PEMBANGKITAN TEKS JUDUL SPESIFIK UNTUK GAMBAR PRODUK MENGGUNAKAN ATRIBUT SEMANTIK

TESIS

Karya tulis sebagai salah satu syarat
untuk memperoleh gelar Magister dari
Institut Teknologi Bandung

Oleh
IRFAN IHSANUL AMAL
NIM: 23520025
(Program Studi Magister Informatika)

INSTITUT TEKNOLOGI BANDUNG
Januari 2016

PEMBANGKITAN TEKS JUDUL SPESIFIK UNTUK GAMBAR PRODUK MENGGUNAKAN ATRIBUT SEMANTIK

Oleh
Irfan Ihsanul Amal
NIM: 23520025
(Program Studi Magister Informatika)

Institut Teknologi Bandung

Menyetujui
Pembimbing

Tanggal 22 Januari 2021

_____

Prof.Ir. Dwi Hendratmo Widyantoro M.Sc.,Ph.D.

DAFTAR ISI

HALAMAN PENGESAHAN i
DAFTAR ISI ii
DAFTAR GAMBAR DAN ILUSTRASI iii
DAFTAR TABEL iv

Bab I Pendahuluan 1
I.1 Latar Belakang 1
I.2 Masalah Penelitian 3
I.3 Tujuan 3
I.4 Hipotesis 3
I.5 Batasan Masalah 3
I.6 Metodologi 4

Bab II Tinjauan Pustaka 5
II.1 Image Captioning 5
II.2 Attention-based Image Captioning 7
II.3 Transformer 9
II.4 Ekstraksi Atribut untuk Image Captioning 12
II.5 Penelitian Terkait 14

Bab III Analisis Masalah dan Rancangan Solusi 21
III.1 Analisis Masalah 21
III.2 Rancangan Solusi 22
III.3 Alur Pembangunan Solusi 24

DAFTAR PUSTAKA 28

DAFTAR GAMBAR DAN ILUSTRASI

Gambar II.1 *Neural Image Caption* (Vinyals dkk., 2015) 5
Gambar II.2 Arsitektur model NIC dengan *visual attention* (Xu dkk., 2015) ... 7
Gambar II.3 Arsitektur transformer (Vaswani dkk., 2017) 9
Gambar II.4 Proses komputasi *attention* pada transformer (Vaswani dkk., 2017) 10
Gambar II.5 Arsitektur SCN-LSTM (Gan dkk., 2017) 12
Gambar II.6 Arsitektur DaE (Kim dkk., 2018) 14
Gambar II.7 Arsitektur transformer untuk *image captioning* (Yu dkk., 2019) 16
Gambar II.8 Skema AMV (Yu dkk., 2019) 17
Gambar II.9 Skema UMV (Yu dkk., 2019) 17
Gambar II.10 Arsitektur ViT (Dosovitsky dkk., 2020) 19
Gambar III.1 Judul produk hasil NIC 21
Gambar III.2 Arsitektur model klasifikasi atribut 23
Gambar III.3 Atsitektur model *image captioning* 24
Gambar III.4 Alur pembangunan solusi 25

DAFTAR TABEL

Tabel II.1 Skor NIC pada MSCOCO (Vinyals dkk., 2015) 6
Tabel II.2 Skor BLEU-1 NIC pada berbagai dataset (Vinyals dkk., 2015) 6
Tabel II.3 Skor model NIC dengan *attention* (Xu dkk., 2015) 8
Tabel II.4 Skor model transformer (Vaswani dkk., 2017) 11
Tabel II.5 Skor SCN-LSTM pada dataset COCO (Gan dkk., 2017)..... 13
Tabel II.6 Skor DaE pada dataset COCO (Kim dkk., 2018)..... 15
Tabel II.7 Skor MT pada dataset Flickr30k (Yu dkk., 2019)..... 18
Tabel II.8 Skor akurasi ViT pada berbagai dataset (Dosovitsky dkk., 2020) . 20

Bab I Pendahuluan

I.1 Latar Belakang

Image captioning merupakan suatu *task* yang memetakan *input* gambar menjadi suatu teks deskripsi atau *caption*. *Task* tersebut dapat dilakukan untuk gambar-gambar yang bersifat umum atau biasa disebut dengan *general image captioning*. Selain itu *image captioning* juga dapat dilakukan pada domain spesifik seperti pada gambar artikel berita (Biten dkk., 2019) dan *remote sensing* (Hoxha dkk., 2020).

