BAB I
PENDAHULUAN

1.1 Latar Belakang

Video captioning bertujuan untuk menghasilkan deskripsi singkat (*caption*) secara otomatis berdasarkan video yang diberikan seperti yang terlihat pada gambar 1.1. Hal ini akan sangat membantu dalam aplikasi interaksi manusia dan robot (Das dkk., 2017), alat bantu bagi penyandang tunanetra (Voykinska dkk., 2016), dan *video surveillance* (Namjoshi & Khurana, 2021). Aplikasi interaksi manusia dan robot seperti Alexa yang dapat mendeskripsikan keadaan kamar bayi berdasarkan masukan video CCTV. Selain itu, aplikasi alat bantu penyandang tunanetra dicontohkan pada alat yang mampu mendeskripsikan secara singkat keadaan sekitar atau konten *social media*. Lebih jauh lagi, aplikasi *video captioning* dapat membantu pengambilan keputusan pada *video surveillance* dengan jumlah data yang sangat banyak. Oleh karena itu, menghasilkan deskripsi singkat yang menjelaskan isi dari video yang diberikan secara otomatis sangat membantu kehidupan manusia.

Gambar 1.1 Contoh hasil generasi kalimat tugas *video captioning* yang ditunjukan dengan *baseline* dan ORG-TRL (Zhang dkk., 2020). *Ground truth* merupakan target kalimat.

Dalam bebesrapa tahun terakhir, metode *neural networks* pada *video captioning* berhasil menjadi *state-of-the-arts* dan pada umumnya mengadopsi arsitektur *encoder-decoder* (Pan dkk., 2016; Venugopalan dkk., 2015; Xu dkk., 2017; Yang dkk., 2021; Yao dkk., 2015). Arsitektur ini terlebih dahulu mengekstrak (*encode*) fitur visual video menjadi sekuens vektor menggunakan *Convolutional Neural Networks* (CNN). Lalu, sekuens vektor tersebut dijadikan dasar untuk

membangun kalimat deskripsi singkat menggunakan *Recurrent Neural Networks* (RNN) atau *Transformer* (Hori dkk., 2017). Metode *video captioning* dikatakan baik jika dapat mendeskripsikan hubungan antar objek dan memiliki waktu latensi *inference* yang rendah. Namun metode saat ini masih memiliki kekurangan pada kedua aspek tersebut.

Pada aspek representasi visual, metode-metode saat ini (Liu dkk., 2021; Yang dkk., 2021) umumnya hanya memanfaatkan *appearance features* dari *keyframes* dan *motion features* dari beberapa segmen *frames* untuk merepresentasikan sebuah video. Representasi ini menangkap informasi video secara global tetapi sulit untuk memperoleh informasi dinamis objek secara detail. Disebutkan oleh (Zhang dkk., 2020) bahwa informasi objek secara detail khususnya hubungan antar objek merupakan hal yang penting untuk menghasilkan deskripsi yang lebih detail dan beragam. Penelitian terbaru (Zhang dkk., 2020) memanfaatkan *object detector* untuk memperoleh informasi objek dan *Graph Convolutional Network* (GCN) untuk mempelajari hubungan antar objek. Berdasarkan hasil penelitian tersebut, ditunjukan bahwa menyertakan informasi objek dan hubungan antar objek merupakan hal yang dibutuhkan untuk pemahaman konten video agar menghasilkan deskripsi yang lebih baik.

Pada aspek pembangkitan *caption*, hampir seluruh penelitian menggunakan pendekaatan *autoregressive* (AR) seperti penelitian (Zhang dkk., 2020). Pendekatan *autoregressive* (AR) merupakan cara pembangkitan *caption* dimana mengkondisikan setiap kata berdasarkan keluaran yang dihasilkan sebelumnya. Pendekatan secara sekuensial tersebut mengakibatkan waktu *inference* yang lebih tinggi pada kalimat yang panjang, terutama untuk *caption* yang detail dan mengandung kata deskriptif yang lebih banyak (Gella dkk., 2018). Oleh karena itu, peneliti (Liu dkk., 2021; Yang dkk., 2021) mengadopsi pendekatan *non-autoregressive* (NA) untuk *video captioning* dan berhasil mengurangi waktu *inference* secara signifikan. Hal ini dikarenakan pendekatan tersebut membangkitkan *caption* secara paralel dibandingkan cara sekuensial AR. Hal ini juga yang menyebabkan pendekatan NA memiliki unjuk kerja yang masih kurang baik dibandingkan dengan pendekatan AR. Meskipun demikian, pendekatan NA ini

penting untuk aplikasi *video captioning* seperti interaksi robot dan manusa atau alat bantu tunanetra karena memiliki waktu *inference* yang rendah.

