

De Onderzoeksgroep
Wiskunde & Data Science

nodigt U graag uit op de openbare verdediging van het proefschrift van

Willy Carlos Tchuitcheu

ter behaling van de graad van Doctor in de wetenschappen

Titel van het proefschrift:

**Representation Learning for Table Understanding in
Intelligent Document Processing**

Promotor:

Prof. dr. Ann Dooms (VUB)

Co-promotor:

Prof. dr. Tan Lu (VUB)

De verdediging heeft plaats op

vrijdag 5 december 2025 om 15u

Campus Etterbeek VUB, Pleinlaan 2, Elsene,
Learning Theatre, Learning & Innovation Center
(LIC)?

De verdediging is ook te volgen via een

livestream: Stuur een e-mail naar

willy.carlos.tchuitcheu@vub.be om de link te ontvangen.

Samenstelling van de jury

Prof. dr. Jan De Beule (VUB, voorzitter)

Prof. dr. Ann Nowé (VUB)

Prof. dr. Paul Van Eecke (VUB)

Prof. dr. Lynn Houthuys (VUB)

Prof. dr. Apostolos Antonacopoulos (University of
Salford, VK)

Dr. Jeroen Jordens (Flanders Make)

Curriculum vitae

Willy Carlos Tchuitcheu behaalde in 2019 zijn masterdiploma in de wiskundige wetenschappen aan het African Institute for Mathematical Sciences (AIMS) Rwanda. Tussen 2015 en 2020 deed hij drie jaar ervaring op als onderzoeksingenieur bij Camertronix, een onderzoekscentrum in Kameroen. In 2021 startte hij zijn doctoraat aan de vakgroep Wiskunde en Data Science van de VUB. Zijn doctoraatsonderzoek heeft geleid tot drie artikelen als eerste auteur in internationale tijdschriften (waarvan één reeds gepubliceerd), één octrooiaanvraag en een Best Poster Award tijdens de Flanders AI Research (FAIR) Day 2021. Daarnaast is Willy mede-auteur van twee bijkomende tijdschriftartikelen die niet verbonden zijn aan dit proefschrift, waaronder één als eerste auteur.

Abstract van het doctoraatsonderzoek

Het volume aan zakelijke en wetenschappelijke documenten neemt snel toe, wat leidt tot een groeiende vraag naar methoden voor Intelligent Document Processing (IDP). Deze documenten bevatten een mix van datatypes, zoals tekst, tabellen en figuren, elk met hun eigen structuur en semantiek. Tabellen verschillen daarbij duidelijk van tekst: ze zijn tweedimensionaal, hun betekenis volgt niet de grammaticale regels van natuurlijke taal, en ze zijn vaak permutatie-invariant, omdat het verwisselen van rijen of kolommen de interpretatie niet hoeft te veranderen. Een groot deel van de cruciale informatie in documenten staat in tabellen (bv. spreadsheets, rapporten, facturen, datasheets en wetenschappelijke artikelen), die mensen op een natuurlijke manier begrijpen door elke cel te koppelen aan de bijbehorende kop. Voor dit menselijk vermogen om tabellen te lezen introduceren we de term *Table Understanding (TU) principle*.

Het succes van LLM's in tekstbegrip steunt in grote mate op rijke vectorvoorstellingen, zogenaamde *word embeddings*. Betekenisvolle *table embeddings* zijn daarentegen nog onderontwikkeld. In de praktijk lineariseren LLM's tabellen tot lange tekststrings, waarbij ze de tabelstructuur verwaarlozen. Hierdoor gaan lay-outinformatie en kolom- en rijkoppen verloren, wat tot een prestatiedaling leidt, zeker omdat tabellen eigenschappen vereisen zoals permutatie-invariantie, die in contrast staan met het volgordegevoelige karakter van de tekst waarop LLM's zijn getraind.

Om dit aan te pakken, introduceren we een structuurbewuste methode voor representatieleren op tabellen. We modelleren eerst elke tabel als een *Heterogeneous Graph Table (HGT)* die de tweedimensionale lay-out, celtypes en relaties tussen koppen bewaart. Op basis van de HGT leren we contextbewuste celembeddings die cellen expliciet koppelen aan hun koppen, in lijn met het TU-principe. Daarnaast dragen we taalkundige kennis uit LLM's over naar deze embeddings zonder de tabel te flattenen, waardoor we oppervlaktetekst en structuur combineren. Ten slotte onderzoeken we permutatie-invariantie op een systematische manier door rijen en kolommen te herschikken en de stabiliteit van de table embeddings te evalueren.

Onze methode toont competitieve prestaties in twee hoofdcategorieën van TU-taken: *Column Type Annotation (CTA)*, dat onder de semantische categorie valt, en *Table Question Answering (TQA)*, dat onder de redeneercategorie valt. Bovendien vermindert onze aanpak de complexiteit en verhoogt ze de robuustheid tegen permutaties van tabellen. Deze resultaten wijzen erop dat het gebruik van structuurbewuste table embeddings, gestoeld op het TU-principe, een praktische oplossing kan bieden voor betrouwbare, low-latency informatie-extractie binnen het bredere IDP-domein.