

**IMPLEMENTASI *VIDEO CAPTIONING* MENGGUNAKAN  
*OBJECT RELATIONAL GRAPH*  
DENGAN PENDEKATAN *NON-AUTOREGRESSIVE***

**SKRIPSI**

Diajukan untuk Memenuhi Sebagian dari  
Syarat Memperoleh Gelar Sarjana Komputer  
Program Studi Ilmu Komputer



Oleh  
Muhammad Ilham Malik  
1902563

**PROGRAM STUDI ILMU KOMPUTER  
FAKULTAS PENDIDIKAN MATEMATIKA DAN ILMU PENGETAHUAN  
ALAM  
UNIVERSITAS PENDIDIKAN INDONESIA  
BANDUNG  
2022**

**IMPLEMENTASI *VIDEO CAPTIONING* MENGGUNAKAN  
*OBJECT RELATIONAL GRAPH*  
DENGAN PENDEKATAN NON-AUTOREGRESSIVE**

Oleh  
Muhammad Ilham Malik  
NIM 1902563

Diajukan untuk memenuhi salah satu syarat memperoleh gelar Sarjana Komputer  
pada Fakultas Pendidikan Matematika dan Ilmu Pengetahuan Alam

© Muhammad Ilham Malik  
Universitas Pendidikan Indonesia  
Agustus 2023

Hak cipta dilindungi Undang-Undang  
Skripsi ini tidak boleh diperbanyak seluruhnya atau sebagian, dengan  
dicetak ulang, difotokopi, atau cara lainnya tanpa izin dari penulis.

MUHAMMAD ILHAM MALIK

**IMPLEMENTASI *VIDEO CAPTIONING* MENGGUNAKAN  
*OBJECT RELATIONAL GRAPH*  
DENGAN PENDEKATAN NON-AUTOREGRESSIVE**

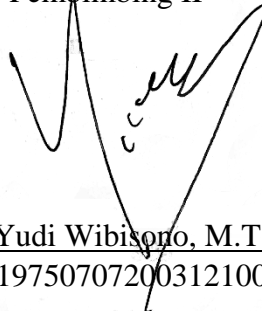
disetujui dan disahkan oleh pembimbing:

Pembimbing I



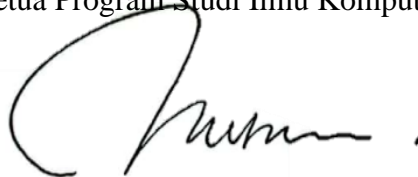
Yaya Wihardi, M.Kom  
NIP: 198903252015041001

Pembimbing II



Dr. Yudi Wibisono, M.T.  
NIP: 197507072003121003

Mengetahui  
Ketua Program Studi Ilmu Komputer



Dr. Muhamad Nursalman, M.T.  
NIP. 197909292006041002

## SURAT PERNYATAAN

Dengan ini saya menyatakan bahwa skripsi/tesis/disertasi dengan judul "Implementasi *Video Captioning* Menggunakan *Object Relational Graph* dengan Pendekatan *Non-Autoregressive*" ini beserta seluruh isinya adalah benar-benar karya saya sendiri. Saya tidak melakukan penjiplakan atau pengutipan dengan cara-cara yang tidak sesuai dengan etika ilmu yang berlaku dalam masyarakat keilmuan. Atas pernyataan ini, saya siap menanggung risiko/sanksi apabila di kemudian hari ditemukan adanya pelanggaran etika keilmuan atau ada klaim dari pihak lain terhadap keaslian karya saya ini.

Bandung, Juli 2023  
Yang membuat pernyataan

Muhammad Ilham Malik  
1902563

## KATA PENGANTAR

Puji dan syukur penulis panjatkan kepada kehadirat Allah SWT. karena atas kehendak-Nya serta karunia-Nya penyusunan skripsi yang berjudul “Implementasi *Video Captioning* Menggunakan *Object Relational Graph* dengan Pendekatan *Non-Autoregressive*” ini dapat diselesaikan tepat waktu. Penyusunan skripsi ini ditujukan untuk memenuhi dan melengkapi salah satu syarat untuk memperoleh gelar sarjana komputer atas jenjang S1 pada Program Studi Ilmu Komputer Universitas Pendidikan Indonesia.

Penulis menyadari bahwa selama penyusunan skripsi ini tidak terlepas dari peran, dukungan, dan bantuan dari berbagai pihak baik secara langsung atau tidak langsung. Oleh karena itu, penulis sampaikan terima kasih yang sebesar-besarnya kepada:

1. Kedua orang tua yang senantiasa berdoa demi kelancaran dalam penyusunan skripsi ini serta dukungan secara moral ataupun materi.
2. Bapak Yaya Wihardi, M.Kom., dan Bapak Dr. Yudi Wibisono, M.T., sebagai pembimbing, atas segala ilmu, motivasi dan waktu yang telah dicurahkan kepada penulis serta sumber daya komputer yang telah dipinjamkan kepada penulis.
3. Ibu Rosa Ariani Sukamto, M.T., atas segala ilmu dasar pemrograman hingga pemrograman tingkat lanjut, motivasi, dan serta kesediannya untuk berdiskusi.
4. Ibu Dr. Rani Megasari, M.Kom., dan Bapak Dr. Muhammad Nursalman, M.T., dan segenap dosen Prodi Ilmu Komputer atas ilmu yang sangat bermanfaat yang telah ditanamkan kepada penulis.
5. Rekan-rekan mahasiswa Prodi Ilmu Komputer 2019 yang telah bersama berjuang menimba ilmu.
6. Semua pihak yang telah membantu penulis dalam menyelesaikan skripsi ini yang tidak bisa disebutkan satu-persatu.