Pada domain *e-commerce*, *image captioning* dapat digunakan untuk menghasilkan judul yang tepat untuk gambar produk. Pada domain tersebut caption memiliki karakteristik khusus seperti: umumnya tersusun sebagai frasa nomina, memiliki spesifikasi merek, memiliki spesifikasi variasi dan model, dan juga memiliki spesifikasi warna. Karakteristik tersebut penting untuk muncul dalam judul produk agar dapat mendeskripsikan gambar produk dengan tepat.

Image captioning juga dapat dimengerti sebagai translasi gambar menjadi teks. Berdasarkan konsep tersebut, arsitektur model *image captioning* dapat diadaptasi dari model *machine translation* dengan mengubah *encoder* untuk teks menjadi *encoder* untuk gambar. Salah satu model *image captioning* yang sederhana yaitu menggunakan arsitektur CNN-LSTM (Vinyals dkk., 2015). Komponen CNN berfungsi untuk mengekstraksi fitur dari gambar dan merepresentasikannya sebagai suatu vektor. Komponen LSTM berfungsi untuk menerima vektor representasi gambar dan menghasilkan *caption* yang sesuai dengan gambar yang diberikan. Namun, model tersebut tidak mampu mengidentifikasi fitur spesifik dari gambar atau pun *caption*. Hal tersebut mengakibatkan *caption* yang dihasilkan tidak dapat mendeskripsikan gambar secara detail.

Pada perkembangannya, model *image captioning* menggunakan berbagai pendekatan agar dapat mengekstraksi fitur spesifik baik dari gambar maupun dari *caption*. Xu dkk. (2015) menggunakan *visual attention* pada komponen CNN sebagai masukan untuk LSTM *decoder*. Gan dkk. (2017) menggunakan caption

untuk memperoleh konsep semantik dari masukan gambar yang kemudian digunakan sebagai input tambahan pada LSTM *decoder*. Melanjutkan penelitian Gan dkk. (2017), Kim dkk. (2018) mengembangkan teknik serupa dengan menggunakan tf-idf untuk memberikan bobot yang lebih sesuai pada konsep semantik yang diprediksi.

Penggunaan *attention* seperti pada penelitian Xu dkk. (2015) juga menjadi teknik yang banyak digunakan pada *task* pemrosesan bahasa alami (NLP) seperti *machine translation*. Model-model yang menjadi *state-of-the-art* pada *task machine translation* tersebut berbasis pada arsitektur transformer (Vaswani dkk., 2017) yang mana menggunakan konsep *attention*. Berdasarkan perkembangan tersebut, beberapa penelitian di luar domain NLP pun menggunakan transformer dan memperoleh hasil yang baik.

Pada *task image classification*, Dosovitskiy dkk. (2020) menggunakan bagian *encoder* transformer sebagai hidden layer dari model. Transformer tersebut menerima input gambar dengan cara membagi gambar tersebut menjadi beberapa bagian. Model yang diberi nama Visual Transformer (ViT) tersebut kemudian menjadi *state-of-the-art* untuk task *image classification.*

Pada *task image captioning,* Yu dkk. (2019) menggunakan transformer sebagai arsitektur dasar model *image captioning* mereka. Mereka melakukan modifikasi pada bagian *encoder* yaitu pada bagian multi head attention. Selain itu, gambar dimasukkan terlebih dahulu pada model object detection sebelum diteruskan ke arsitektur transformer.

Penelitian tesis ini memanfaatkan *Distinctive-attribute Extraction* (Kim dkk., 2018) untuk menyelesaikan masalah pembangkitan judul dari produk pada *e-commerce* dengan masukan gambar produk. Penggunaan *Distinctive-attribute Extraction* (DaE) diharapkan dapat memberikan bobot yang sesuai pada konsep penting di domain produk *e-commerce* seperti merek, variasi, atau model produk. Arsitektur model *image captioning* yang digunakan berbasis pada arsitektur transformer yang

diusulkan Vaswani dkk. (2017). Hal ini didasari pada baiknya hasil yang diperoleh ViT untuk *task image classification*. Berdasarkan konsep *image captioning* yang terdiri atas *visual understanding* dan *caption generation*, maka arsitektur menggunakan ViT sebagai komponen *visual understanding* dengan komponen *caption generation* berupa transformer *decoder*.