Pada penelitian ini, saya mengajukan sebuah model *video captioning* yang bernama *Object Relational Graph* dengan pendekatan *Non*-autoregressive (ORG-NA) untuk mengatasi masalah *video captioning* pada kedua aspek tersebut. Modul ORG (Zhang dkk., 2020) digunakan untuk memperoleh informasi objek secara detail dan mempelajari hubungan antar objek. Sedangkan modul NA (Yang dkk., 2021) digunaksan untuk membangkitkan *caption* dengan waktu *inference* yang lebih rendah. Diharapkan penggabungan tersebut menghasilkan model *video captioning* yang dapat menangkap informasi hubungan antar objek dan memiliki waktu *inference* yang rendah.

1.2 Rumusan Masalah

Berdasarkan latar belakang masalah yang telah diuraikan maka dirumuskan masalah sebagai berikut:

1) Bagaimana implementasi model *Object Relational Graph* dengan pendekatan *Non-autoregressive?*
2) Bagaimana hasil unjuk kerja model *Object Relational Graph* dengan pendekatan *Non-autoregressive?*
3) Bagaimana kecepatan waktu *inference* model *Object Relational Graph* dengan pendekatan *Non-autoregressive*?

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah dibuat, tujuan dilakukannya tugas akhir ini disebutkan sebagai berikut:

1) Mengimplementasikan model *Object Relational Graph* dengan pendekatan *Non-autoregressive*.
2) Menganalisis unjuk kerja model *Object Relational Graph* dengan pendekatan *Non-autoregressive*.
3) Menganalisis kecepatan waktu *inference* model *Object Relational Graph* dengan pendekatan *Non-autoregressive*.

1.4 Manfaat Penelitian

Adapun manfaat yang diharapkan dari penelitian ini sebagai berikut:

1) Bagi Peneliti

Peneliti diharapkan mampu mendapatkan pengetahuan baru mengenai penyelesaian masalah *video captioning* pada bidang *computer vision* dan pengolahan bahasa alami.

2) Bagi Pihak Lain

Hasil penelitian ini diharapkan mampu diimplementasikan dalam penyelesaian masalah *video captioning* pada bidang *computer vision* dan pengolahan bahasa alami dapat dijadikan rujukan pada penelitian selanjutnya.

1.5 Batasan Masalah

Batasan masalah ditentukan agar penelitian yang dilakukan fokus terhadap bidang yang diteliti dan disebutkan sebagai berikut:

1) Setiap video yang digunakan pada penelitian ini memiliki durasi selama 10 detik hingga 30 detik.
2) *Caption* didefinisikan sebagai deskripsi singkat berbahasa inggris yang menjelaskan isi dari video secara menyeluruh.
3) *Dataset* yang digunakan merupakan dataset standar *video captioning* bernama MSR-VTT dan dapat diperoleh di Kaggle.com.

1.6 Sistematika Penulisan

Sistematika Penulisan Skripsi terdiri dari 5 bab dengan struktur sebagai berikut:

BAB I PENDAHULUAN

Pada bab ini terdiri dari latar belakang dari topik *video captioning*, rumusan masalah, tujuan penelitian, manfaat penelitian, batasan penelitian, dan sistematika penulisan pada skripsi ini.

BAB II KAJIAN PUSTAKA

Pada bab ini memuat kajian pustaka terkait *video captioning* dan metode yang digunakan untuk menyelesaikan masalah tugas ini. Dimulai dari penjelasan tentang *video captioning* dan penjelasan metode yang sudah umum digunakan untuk menyelesaikan masalah ini yaitu metode *encoder-decoder*. Dilanjutkan

dengan membahas komponen yang digunakan sebagai *encoder* meliputi *convolutional neural network* (CNN), ResNet, InceptionResNetV2, ResNeXt, Faster R-CNN, dan *Object Relational Graph* (ORG). Komponen *decoder* yang dibahas meliputi *Long Short Term Memory* (LSTM) dan *Transformer*. Setelah itu, pembahasan paradigma pembangkitan *caption* seperti *autoregressive, non-autoregressive,* dan *Teacher Recommended Learning* (TRL). Terakhir, dijelaskan tentang dataset, metrik evaluasi, dan penelitian terkait tentang *video captioning*.

BAB III METODE PENELITIAN

Pada bab ini dijelaskan desain penelitian atau alur penelitian dari awal penelitian yaitu perumusan masalah, perancangan dan pemodelan model *video captioning*, perancangan dan implementasi skenario eksperimen, hingga analisis dan evaluasi dari model *video captioning*.

BAB IV TEMUAN DAN PEMBAHASAN

Pada bab ini dibahas praproses dan pengolahan data yang dilakukan, penjelasan cara pemodelan model yang diajukan, dan menganalisis hasil model *video captioning* yang diperoleh. Hasil yang diperoleh kemudian dibandingkan dengan model-model *video captioning* saat ini.

BAB V KESIMPULAN DAN SARAN

Pada bab ini disebutkan kesimpulan yang didapat dari penelitian tentang *video captioning* berdasarkan hasil temuan yang diperoleh. Peneliti juga menyampaikan saran bagi peneliti selanjutnya yang akan mengembangkan penelitian ini.