Akhirnya, penulis menyampaikan permohonan maaf atas segala ketidaksempurnaan. Penulis sangat mengharapkan setiap saran dan kritik yang

membangun agar skripsi ini dapat bermanfaat bagi kemajuan ilmu pengetahuan dan teknologi.

Bandung, Juli 2023

Penulis

**IMPLEMENTASI VIDEO CAPTIONING MENGGUNAKAN  
OBJECT RELATIONAL GRAPH  
DENGAN PENDEKATAN NON-AUTOREGRESSIVE**

Oleh

Muhammad Ilham Malik – muh.ilham.malik@upi.edu

1902563

**ABSTRAK**

Kemampuan video captioning dalam menghasilkan deskripsi singkat yang menjelaskan isi dari video secara detail dengan waktu inference yang rendah merupakan hal yang penting. Namun, metode-metode yang ada saat ini memiliki kekurangan pada kedua aspek tersebut. Pada penelitian ini, kami mengajukan sebuah model video captioning yang bernama *Object Relational Graph* dengan pendekatan *Non-autoregressive Coarse-to-Fine* (ORG-NACF) untuk mengatasi masalah video captioning pada kedua aspek tersebut. Modul ORG digunakan untuk memperoleh informasi objek secara detail dan mempelajari hubungan antar objek. Modul NACF bersama sequential cross attention digunakan untuk menyelesaikan masalah waktu *inference* yang tinggi dan menjaga kualitas caption ketika pembangkitan *caption*. Evaluasi dengan Dataset MSR-VTT menunjukkan hasil unjuk kerja Model ORG-NACF yang setara bahkan melebihi dari model *state-of-the-art* pada beberapa metrik serta memiliki kelebihan lain yaitu waktu *inference* pembangkitan *caption* yang lebih rendah. Hasil ini menunjukkan bahwa Model ORG-NACF mampu membangkitkan *caption* yang deskriptif dengan waktu *inference* yang lebih rendah dibandingkan dengan metode yang ada saat ini.

Kata Kunci: *Video Captioning, Transformer, Object Detection, Graph Convolutional Network, Convolutional Neural Network.*

**IMPLEMENTASI *VIDEO CAPTIONING* MENGGUNAKAN  
*OBJECT RELATIONAL GRAPH*  
DENGAN PENDEKATAN *NON-AUTOREGRESSIVE***

*Arranged by*

Muhammad Ilham Malik – muh.ilham.malik@upi.edu

1902563

**ABSTRACT**

The ability of video captioning to generate a detailed caption that explains the content of the video with low inference is important. However, existing methods have limitations in both aspects. In this paper, we propose a video captioning model Object Relational Graph with Non-autoregressive Coarse to Fine (ORG-NACF) approach to tackle the video captioning problem in both aspects. The ORG module is used to obtain detailed object information and learn the relationship between the objects. The NACF module along with sequential cross attention is used to solve the problem of high inference time and maintain caption quality during caption generation. Experimental evaluation on benchmark MSR-VTT dataset shows that the performance of the ORG-NACF model is competitive and even exceeds the state-of-the-art model on several metrics and has the advantage of faster inference time. This model achieved 7 times more faster inference time than the baseline model. These results show that the ORG-NACF Model is able to generate descriptive and detailed captions with lower inference time compared to existing methods.

*Kata Kunci: Video Captioning, Transformer, Object Detection, Graph Convolutional Network, Convolutional Neural Network.*



## DAFTAR ISI

<b>KATA PENGANTAR.....</b>	<b>IV</b>
<b>ABSTRAK .....</b>	<b>VI</b>
<b>ABSTRACT .....</b>	<b>VII</b>
<b>DAFTAR ISI.....</b>	<b>VIII</b>
<b>DAFTAR TABEL .....</b>	<b>X</b>
<b>DAFTAR GAMBAR.....</b>	<b>XII</b>
<b>BAB I PENDAHULUAN.....</b>	<b>1</b>
1.1 LATAR BELAKANG.....	1
1.2 RUMUSAN MASALAH .....	3
1.3 TUJUAN PENELITIAN .....	3
1.4 MANFAAT PENELITIAN.....	4
1.5 BATASAN MASALAH .....	4
1.6 SISTEMATIKA PENULISAN .....	4
<b>BAB II KAJIAN PUSTAKA .....</b>	<b>6</b>
2.1 PETA LITERATUR .....	6
2.2 VIDEO CAPTIONING.....	6
2.3 CONVOLUTIONAL NEURAL NETWORK .....	10
2.3.1 <i>ResNet</i> .....	19
2.3.2 <i>InceptionResNetV2</i> .....	22
2.3.4 <i>3D Convolution Neural Network</i> .....	27
2.3.5 <i>ResNeXt</i> .....	31
2.3.6 <i>Faster R-CNN</i> .....	34
2.3.7 <i>Object Relational Graph</i> .....	38
2.4 <i>LONG SHORT TERM MEMORY</i> .....	42
2.5 MEKANISME ATENSI .....	43
2.6 TRANSFORMER.....	46
2.6.1 <i>Attention Layer</i> .....	47
2.6.2 <i>Feed-Forward Layer</i> .....	51
2.6.3 <i>Positional Embedding</i> .....	52
2.6.4 <i>Skip Connection dan Normalization Layer</i> .....	54
2.6 DISTILBERT .....	56
2.7 PEMBANGKITAN CAPTION .....	57
2.7.1 <i>Autoregressive Decoding</i> .....	57
2.7.2 <i>Teacher Recommended Learning</i> .....	59
2.7.3 <i>Non-Autoregressive Coarse-to-Fine</i> .....	61
2.8 DATASET MSR-VTT .....	64
2.9 METRIK EVALUASI.....	66
2.9.1 <i>BLEU</i> .....	66
2.9.2 <i>METEOR</i> .....	69