I.2 Masalah Penelitian

Masalah yang dibahas pada penelitian tesis ini adalah mengenai bagaimana cara agar model *image captioning* dapat menghasilkan judul produk yang spesifik hingga merek, variasi, atau pun model pada *image captioning* untuk domain produk *e-commerce*.

I.3 Tujuan

Tujuan dari penelitian tesis ini adalah menghasilkan model *image captioning* yang dapat menghasilkan judul produk yang detail hingga merek, variasi, atau pun model sehingga didapatkan skor metrik yang baik.

I.4 Hipotesis

Konsep seperti variasi dan model produk yang spesifik mengakibatkan probabilitasnya untuk dihasilkan pada proses *decoding* menjadi lebih kecil. DaE dapat mengidentifikasi konsep semantik yang spesifik dari teks judul dan memberikan bobot yang sesuai menggunakan tf-idf. Pemberian bobot yang besar pada konsep spesifik dapat meningkatkan probabilitas konsep tersebut untuk dihasilkan model.

I.5 Batasan Masalah

Batasan masalah dalam penelitian tesis ini adalah sebagai berikut.

1. Produk *e-commerce* yang digunakan hanya produk dari kategori *fashion* wanita dari *e-commerce* Zalora.
2. Evaluasi hanya dilakukan secara kuantitatif menggunakan metrik evaluasi untuk *image captioning*.

I.6 Metodologi

Metodologi pengerjaan tesis ini adalah sebagai berikut.

1. Preparasi Dataset

Pada tahap preparasi dataset, dilakukan *crawling* data dari *website e-commerce.* Data berupa pasangan gambar produk dan judul produk dengan kategori *fashion* wanita.

2. Desain Eksperimen

Pada tahap desain eksperimen dilakukan penetuan tujuan, scenario, dan skema eksperimen. Selain itu, dilakukan juga perancangan arsitektur model yang akan diimplementasi berdasarkan pendekatan yang akan digunakan.

3. Implementasi

Pada tahap implementasi, arsitektur model yang sudah dirancang diimplementasi sehingga sesuai dengan desain eksperimen yang sudah didefinisikan sebelumnya.

4. Eksperimen

Pada tahap eksperimen, dilakukan proses pelatihan model dan pengujian terhadap dataset. Hasil dari tahap ini merupakan ukuran performa berupa skor metrik evaluasi.

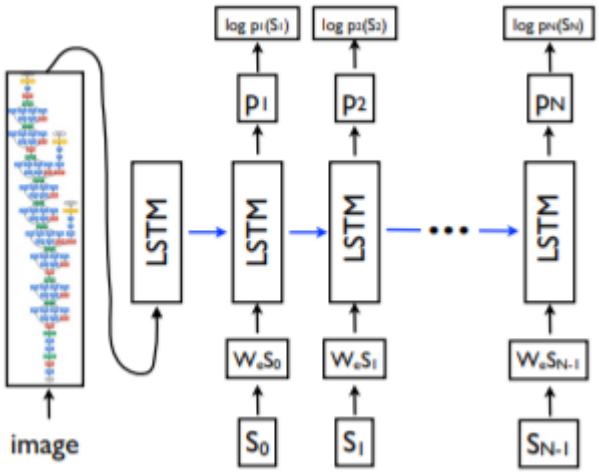
5. Evaluasi

Pada tahap evaluasi, dilakukan perbandingan antara model yang diusulkan penelitian tesis terhadap model *baseline*. Setelah itu dapat ditentukan kelebihan dan kekurangan dari model yang diusulkan agar dapat dikembangkan pada penelitian mendatang.

Bab II Tinjauan Pustaka

II.1 Image Captioning

Image captioning merupakan *task* dengan tujuan menghasilkan teks deskripsi yang menjelaskan masukan gambar. Secara umum terdapat dua komponen dalam model *image captioning* yaitu *visual understanding* dan *caption generation*. Pada *visual understanding*, masukan gambar akan diproses sehingga dapat dikenali oleh model. Masukan gambar yang telah dikenali akan menjadi petunjuk bagi model dalam menghasilkan teks deskripsi yang sesuai. Proses tersebut dilakukan pada komponen *caption generation*.



