Alternatif Topik Tesis : Penerapan Representasi Kontekstual pada Pemrosesan Data Genetik

Masalah Umum

Proses produksi sekuens DNA dari mesin *sequencer* semakin efisien sehingga data sekuens genetik (DNA/RNA) semakin banyak sehingga dibutuhkan metode komputasi yang mampu memproses data tersebut. Metode yang sedang populer digunakan adalah *deep learning*. Data genetik dapat digunakan untuk berbagai penelitian atau *task*. Beberapa diantaranya adalah *drug discovery*, klasifikasi organisme, identifikasi protein, dan prediksi kanker. Tiap penelitian menggunakan atau mengembangkan metode *deep learning* yang khusus untuk memecahkan masalah tertentu sehingga model *deep learning* suatu penelitian tidak dapat digunakan pada penelitian yang lain meskipun menggunakan data genetik yang sama.

Masalah Khusus

Tidak ada metode komputasi yang dapat memproses data genetik menjadi suatu representasi tertentu yang dapat digunakan pada berbagai penelitian.

Referensi

Data genetik dapat dipandang sebagai sebuah teks dengan bahasa tertentu yang berisi informasi tentang makhluk hidup. Sebuah sekuens dapat dianalogikan sebagai teks dan tiap-tiap subsekuens dapat dipandang sebagai kata. Dengan demikian, data genetik dapat diproses sebagaimana layaknya bahasa (Iuchi et. al., 2021).

Arsitektur BERT dapat menghasilkan representasi semantik dari sebuah teks yang bersifat independen terhadap *task* pemrosesan bahasa secara khusus. Hal ini menjadikan arsitektur BERT dapat digunakan untuk berbagai *task* pada pemrosesan bahasa seperti *question answering* dan klasifikasi teks (Devlin et. al., 2018). DNABert dikembangkan berdasarkan arsitektur BERT untuk memproses data genetik untuk memecahkan berbagai masalah penelitian seperti identifikasi protein dan klasifikasi gen (Ji et. al., 2021).

Hal-Hal Pendukung

1. *Library* atau perangkat lunak untuk membuat arsitektur *deep learning* dapat diakses dengan mudah dan terdapat banyak tutorial cara penggunaan perangkat lunak tersebut.
2. Data genetik mudah diakses dan gratis.

Tujuan

1. Menguji representasi semantik BERT untuk berbagai *task* pada bidang genetik.
2. Menemukan representasi data genetik yang *task-independent* dan memiliki performa yang bagus untuk berbagai *task*.

Manfaat

1. Penelitian-penelitian *deep learning* yang menggunakan data sama dapat memanfaatkan representasi yang sama sehingga dapat meminimalisasi waktu yang digunakan untuk proses *training* model dari awal.