UNIVERSITAS
GADJAH MADA
Model Paralel CNN dan BLIP-2 Untuk Identifikasi Jenis Subject pertanyaan Matematika Berbahasa Inggris
Muhammad Ashabul Kahfi, Prof. Dr. Drs. Azhari, MT
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

INTISARI

MODEL PARALEL CNN DAN BLIP-2 UNTUK IDENTIFIKASI JENIS SUBJEK PERTANYAAN MATEMATIKA BERBAHASA INGGRIS

Oleh

MUHAMMAD ASHABUL KAHFI
24/537433/PPA/06796

Penelitian ini membahas klasifikasi subjek soal matematika berbasis data multimodal yang terdiri atas teks pertanyaan dan citra soal dari dataset MathVision. Tantangan utama dalam penelitian ini adalah kebutuhan model untuk memahami hubungan antara informasi visual dan teks, serta adanya ketidakseimbangan kelas. Dataset awal terdiri atas 3.040 sampel dalam 16 kelas, kemudian disaring menjadi 2.718 sampel dalam 11 kelas berdasarkan batas minimum 100 sampel untuk setiap kelas.

Penelitian ini menerapkan tujuh skenario eksperimen, yaitu BLIP-2 *baseline*, Parallel CNN-only, Parallel CNN + BLIP-2, CLIP *baseline*, Parallel CNN + CLIP, Single CNN 7×7 + BLIP-2, dan Single CNN 5×5 + BLIP-2. Selain itu, metode SMOTE diterapkan pada ruang fitur untuk eksperimen E1 hingga E3, khususnya pada data pelatihan. Evaluasi dilakukan menggunakan metrik *accuracy*, macro-F1, weighted-F1, dan *confusion matrix*.

Hasil eksperimen menunjukkan bahwa Parallel CNN + BLIP-2 mampu meningkatkan performa dibandingkan BLIP-2 *baseline*, dengan nilai *accuracy* sebesar 0,6618, macro-F1 sebesar 0,6150, dan weighted-F1 sebesar 0,6528. Model Single CNN 5×5 + BLIP-2 memperoleh *accuracy* tertinggi sebesar 0,6740 dan weighted-F1 tertinggi sebesar 0,6614. Sementara itu, penerapan SMOTE pada fitur gabungan Parallel CNN + BLIP-2 menghasilkan macro-F1 tertinggi sebesar 0,6198. Analisis kesalahan menunjukkan bahwa model masih mengalami kesulitan dalam membedakan kelas-kelas yang memiliki kemiripan visual dan konseptual, terutama pada subkategori geometri metrik. Oleh karena itu, integrasi fitur visual CNN dan representasi multimodal dapat meningkatkan performa klasifikasi, tetapi pengembangan lebih lanjut pada representasi visual-teks dan penanganan ketidakseimbangan kelas masih diperlukan.

Kata kunci: Klasifikasi, multimodal, BLIP-2, CLIP, Parallel CNN, SMOTE, MathVision.

UNIVERSITAS
GADJAH MADA
Model Paralel CNN dan BLIP-2 Untuk Identifikasi Jenis Subject pertanyaan Matematika Berbahasa Inggris
Muhammad Ashabul Kahfi, Prof. Dr. Drs. Azhari, MT
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

ABSTRACT

A Parallel CNN and BLIP-2 Model for Subject Category Identification in English Mathematical Questions

By

MUHAMMAD ASHABUL KAHFI
24/537433/PPA/06796

This study discusses subject classification for mathematical questions based on multimodal data consisting of question text and problem images from the MathVision dataset. The main challenges in this task are the need for a model to understand visual-text relationships and the presence of class imbalance. The initial dataset consists of 3,040 samples with 16 classes, which are then filtered into 2,718 samples with 11 classes based on a minimum threshold of 100 samples per class.

This study implements seven experimental scenarios, namely BLIP-2 baseline, Parallel CNN-only, Parallel CNN + BLIP-2, CLIP baseline, Parallel CNN + CLIP, Single CNN 7×7 + BLIP-2, and Single CNN 5×5 + BLIP-2. In addition, SMOTE is applied to the feature space for E1 to E3, specifically on the training data. Evaluation is conducted using accuracy, macro-F1, weighted-F1, and confusion matrix.

The experimental results show that Parallel CNN + BLIP-2 improves performance compared with the BLIP-2 baseline, achieving an accuracy of 0.6618, macro-F1 of 0.6150, and weighted-F1 of 0.6528. The Single CNN 5×5 + BLIP-2 model achieves the highest accuracy of 0.6740 and the highest weighted-F1 of 0.6614, while SMOTE on the combined features of Parallel CNN + BLIP-2 produces the highest macro-F1 of 0.6198. Error analysis indicates that the model still struggles to distinguish classes with visual and conceptual similarities, especially within the metric geometry subcategories. Therefore, the integration of CNN visual features and multimodal representations can improve classification performance, but further enhancement of visual-text representation and class imbalance handling is still required.

Keywords: Classification, multimodal, BLIP-2, CLIP, Parallel CNN, SMOTE, MathVision.