Gambar II.1 *Neural Image Caption* (Vinyals dkk., 2015)

Image captioning juga dapat dipahami sebagai translasi gambar menjadi teks. Konsep tersebut diadaptasi dari konsep *machine translation* yaitu translasi teks dari bahasa sumber menjadi bahasa target. Berdasarkan hal tersebut, maka model yang umum digunakan pada *task image captioning* pun diadaptasi dari model untuk *machine translation.*

Vinyals dkk. (2015) mengembangkan *Neural Image Caption* (NIC) sebagai model untuk *task image captioning* dengan arsitektur *end-to-end* yang diadaptasi dari model *machine translation*. Model *machine translation* yang menggunakan prinsip

encoder-decoder diadaptasi dengan cara mengubah *encoder* agar dapat menerima masukan gambar.

Tabel II.1 Skor NIC pada MSCOCO (Vinyals dkk., 2015)

Pendekatan	BLEU-4	METEOR	CIDER
NIC	27.7	23.7	85.5
Random	4.6	9.0	5.1
Nearest Neighbor	9.9	15.7	36.5
Human	21.7	25.2	85.4

Tabel II.2 Skor BLEU-1 NIC pada berbagai dataset (Vinyals dkk., 2015)

Pendekatan	Pascal	Flickr30k	Flickr8k	SBU
Im2Text				11
TreeTalk				19
BabyTalk	25			
Tri5Sem			48	
m-RNN		55	58	
MNLM		56	51	
SOTA	25	56	58	19
NIC	59	66	63	28
Human	69	68	70	

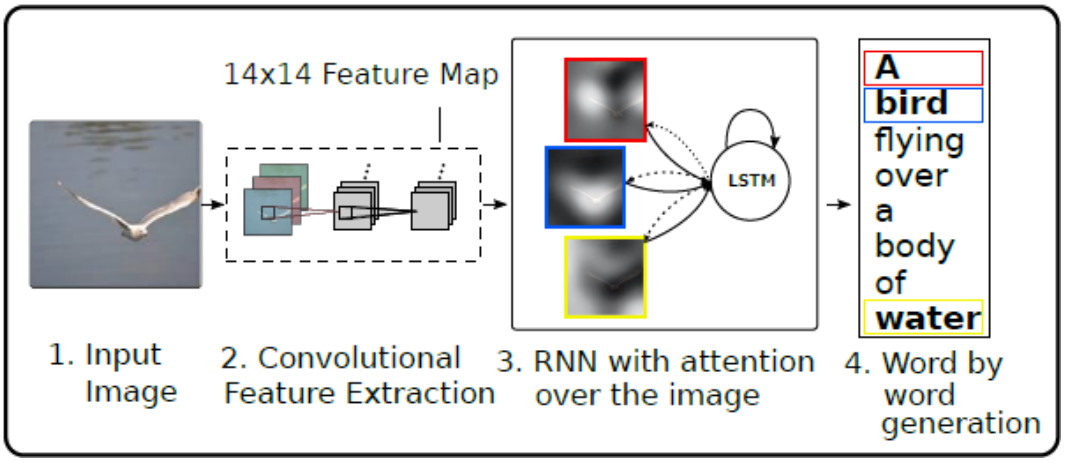
Secara umum NIC terdiri dari komponen *visual understanding* berupa CNN dan komponen *caption generation* berupa LSTM seperti pada Gambar II.1. Masukan gambar diterima oleh CNN sehingga diperoleh vektor representasi dari gambar tersebut. Vektor yang diperoleh kemudian diteruskan ke LSTM sebagai masukan. LSTM kemudian menghasilkan teks deskripsi secara bertahap kata per kata.

NIC menjadi *state-of-the-art* pada saat publikasinya untuk beberapa dataset seperti yang terlihat pada Tabel II.1 dan Tabel II.2. Berdasarkan hasil tersebut, NIC kemudian menjadi model dasar untuk *task image captioning* yang menggunakan

arsitektur *encoder-decoder*. Secara spesifik arsitektur CNN-LSTM milik NIC menjadi arsitektur dasar model *image captioning* yang kemudian banyak dikembangkan di penelitian-penelitian setelahnya.

II.2 Attention-based Image Captioning

Pada perkembangan yang lebih lanjut, *image captioning* menggunakan mekanisme *attention* pada proses pembangkitan *caption.* Mekanisme tersebut bertujuan agar ketika melakukan prediksi kata, model memperhatikan fitur tertentu dari gambar sehingga kata yang dibangkitkan akan berkorelasi dengan fitur tersebut.