2.9.3	<i>ROGUE-L</i> .....	73
2.9.4	<i>CIDEr</i> .....	75
2.10	PENELITIAN TERKAIT <i>VIDEO CAPTIONING</i> .....	79
<b>BAB III METODE PENELITIAN .....</b>		<b>83</b>
3.1	DESAIN PENELITIAN .....	83
3.1.1	<i>Perumusan Masalah</i> .....	83
3.1.2	<i>Studi Literatur</i> .....	84
3.1.3	<i>Pengumpulan Data</i> .....	84
3.1.4	<i>Perancangan Metode</i> .....	85
3.1.5	<i>Pemodelan Video Captioning</i> .....	88
3.1.6	<i>Perancangan Skenario Eksperimen</i> .....	89
3.1.7	<i>Eksperimen</i> .....	90
3.1.8	<i>Analisis dan Evaluasi</i> .....	90
3.1.9	<i>Penarikan Kesimpulan</i> .....	91
3.2	ALAT DAN BAHAN PENELITIAN.....	91
3.2.1	<i>Alat Penelitian</i> .....	91
3.2.2	<i>Bahan Penelitian</i> .....	92
<b>BAB IV TEMUAN DAN PEMBAHASAN .....</b>		<b>93</b>
4.1	PRAPROSES PENGOLAHAN DATA .....	93
4.1.1	<i>Pengambilan Frame dari Video</i> .....	93
4.1.2	<i>Ekstraksi Appearance Feature</i> .....	95
4.1.3	<i>Ekstraksi Motion Feature</i> .....	97
4.1.4	<i>Ekstraksi Object Feature</i> .....	98
4.1.5	<i>Praproses Caption Video</i> .....	100
4.2	PEMODELAN VIDEO CAPTIONING .....	103
4.2.1	<i>Pemodelan IEL-TRL</i> .....	103
4.2.2	<i>Pemodelan ORG-NACF</i> .....	126
4.3	ANALISIS HASIL .....	157
4.3.1	<i>Perbandingan Unjuk Kerja</i> .....	157
4.3.2	<i>Analisis Kualitatif</i> .....	161
4.3.3	<i>Analisis Waktu Inference Model</i> .....	164
<b>BAB V KESIMPULAN DAN SARAN .....</b>		<b>167</b>
5.1	KESIMPULAN.....	167
5.2	SARAN.....	167

## DAFTAR TABEL

Tabel 2.1 Rincian Perhitungan Metrik Evaluasi METEOR.....	72
Tabel 2.2 Contoh Daftar Kalimat Rujukan dan Kalimat Kandidat .....	77
Tabel 2.3 Contoh Representasi Unigram Setiap Kalimat .....	77
Tabel 2.4 Tabel Perhitungan Term Frequency (TF) dan Inverse Document Frequency (IDF).....	78
Tabel 3.1 Rincian Skenario Penggunaan Jumlah Data .....	90
Tabel 4.1 Command yang Digunakan Ketika Pengambilan Frame .....	94
Tabel 4.2 Implementasi Setiap Layer Modul IEL Menggunakan Framework Pytorch.....	108
Tabel 4.3 Variasi Jumlah Data dan Rasio Pembagian Himpunan Data .....	112
Tabel 4.4 Model yang Digunakan Untuk Ekstraksi Fitur dan Pembangkitan Soft Target .....	112
Tabel 4.5 Daftar Hyperparameter dan Nilai Yang Digunakan Untuk Fine-Tuning Model DistilBERT.....	113
Tabel 4.6 Daftar Hyperparameter dan Nilai Yang Digunakan Untuk Training Model IEL-TRL.....	113
Tabel 4.7 Unjuk Kerja Pengembangan Model Pada 100 Data Klip Video.....	116
Tabel 4.8 Unjuk Kerja Pengembangan Model IEL-TRL Pada 1.000 Data Klip Video .....	121
Tabel 4.9 Hasil Prediksi Kata jika Diberikan Sebuah Kalimat yang Sudah Ditentukan .....	122
Tabel 4.10 Unjuk Kerja Pengembangan Model IEL-TRL Pada 1.000 Data Klip Video .....	124
Tabel 4.11 Unjuk Kerja Pengembangan Model IEL-TRL Terhadap Baseline Pada 100 Data Klip Video .....	124
Tabel 4.12 Unjuk Kerja Pengembangan Model IEL-TRL Terhadap Baseline Pada 1.000 Data Klip Video .....	125
Tabel 4.13 Unjuk Kerja Pengembangan Model IEL-TRL Terhadap Baseline Pada 10.000 Data Klip Video .....	126
Tabel 4.14 Variasi Jumlah Data dan Rasio Pembagian Himpunan Data.....	136
Tabel 4.15 Model yang Digunakan Untuk Ekstraksi Fitur Dari Klip Video. ....	136

