MATA KULIAH KEAMANAN JARINGAN

SPAM FILTERING METODE DEEP LEARNING

Oleh:

KELOMPOK 1

Putu Widyantara Artanta Wibawa	**(2108561005)**
Kenny Belle Lesmana	**(2108561015)**
I Made Ari Madya Santosa	**(2108561020)**
I Nyoman Dheva Surya	**(2108561025)**

PROGRAM STUDI INFORMATIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS UDAYANA

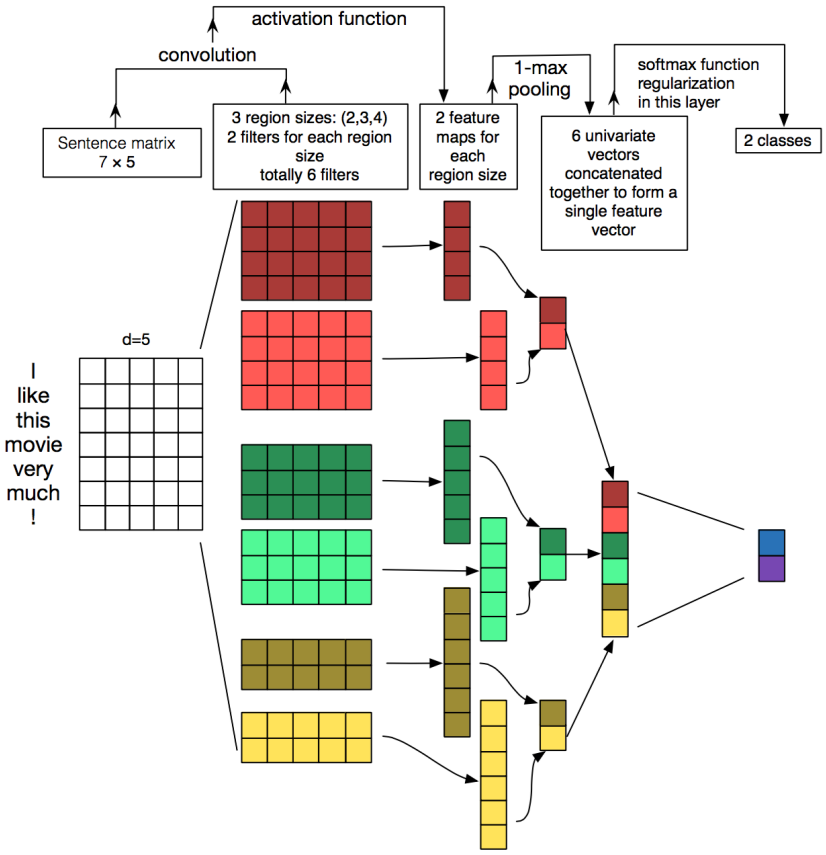
2023

Deep Learning Algorithms (CNN)

CNN (Convolutional Neural Network) adalah algoritma deep learning yang biasanya digunakan dalam melakukan klasifikasi gamba. Akan tetapi, baru-baru ini, CNN juga mulai diterapkan pada masalah NLP dan mendapatkan beberapa hasil yang cukup baik. Berdasarkan asal katanya, CNN terdiri dari Convolutional dan Neural Network. Convolutional secara sederhana dapat diartikan sebagai sebuah jendela bergerak atau filter (kernel) yang diaplikasikan kepada matriks. Sementara Neural Network atau jaringan syaraf tiruan adalah metode kecerdasan buatan yang mengadopsi cara kerja otak manusia, dengan menggunakan neuron yang saling terhubung antara satu lapisan (layer) dengan lapisan lainnya.

CNN pada dasarnya terdiri dari beberapa lapisan convolutional dengan dengan fungsi aktivasi nonlinier seperti ReLU diterapkan pada hasilnya. Pada jaringan saraf tiruan, setiap neuron masukan dihubungkan dengan setiap neuron keluaran pada lapisan berikutnya. Pada CNN, digunakan lapisan convolutional pada masukan untuk menghitung keluaran. Selain itu, setiap lapisan menerapkan filter yang berbeda, biasanya ratusan atau ribuan yang nantinya hasilnya digabungkan. Selama fase pelatihan, CNN akan mempelajari nilai dari setiap filter sesuai dengan tujuan yang diinginkan.

Pada permasalahan NLP, kalimat atau dokumen akan direpresentasikan sebagai matriks. Setiap baris matriks sesuai dengan satu token, biasanya sebuah kata, tetapi bisa juga berupa karakter. Dapat diartikan bahwa setiap baris adalah vektor yang merepresentasikan sebuah kata. Biasanya, vektor-vektor ini direpresentasikan dengan word embedding seperti word2vec atau GloVe. Untuk sebuah kalimat dengan 10 kata menggunakan embedding 100 dimensi, akan dihasilkan sebuah matriks 10x100 sebagai input. Matriks tersebut dapat diandaikan sebagai “gambar” dalam hal ini. Perbedaan dalam kasus NLP adalah biasanya filter yang digunakan bergeser pada seluruh baris matriks (kata-kata). Dengan demikian, “lebar” filter sama dengan lebar matriks input. Adapun model dari CNN yang akan dibangun terlihat pada gambar berikut:



Gambar di atas adalah ilustrasi dari arsitektur CNN untuk melakukan proses klasifikasi teks dalam hal ini yang akan dibuat adalah klasifikasi spam. Masukan yang diberikan adalah sebuah kalimat dengan panjang 7 kata dengan embedding 5 dimensi. Dari masukan tersebut akan diaplikasikan 6 buah filter, masing-masing dengan ukuran 2, 3, dan 4 sebanyak 2 buah. Setiap filter akan melakukan konvolusi pada matriks input sehingga akan dihasilkan masing-masing sebanyak 2 feature map. Kemudian pada setiap feature map akan dilakukan max pooling, yaitu mencari nilai tertinggi. Selanjutnya hasil dari max pooling akan digabungkan (dense) menjadi vektor fitur untuk melalui tahap terakhir untuk melakukan klasifikasi.

Dalam melakukan implementasi, akan digunakan bahasa pemrograman Python serta deep learning library Keras untuk mempermudah pembuatan arsitektur CNN ini. Dataset

yang digunakan berasal dari UCI Machine Learning yang didapatkan melalui platform Kaggle. Implementasi arsitektur CNN dibuat sangat mirip dengan gambar yang telah dijelaskan sebelumnya. Perbedaannya hanyalah pada panjang maksimal kata dan dimensi yang digunakan, hal ini disebabkan karena adanya penyesuaian terhadap rata-rata panjang teks dari dataset digunakan.

Penggunaan CNN dalam melakukan spam filtering menghasilkan model dengan akurasi yang sangat baik. Dalam kasus ini dengan dataset yang digunakan, arsitektur CNN berhasil memperoleh akurasi diatas 98%. Akan tetapi dalam membangun model deep learning diperlukan cukup banyak sumber daya komputasi serta jumlah dataset yang cukup banyak. Berdasarkan dataset yang digunakan, dapat diketahui pula bahwa CNN mampu melakukan klasifikasi dengan sangat baik walaupun pada jumlah data yang tidak seimbang.