Gambar II.2 Arsitektur model NIC dengan *visual attention* (Xu dkk., 2015)

Penelitian Xu dkk. (2015) merupakan salah satu perkembangan model *image captioning* dengan menggunakan mekanisme *attention*. Secara umum, arsitektur dasar yang digunakan adalah arsitektur yang dikenalkan oleh Vinyals dkk. (2015). Namun, terdapat perubahan pada komponen visual CNN yang mana tidak menghasilkan vektor yang merepresentasikan gambar secara keseluruhan melainkan fitur dari region tertentu pada gambar.

Secara umum, model yang diusulkan Xu dkk. (2015) dapat dilihat pada Gambar II.2. Komponen visual berupa CNN digunakan untuk menerima masukan gambar. Kemudian, fitur yang diekstraksi pada layer konvolusi yang lebih rendah digunakan sebagai input *visual attention* pada LSTM. Proses pembangkitan *caption*

menggunakan LSTM secara garis besar sama seperti pada NIC, namun dengan tambahan input *attention* pada setiap *step*.

Tabel II.3 Skor model NIC dengan *attention* (Xu dkk., 2015)

Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Flickr8k	NIC	63	41	27	-	-
	Log Bilinear	65.6	42.4	27.7	17.7	17.31
	Soft-Attention	67	44.8	29.9	19.5	18.93
	Hard-Attention	67	45.7	31.4	21.3	20.30
Flickr30k	NIC	66.3	42.3	27.7	18.3	-
	Log Bilinear	60	38	25.4	17.1	16.88
	Soft-Attention	66.7	43.4	28.8	19.1	18.49
	Hard-Attention	66.9	43.9	29.6	19.9	18.46
COCO	CMU/MS	-	-	-	-	20.41
	MS	-	-	-	-	20.71
	BRNN	64.2	45.1	30.4	20.3	-
	NIC	66.6	46.1	32.9	24.6	-
	Log Bilinear	70.8	48.9	34.4	24.3	20.03
	Soft-Attention	70.7	49.2	34.4	24.3	23.90
	Hard-Attention	71.8	50.4	35.7	25.0	23.04

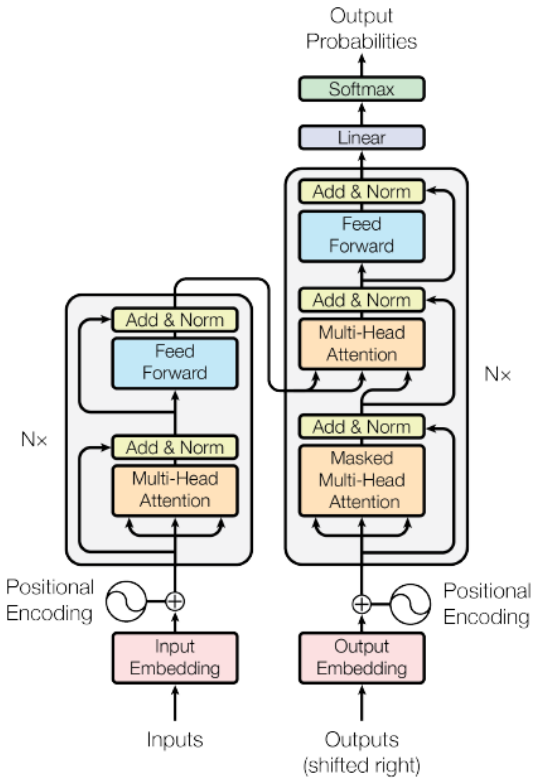
Xu dkk. (2015) mengenalkan dua pendekatan *attention* untuk modelnya yaitu *hard attention* dan *soft attention.* Secara garis besar pada *hard attention* model akan

memilih fitur gambar mana yang akan digunakan sebagai masukan *visual attention* pada *step* tertentu untuk menghasilkan kata pada *step* tersebut. Adapun pada *soft attention* dilakukan pembobotan pada setiap fitur gambar untuk kemudian digunakan sebagai masukan *visual attention.*

Hasil dari pendekatan yang digunakan oleh Xu dkk. (2015) yaitu performa yang lebih baik dibandingkan NIC. Hal tersebut berdasarkan skor metrik yang dapat dilihat pada Tabel II.3. Baik *hard attention* maupun *soft attention* memberikan skor metrik yang lebih baik dibandingkan NIC. Hal ini menunjukkan adanya pengaruh yang signifikan atas penggunaan *attention* pada model *image captioning.*