Tabel 4.16 Implementasi <i>Layer</i> Modul ORG Menggunakan <i>Framework</i> Pytorch .....	137
Tabel 4.17 Implementasi Bagian <i>Decoder</i> dari Model ORG-LSTM Menggunakan <i>Framework</i> Pytorch.....	137
Tabel 4.18 Implementasi Bagian <i>Decoder</i> dari Model ORG-NACF Menggunakan <i>Framework</i> Pytorch.....	137
Tabel 4.19 Daftar <i>Hyperparameter</i> Dan Nilai Yang Digunakan Untuk <i>Training</i> Model ORG-NACF. ....	138
Tabel 4.20 Unjuk Kerja Pengembangan Model Pada 100 Data Klip Video.....	143
Tabel 4.21 Unjuk Kerja Pengembangan Modul ORG Menggunakan <i>Dropout</i> <i>Layer</i> .....	147
Tabel 4.22 Unjuk Kerja Pengembangan Modul ORG Menggunakan <i>Activation</i> <i>Function</i> .....	149
Tabel 4.23 Unjuk Kerja Pengembangan Model Pada 1.000 Data Klip Video...	152
Tabel 4.24 Unjuk Kerja Model ORG-NACF Pada 10.000 Data Klip Video.....	155
Tabel 4.25 Unjuk Kerja Tiap Pengembangan Model ORG-NACF Pada 100 Data Klip Video .....	155
Tabel 4.26 Unjuk Kerja Tiap Pengembangan Model ORG-NACF Pada 1000 Data Klip Video .....	156
Tabel 4.27 Unjuk Kerja Tiap Pengembangan Model ORG-NACF Pada 10.000 Data Klip Video .....	157
Tabel 4.28 Hasil Evaluasi Unjuk Kerja Setiap Model Pada 100 Data Video...	158
Tabel 4.29 Hasil Evaluasi Unjuk Kerja Setiap Model Pada 1.000 Data Video MSRVTT.....	158
Tabel 4.30 Hasil Evaluasi Unjuk Kerja Setiap Model Pada 10.000 Data Video MSR-VTT. ....	159
Tabel 4.31 Perbandingan Waktu <i>Inference</i> Model Yang Diajukan Terhadap <i>Baseline</i> .....	164

## DAFTAR GAMBAR

Gambar 1.1 Contoh hasil generasi kalimat tugas <i>video captioning</i> yang ditunjukkan dengan <i>baseline</i> dan ORG-TRL (Zhang dkk., 2020). <i>Ground truth</i> merupakan target kalimat.....	1
Gambar 2.1 Gambar peta literatur implementasi model <i>video vaptioning</i> .....	6
Gambar 2.2 Diagram salah satu contoh ekstraksi fitur visual.....	7
Gambar 2.3 Diagram salah satu contoh operasi agregasi fitur.....	8
Gambar 2.4 Diagram <i>Layer-Layer</i> yang Membangun <i>Convolutional Neural Network</i> (O’Shea & Nash, 2015).....	11
Gambar 2.5 Representasi visual operasi dot product antara kernel terhadap receptive field citra pada convolution layer.....	12
Gambar 2.6 Ilustrasi perubahan alur citra digital setelah melewati <i>convolutional layer</i> yang memiliki 16 <i>kernel</i> dengan dimensi 3×3.....	13
Gambar 2.7 Ilustrasi penggunaan <i>padding</i> pada citra.....	14
Gambar 2.8 Contoh citra sebelum dilakukan operasi konvolusi.....	15
Gambar 2.9 Contoh visualisasi <i>feature maps</i> hasil operasi konvolusi menggunakan Model ResNet-18. ....	15
Gambar 2.10 Contoh Operasi <i>Max Pooling</i> Pada Matriks <i>Activation Map</i> .....	16
Gambar 2.11 (Géron, 2022) Menunjukkan invarians terhadap translasi. ....	17
Gambar 2.12 Ilustrasi dua buah layer FC pada arsitektur CNN. ....	18
Gambar 2.13 Ilustrasi dari <i>residual learning</i> (Géron, 2022). ....	19
Gambar 2.14 Ilustrasi jaringan yang belum memiliki kemajuan karena terhalang <i>layer</i> (kiri) dan jaringan yang memiliki kemajuan meskipun terhalang <i>layer</i> melalui <i>skip connection</i> (kanan) (Géron, 2022). ....	20
Gambar 2.15 Representasi sederhana arsitektur ResNet (He dkk., 2016) dari (Géron, 2022).....	21
Gambar 2.16 (Szegedy dkk., 2017) Tiga varian arsitektur <i>Inception</i> yang digunakan Inception-ResNet-v2 diantaranya adalah <i>Inception-A</i> (Kiri), <i>Inception-B</i> (Tengah), dan <i>Inception-C</i> (Kanan).....	22
Gambar 2.17 Modul <i>Inception</i> (ditunjukkan dengan garis hijau) dan <i>filter expansion layer</i> (ditunjukkan dengan garis oranye).....	23

