼	▼	▼	–	▼	▼	▼	▼
SChuBERT (2)	▲	–	▼	▼	▼	▲	–	▼	▼	▼	▲	–	▼	▼	▼
MultiSChuBERT$_{GU}$ (3)	▲	▲	–		▼	▲	▲	–		▼	▲	▲	–	▲	
SChuBERT$_{SR2.0}$(4)	▲	▲		–	▼	▲	▲		–	▼	▲	▲	▼	–	▼
SChuBERT$_{GU_SR2.0}$ (5)	▲	▲	▲	▲	–	▲	▲	▲	▲	–	▲	▲		▲	–

Conclusions

- All SChuBERT-based methods outperform the baseline models
- $\text{MultiSChuBERT}_{\text{GU}}$ significantly outperforms SChuBERT, MultiSChuBERT and is the best model overall
  - Gradual Unfreezing helps in mitigating parameter imbalance
- The concatenation method makes a difference, but there are multiple alternatives that perform the same (no statistically significant difference)
- The SPECTER 2.0 domain specialized text embedding further improves performance (statistically significant and while avoiding label leakage)