II.3 Transformer

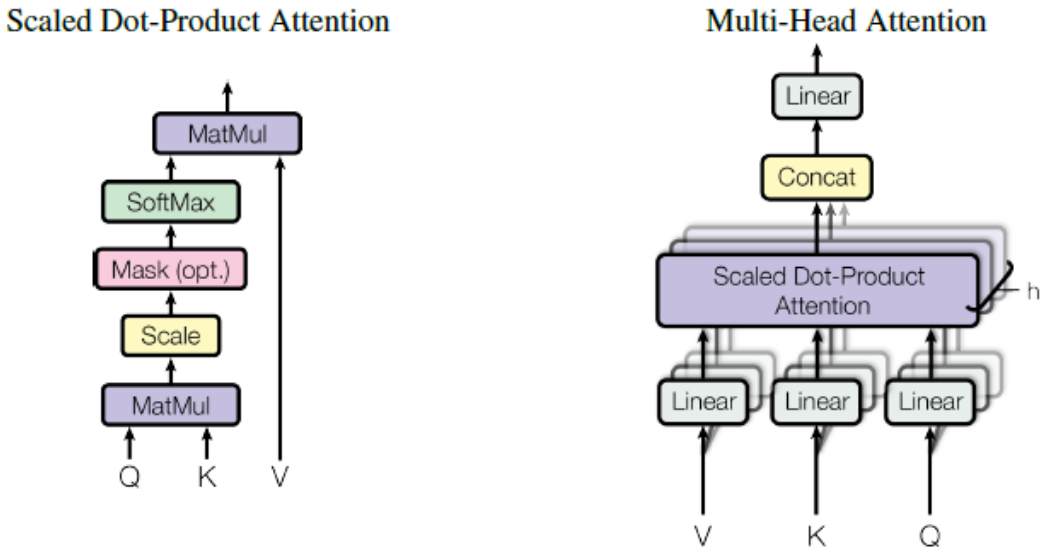
Mekanisme *attention* umum digunakan pada bidang NLP, khususnya pada *task machine translation*. Salah satu arsitektur model *machine translation* yang hingga kini menjadi *state-of-the-art* yaitu Transformer (Vaswani dkk., 2017).



Gambar II.3 Arsitektur transformer (Vaswani dkk., 2017)

Transformer menggunakan suatu mekanisme *attention* pada arsitekturnya yang diberi nama *self-attention*. Pada transformer mekanisme *attention* tidak hanya diterapkan antara keluaran terhadap masukannya saja, melainkan juga antar bagian dari masukan dan antar bagian dari keluaran.

Secara umum arsitektur dari transformer seperti yang terlihat pada Gambar II.3. Transformer terdiri dari dua bagian yaitu bagian *encoder* dan *decoder*. Masing-masing bagian menggunakan mekanisme *self-attention* pada arsitekturnya. Transformer bukanlah *sequence model* sebagaiman *state-of-the-art* sebelumnya. Semantik mengenai posisi kata dalam kalimat direpresentasikan menggunakan suatu *positional encoding*. Baik *encoder* atau pun *decoder*, masing-masing bagian dapat ditumpuk hingga beberapa lapis.



Gambar II.4 Proses komputasi *attention* pada transformer (Vaswani dkk., 2017)

Positional *encoding* menggunakan suatu fungsi sin dan cos dalam berbagai frekuensi berbeda. Nilai dari fungsi tersebut dijumlahkan pada vektor masukan di bagian *encoder* dan *decoder* pada transformer. Hasil penjumlahan vekor tersebut yang menjadi masukan pada bagian *encoder* dan *decoder*.

Komponen *self-attention* pada transformer tersusun atas beberapa proses komputasi yang dapat dilihat pada Gambar II.4. Relevansi antar kata ditentukan melalui suatu proses komputasi *scaled dot product attention*. Terdapat sejumlah komputasi *scaled dot product attention* berbeda yang dilakukan secara parallel. Jumlah tersebut berdasarkan seberapa banyak semantik yang model perlu pelajari. Hasil dari masing-masing komputasi tersebut disambung menjadi satu vektor sebelum diteruskan ke sebuah fungsi linear. Keseluruhan proses tersebut dinamakan *multi-head attention*.

Transformer memperoleh hasil yang lebih baik dibandingkan *state-of-the-art* sebelumya untuk *task machine translation* pada dataset *English-to-German* dan *English-to-French* seperti yang terlihat pada Tabel II.4. Berdasarkan hasil tersebut, transformer kemudian menjadi arsitektur dasar dari berbagai penelitian setelahnya di bidang NLP.