Gambar 2.18 Modul <i>Inception</i> yang pertama kali diajukan pada (Szegedy et al., 2015) yang direpresentasikan oleh (Géron, 2022).....	24
Gambar 2.19 Komponen Modul <i>Stem</i> (Szegedy dkk., 2017) .....	25
Gambar 2.20 Modul <i>Reduction-A</i> (bagian kiri) dan Modul <i>Reduction-B</i> (bagian kanan) (Szegedy dkk., 2017) .....	26
Gambar 2.21 Skema arsitektur jaringan Inception-ResNet-v2. ....	27
Gambar 2.22 Operasi 2D CNN terhadap sebuah <i>frame</i> (kiri) dan berbagai <i>frame</i> (kanan) menghasilkan keluaran 2-dimensi. ....	28
Gambar 2.23 Operasi 3D CNN terhadap beberapa <i>frame</i> menghasilkan keluaran berbentuk volume. ....	28
Gambar 2.24 Operasi 3D <i>convolution</i> pada input citra dengan menggunakan <i>kernel</i> 3-dimensi dan menghasilkan sebuah <i>feature map</i> 3-dimensi 29	
Gambar 2.25 Operasi matematika dari operasi 3D <i>convolution</i> secara singkat....	30
Gambar 2.26 Operasi 3D <i>max pooling</i> pada <i>feature maps</i> .....	31
Gambar 2.27 Modul dari ResNet (He dkk., 2016) (kiri) dan Modul dari ResNeXt (Xie dkk., 2017) (kanan).....	32
Gambar 2.28 Beberapa blok pembangun yang memiliki ekuivalensi pada bagian kiri memiliki susunan yang sama seperti Gambar 2.27 dan pada bagian kanan menggunakan implementasi <i>grouped convolution</i> ....	33
Gambar 2.29 Arsitektur ResNeXt pada penelitian yang dilakukan oleh (Hara dkk., 2018). ....	34
Gambar 2.30 Jaringan yang membangun <i>object detector</i> Faster R-CNN (Ren dkk., 2015). ....	35
Gambar 2.31 Ilustrasi tahap pertama pengajuan region dari Faster R-CNN .....	36
Gambar 2.32 Visualisasi nilai IoU antara <i>anchor box</i> dengan <i>ground truth box</i> . 37	
Gambar 2.33 Arsitektur Fast R-CNN (Girshick, 2015). ....	38
Gambar 2.34 Alur perubahan data dari modul ORG. ....	40
Gambar 2.35 Mengilustrasikan hasil dari matriks koefisien relasi <i>A</i> pada P-ORG (kiri) dan pada C-ORG (kanan) .....	41
Gambar 2.36 Ilustrasi arsitektur Model LSTM.....	42
Gambar 2.37 Ilustrasi Diagram dari Mekanisme Atensi Bahdanau (Bahdanau dkk., 2014) .....	44

Gambar 2.38 Diagram Arsitektur <i>Transformer</i> (Vaswani dkk., 2017).....	47
Gambar 2.39 Visualisasi langkah-langkah scaled dot product attention (Tunstall dkk., 2022) .....	49
Gambar 2.40 Ilustrasi rangkaian proyeksi menggunakan <i>linear layer</i> yang dinamakan <i>attention head</i> .....	50
Gambar 2.41 Ilustrasi proses transformasi pada operasi <i>feed-forward layer</i> .....	51
Gambar 2.42 Diagram bagian Transformer Encoder (Vaswani dkk., 2017) yang menunjukkan feed-forward layer.....	52
Gambar 2.43 Contoh ilustrasi matriks <i>positional embedding</i> dengan dimensi fitur 4, konstanta bernilai 100, dan panjang kalimat 4. ....	53
Gambar 2.44 Operasi penggabungan <i>token embedding</i> dan <i>positional embedding</i> dimodifikasi dari gambar (Vaswani dkk., 2017) .....	54
Gambar 2.45 Operasi <i>skip connection</i> dan <i>layer normalizaiton</i> pada Model <i>Transformer</i> (Vaswani dkk., 2017).....	54
Gambar 2.46 Pendekatan <i>autoregressive</i> dengan metode <i>greedy decoding</i> .....	58
Gambar 2.47 Ilustrasi pembangkitan caption dengan pendekatan autoregressive menggunakan metode beam search (Tunstall dkk., 2022).....	58
Gambar 2.48 Visualisasi metode TRL dan metode TEL. ....	60
Gambar 2.49 Ilustrasi pendekatan Non-autoregressive Coarse-to-Fine (Yang dkk., 2021). ....	61
Gambar 2.50 Ilustrasi algoritma <i>Mask-Predict</i> pada pendekatan NACF. ....	64
Gambar 2.51 Merupakan contoh pasangan klip dan kalimat yang ada pada dataset MSR-VTT (Xu dkk., 2016) .....	65
Gambar 2.52 Visualisasi perhitungan BLEU <i>unigram</i> dan <i>bigram</i> . ....	67
Gambar 2.53 Ilustrasi modifikasi yang dilakukan terhadap perhitungan Metrik BLEU untuk mengatasi kata berulang pada kalimat kandidat.....	68
Gambar 2.54 Dua contoh hasil <i>alignment</i> antara kalimat rujukan (atas) dan kalimat kandidat (bawah).....	69
Gambar 2.55 Visualisasi contoh perhitungan hasil nilai METEOR pada dua <i>alignment</i> berbeda.....	71
Gambar 2.56 Contoh Perhitungan Metrik Evaluasi Rouge-L.....	74