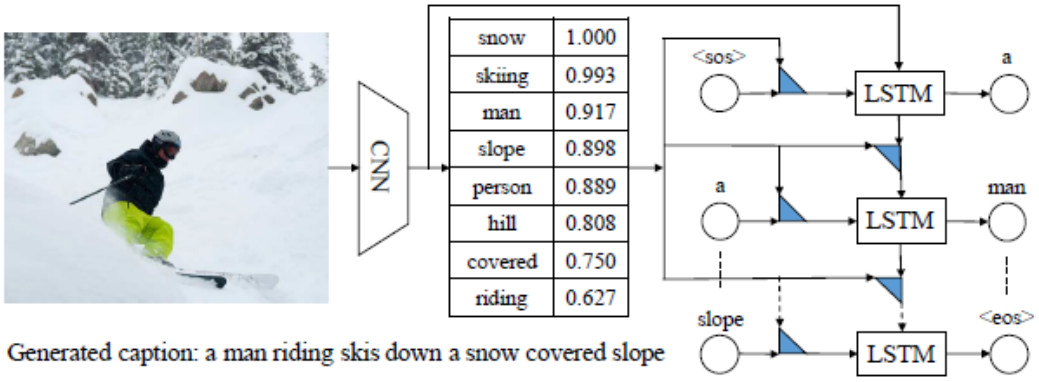
Tabel II.4 Skor model transformer (Vaswani dkk., 2017)

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet	23.75			
DeepAtt+PosUnk		39.2		1.0×10^{20}
GNMT+RL	24.6	39.92	2.3×10^{19}	1.4×10^{20}
ConvS2S	25.16	40.46	9.6×10^{18}	1.5×10^{20}
MoE	26.03	40.56	2.0×10^{19}	1.2×10^{20}
DeepAtt+PosUnk Ensemble		40.4		8.0×10^{20}
GNMT+RL Ensemble	26.30	41.16	1.8×10^{20}	1.1×10^{21}
ConvS2S Ensemble	26.36	41.29	7.7×10^{19}	1.2×10^{21}
Transformer(base)	27.3	38.1	3.3×10^{18}	
Transformer(big)	28.4	41.8	2.3×10^{19}	

II.4 Ekstraksi Atribut untuk Image Captioning

Penelitian terkait *image captioning* terus berkembang hingga melakukan ekstraksi atribut/semantik pada gambar. Pada implementasinya masukan gambar akan diproses terlebih dahulu untuk diperoleh atributnya. Atribut yang berhasil didapatkan kemudian digunakan sebagai masukan komponen *caption generation*. Atribut tersebut dapat menjadi masukan utama menggantikan vektor representasi gambar atau menjadi masukan tambahan di luar arsitektur umum model *image captioning*.

Gan dkk. (2017) mengusulkan suatu pendekatan menggunakan Semantic Compositional Network (SCN) untuk melakukan *visual captioning.* Masukan gambar diterima oleh SCN terlebih dulu untuk memperoleh semantik berupa *tag* apa saja yang terdapat pada gambar. Serupa dengan *soft-attention* yang dijelaskan oleh Xu dkk. (2015), vektor semantik tersebut digunakan sebagai masukan tambahan pada setiap *step* pembangkitan kata di LSTM dengan menggunakan pembobotan untuk setiap *tag* yang terdapat pada vektor.



Gambar II.5 Arsitektur SCN-LSTM (Gan dkk., 2017)

Secara umum arsitektur SCN-LSTM yang diusulkan oleh Gan dkk. (2017) dapat dilihat pada Gambar II.5. Masukan gambar terlebih dahulu diekstraksi fiturnya menggunakan CNN sehingga diperoleh sebuah vektor representasi dari gambar. Vektor tersebut digunakan sebagai masukan untuk LSTM dan untuk SCN.

Keluaran dari SCN diteruskan ke LSTM sebagai masukan tambahan di setiap *step*. Pada LSTM, *caption* dihasilkan kata per kata berdasarkan masukan gambar di awal dan semantik di setiap *step.*

Proses ekstraksi semantik pada SCN dianggap sebagai *multi-label classification*. Label untuk SCN diperoleh dari *caption* di keseluruhan dataset dengan mengambil sejumlah kata yang paling sering muncul. SCN tersusun atas *multi-layer perceptron* (MLP) dengan fungsi aktivasi sigmoid. Keluaran dari SCN berupa vektor berukuran *tag* terdefinisi dengan nilai masing-masing elemen berupa probabilitas *tag* tersebut terdapat pada gambar. Vektor tersebut kemudian dioperasikan dengan menggunakan operator Hadamard bersama dengan kata ke-i menggunakan pembobotan.