Gambar 2.57 Hasil operasi <i>cosine similarity</i> untuk memperoleh nilai CIDEr <i>unigram</i> .....	78
Gambar 3.1 Desain Penelitian yang akan dilakukan.....	83
Gambar 3.2 Rancangan metode .....	85
Gambar 3.3 Arsitektur Model IEL-TRL .....	86
Gambar 3.4 Arsitektur Model ORG-NACF.....	87
Gambar 3.5 Skenario pemodelan IEL-TRL .....	87
Gambar 3.6 Skenario pemodelan ORG-NACF.....	88
Gambar 4.1 <i>Overview</i> proses pengambilan <i>frame</i> .....	94
Gambar 4.2 Visualisasi pengambilan <i>frame</i> dari video .....	95
Gambar 4.3 a). arsitektur umum model CNN yang digunakan untuk memprediksi kelas pada ImageNet b). arsitektur model yang digunakan untuk ekstraksi fitur .....	96
Gambar 4.4 Ilustrasi contoh masukan dan hasil keluaran dari Model ResNet101	97
Gambar 4.5 Proses <i>overview</i> ekstraksi <i>motion feature</i> menggunakan <i>pretrained</i> ResNeXt-101 .....	98
Gambar 4.6 Potongan arsitektur Faster R-CNN dan letak <i>object feature</i> .....	99
Gambar 4.7 Proses ekstraksi <i>object feature</i> menggunakan Model Faster R-CNN .....	100
Gambar 4.8 a). Proses <i>truncate</i> jika kalimat lebih panjang dari batas yang ditetapkan b). Proses <i>padding</i> jika panjang kalimat kurang dari batas. .....	101
Gambar 4.9 Proses merubah <i>caption</i> menjadi <i>lower case</i> , menghilangkan tanda baca, dan membangun <i>vocabulary</i> .....	102
Gambar 4.10 Contoh penggunaan <i>token</i> khusus terhadap <i>caption</i> . .....	103
Gambar 4.11 Arsitektur Model IEL-TRL yang akan diimplementasikan. ....	104
Gambar 4.12 Arsitektur SA-LSTM secara detail.....	104
Gambar 4.13 <i>Overview</i> dari proyeksi fitur pada Modul IEL .....	106
Gambar 4.14 Alur dan arsitektur dari Modul IEL .....	107
Gambar 4.15 Modifikasi <i>Encoder</i> SA-LSTM menjadi <i>Encoder</i> IEL-TRL .....	108
Gambar 4.16 Perbandingan Metode TEL dengan Metode TRL .....	109



Gambar 4.17 Ilustrasi perolehan Model ORG-NACF dimulai dari a) SA-LSTM, b) IEL-LSTM dan d) IEL-TRL.....	111
Gambar 4.18 Hasil <i>training</i> dan unjuk kerja setiap <i>epoch</i> Model SA-LSTM pada 100 data klip video.....	114
Gambar 4.19 Hasil <i>training</i> setiap <i>epoch</i> Model IEL-LSTM pada 100 data klip video.....	115
Gambar 4.20 Hasil unjuk kerja setiap <i>epoch</i> Model IEL-LSTM pada 100 data klip video.....	116
Gambar 4.21 Hasil <i>fine-tuning</i> setiap Model ELM untuk persiapan Metode TRL .....	117
Gambar 4.22 Hasil <i>training</i> Model IEL-TRL menggunakan BERT sebagai guru terhadap 1.000 data klip video.....	117
Gambar 4.23 Hasil unjuk kerja Model IEL-TRL menggunakan BERT sebagai guru terhadap 1.000 data klip video.....	118
Gambar 4.24 Hasil <i>training</i> setiap <i>epoch</i> Model IEL-TRL menggunakan DistilBERT sebagai guru pada 1.000 data klip video.....	119
Gambar 4.25 Hasil unjuk kerja setiap <i>epoch</i> Model IEL-TRL menggunakan DistilBERT sebagai guru pada 1.000 data klip video.....	119
Gambar 4.26 Hasil <i>training</i> setiap <i>epoch</i> Model IEL-TRL menggunakan MobileBERT sebagai guru pada 1.000 data klip video.....	120
Gambar 4.27 Hasil unjuk kerja setiap <i>epoch</i> Model IEL-TRL menggunakan MobileBERT sebagai guru pada 1.000 data klip video.....	121
Gambar 4.28 Hasil <i>fine-tuning</i> model ELM: a) DistilBERT dan b) MobileBERT .....	122
Gambar 4.29 Hasil <i>training</i> IEL-TRL menggunakan DistilBERT sebagai guru terhadap 10.000 data klip video.....	123
Gambar 4.30 Hasil unjuk kerja IEL-TRL menggunakan DistilBERT sebagai guru terhadap 10.000 data klip video.....	123
Gambar 4.31 Gambar Arsitektur ORG-NACF .....	127
Gambar 4.32 <i>Overview</i> dari proses Modul ORG.....	128
Gambar 4.33 Alur perubahan <i>object feature</i> pada operasi Modul <i>Object Relational Graph</i> (ORG) secara detail.....	129

Gambar 4.34 Modifikasi Model SA-LSTM menjadi ORG-LSTM merujuk pada (Zhang dkk., 2020).....	130
Gambar 4.35 Perbandingan bagian <i>decoder</i> dari Model SA-LSTM dan Model ORG-LSTM .....	131
Gambar 4.36 Perbandingan bagian <i>decoder</i> antara a). NACF dan b). modifikasi Model NACF menjadi ORG-NACF .....	133
Gambar 4.37 Perbedaan cara dalam memanfaatkan sequential fusion a).Concatenated feature first (F → R) b).Object feature first (R → F) (Nguyen dkk., 2022) .....	134
Gambar 4.38 Ilustrasi perolehan Model ORG-NACF dimulai dari a) SA-LSTM, b) ORG-LSTM, c) NACF, dan d) ORG-NACF. ....	135
Gambar 4.39 Hasil <i>training</i> menggunakan 100 data klip video oleh a). Model SA-LSTM b). Model ORG-LSTM.....	139
Gambar 4.40 Hasil unjuk kerja model setiap <i>epoch</i> terhadap 100 data klip video a). SA-LSTM b). ORG-LSTM .....	140
Gambar 4.41 Hasil nilai <i>loss training</i> pada 100 data klip video a). Model NACF dan b).ORG-NACF.....	141
Gambar 4.42 Hasil unjuk kerja model setiap <i>epoch</i> terhadap 1.000 data klip video a). NACF b). ORG-NACF .....	142
Gambar 4.43 Hasil <i>training</i> menggunakan 1.000 data klip video pada model a). SA-LSTM dan ORG-LSTM .....	144
Gambar 4.44 Hasil unjuk kerja setiap model pada setiap <i>epoch</i> terhadap 1.000 data klip video a). SA-LSTM dan b). ORG-LSTM.....	145
Gambar 4.45 Hasil <i>training</i> pengembangan Modul ORG menggunakan ragam <i>dropout rate</i> .....	146
Gambar 4.46 Hasil <i>training</i> pengembangan Modul ORG menggunakan ragam fungsi aktivasi .....	148
Gambar 4.47 Hasil <i>training loss</i> pada 1.000 data klip video a). Model NACF dan b). Model ORG-NACF .....	150
Gambar 4.48 Hasil unjuk kerja setiap model di setiap <i>epoch</i> terhadap 1.000 data klip video a). Model NACF dan b). Model ORG-NACF.....	151

Gambar 4.49 Hasil <i>training loss</i> 10.000 data klip video a). <i>Object First</i> dan b). <i>Global First</i> .....	153
Gambar 4.50 Hasil unjuk kerja setiap model di setiap <i>epoch</i> terhadap 10.000 data klip video a). <i>Object First</i> dan b). <i>Global First</i> .....	154
Gambar 4.51 Perbandingan hasil pembangkitan <i>caption</i> setiap model yang diajukan pada dataset MSR-VTT. ....	162
Gambar 4.52 Perbandingan <i>caption</i> terpanjang yang dihasilkan oleh setiap model terbaik .....	162
Gambar 4.53 Perbandingan <i>caption</i> terpendek yang dihasilkan oleh setiap model terbaik .....	163

## DAFTAR PUSTAKA

- Aafaq, N., Akhtar, N., Liu, W., Gilani, S. Z., & Mian, A. (2019). Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2019-June*, 12479–12488. <https://doi.org/10.1109/CVPR.2019.01277>
- Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *ArXiv Preprint ArXiv:1409.0473*.
- Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. *Proceedings of the Acl Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72.
- Chen, D., & Dolan, W. B. (2011). Collecting highly parallel data for paraphrase evaluation. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 190–200.
- Chen, S., & Jiang, Y.-G. (2021). Motion guided region message passing for video captioning. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1543–1552.
- Chen, S., & Jiang, Y.-G. (2019). Motion guided spatial attention for video captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01), 8191–8198.
- Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J. M. F., Parikh, D., & Batra, D. (2017). Visual dialog. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 326–335.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv Preprint ArXiv:1810.04805*.
- Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4), 193–202.

- Gella, S., Lewis, M., & Rohrbach, M. (2018). A dataset for telling the stories of social media videos. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 968–974.
- Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. “O’Reilly Media, Inc.”
- Ghazvininejad, M., Levy, O., Liu, Y., & Zettlemoyer, L. (2019). Mask-predict: Parallel decoding of conditional masked language models. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 6112–6121.  
<https://doi.org/10.18653/v1/d19-1633>
- Girshick, R. (2015). Fast r-cnn. *Proceedings of the IEEE International Conference on Computer Vision*, 1440–1448.
- Hara, K., Kataoka, H., & Satoh, Y. (2018). Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6546–6555.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-Decem*, 770–778.  
<https://doi.org/10.1109/CVPR.2016.90>
- Hori, C., Hori, T., Lee, T. Y., Zhang, Z., Harsham, B., Hershey, J. R., Marks, T. K., & Sumi, K. (2017). Attention-Based Multimodal Fusion for Video Description. *Proceedings of the IEEE International Conference on Computer Vision, 2017-October*, 4203–4212. <https://doi.org/10.1109/ICCV.2017.450>
- Hubel, D. H. (1959). Single unit activity in striate cortex of unrestrained cats. *The Journal of Physiology*, 147(2), 226.
- Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat’s striate cortex. *The Journal of Physiology*, 148(3), 574.
- Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *International Conference on*

*Machine Learning*, 448–456.

- LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324.
- Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 74–81.
- Liu, A. A., Xu, N., Wong, Y., Li, J., Su, Y. T., & Kankanhalli, M. (2017). Hierarchical & multimodal video captioning: Discovering and transferring multimodal knowledge for vision to language. *Computer Vision and Image Understanding*, 163, 113–125. <https://doi.org/10.1016/j.cviu.2017.04.013>
- Liu, F., Ren, X., Wu, X., Yang, B., Ge, S., & Sun, X. (2021). O2NA: An Object-Oriented Non-Autoregressive Approach for Controllable Video Captioning. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 281–292. <https://doi.org/10.18653/v1/2021.findings-acl.24>
- Namjoshi, M., & Khurana, K. (2021). A mask-rcnn based object detection and captioning framework for industrial videos. *International Journal of Advanced Technology and Engineering Exploration*, 8(84), 1466–1478. <https://doi.org/10.19101/IJATEE.2021.874394>
- Nguyen, V.-Q., Suganuma, M., & Okatani, T. (2022). Grit: Faster and better image captioning transformer using dual visual features. *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, 167–184.
- O’Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. *ArXiv Preprint ArXiv:1511.08458*.
- Pan, Y., Mei, T., Yao, T., Li, H., & Rui, Y. (2016). Jointly modeling embedding and translation to bridge video and language. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-Decem*, 4594–4602. <https://doi.org/10.1109/CVPR.2016.497>
- Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual*

- Meeting of the Association for Computational Linguistics*, 311–318.
- Pei, W., Zhang, J., Wang, X., Ke, L., Shen, X., & Tai, Y.-W. (2019). Memory-attended recurrent network for video captioning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8347–8356.
- Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28.
- Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation*, 60(5), 503–520.
- Rohrbach, A., Rohrbach, M., Tandon, N., & Schiele, B. (2015). A dataset for movie description. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3202–3212.
- Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv Preprint ArXiv:1910.01108*.
- Schmidhuber, J., & Hochreiter, S. (1997). Long short-term memory. *Neural Comput*, 9(8), 1735–1780.
- Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *ArXiv Preprint ArXiv:1409.1556*.
- Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1–9.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2818–2826.

- Torabi, A., Pal, C., Larochelle, H., & Courville, A. (2015). Using descriptive video services to create a large data source for video annotation research. *ArXiv Preprint ArXiv:1503.01070*.
- Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3D convolutional networks. *Proceedings of the IEEE International Conference on Computer Vision, 2015 Inter*, 4489–4497. <https://doi.org/10.1109/ICCV.2015.510>
- Tunstall, L., Von Werra, L., & Wolf, T. (2022). *Natural language processing with transformers*. “O’Reilly Media, Inc.”
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
- Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). Cider: Consensus-based image description evaluation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4566–4575.
- Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R., & Saenko, K. (2015). Translating videos to natural language using deep recurrent neural networks. *NAACL HLT 2015 - 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, 1494–1504. <https://doi.org/10.3115/v1/n15-1173>
- Voykinska, V., Azenkot, S., Wu, S., & Leshed, G. (2016). How blind people interact with visual content on social networking services. *Proceedings of the 19th Acm Conference on Computer-Supported Cooperative Work & Social Computing*, 1584–1595.
- Wang, X., Wu, J., Chen, J., Li, L., Wang, Y.-F., & Wang, W. Y. (2019). Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4581–4591.
- Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual



- transformations for deep neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1492–1500.
- Xu, J., Mei, T., Yao, T., & Rui, Y. (2016). Msr-vtt: A large video description dataset for bridging video and language. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5288–5296.
- Xu, J., Yao, T., Zhang, Y., & Mei, T. (2017). Learning multimodal attention LSTM networks for video captioning. *MM 2017 - Proceedings of the 2017 ACM Multimedia Conference*, 537–545.  
<https://doi.org/10.1145/3123266.3123448>
- Yang, B., Zou, Y., Liu, F., & Zhang, C. (2021). Non-autoregressive coarse-to-fine video captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(4), 3119–3127.
- Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., & Courville, A. (2015). Describing videos by exploiting temporal structure. *Proceedings of the IEEE International Conference on Computer Vision*, 4507–4515.
- Zhang, J., & Peng, Y. (2019). Object-aware aggregation with bidirectional temporal graph for video captioning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8327–8336.
- Zhang, Z., Shi, Y., Yuan, C., Li, B., Wang, P., Hu, W., & Zha, Z. (2020). Object relational graph with teacher-recommended learning for video captioning. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 13275–13285.  
<https://doi.org/10.1109/CVPR42600.2020.01329>