Tabel II.5 Skor SCN-LSTM pada dataset COCO (Gan dkk., 2017)

Metode	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDER
NIC	0.666	0.451	0.304	0.203	-	-
m-RNN	0.67	0.49	0.35	0.25	-	-
Hard-Attention	0.718	0.504	0.357	0.250	0.230	-
ATT	0.709	0.537	0.402	0.304	0.243	-
Att-CNN+LSTM	0.74	0.56	0.42	0.31	0.26	0.94
LSTM-R	0.698	0.525	0.390	0.292	0.238	0.889
LSTM-T	0.716	0.546	0.411	0.312	0.250	0.952
LSTM-RT	0.724	0.555	0.419	0.316	0.252	0.970
LSTM-RT_2	0.730	0.568	0.430	0.322	0.249	0.977
SCN-LSTM	0.728	0.566	0.433	0.330	0.257	1.012
SCN-LSTM Ensemble 5	0.741	0.578	0.444	0.341	0.261	1.041

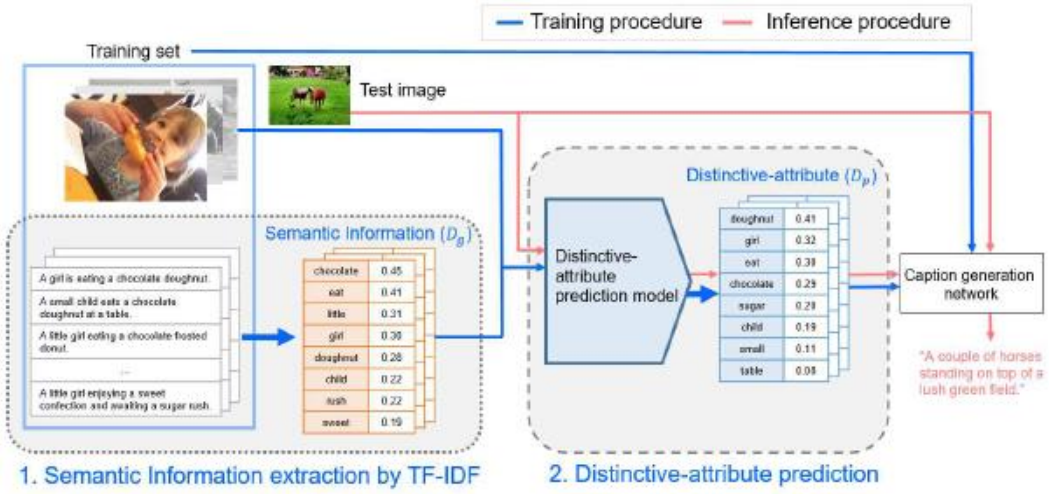
SCN-LSTM memperoleh hasil yang baik berdasarkan skor pada Tabel II.5. Ekstraksi semantik yang dilakukan oleh SCN mengakibatkan kata-kata yang lebih

tepat dapat dihasilkan. Hal tersebut terlihat pada skor SCN-LSTM untuk metrik CIDER yang lebih tinggi dibandingkan model yang lain. Hasil tersebut menunjukan pemberian informasi dengan level yang lebih tinggi secara eksplisit terkait masukan gambar dapat meningkatkan kualitas *caption* yang dihasilkan.

II.5 Penelitian Terkait

II.5.1 Kim dkk. (2018)

Kim dkk. (2018) mengusulkan suatu metode *image captioning* dengan ekstraksi atribut yang berbeda dengan yang diusulkan oleh Gan dkk. (2017). Metode tersebut diberi nama *Distinctive-attribute Extraction* (DaE). Perbedaan utama antara DaE dengan SCN terletak pada nilai yang digunakan oleh vektor atribut. Pada SCN nilai vektor atribut berupa probabilitas atribut ada pada masukan gambar, sementara pada DaE nilai vektor berupa tf-idf dari atribut tersebut.



Gambar II.6 Arsitektur DaE (Kim dkk., 2018)

Fokus utama dari penelitian Kim dkk. (2018) yaitu mengenai metode ekstraksi atribut, sehingga tidak ada pembahasan secara khusus terkait model *image captioning*. Metode tersebut dikatakan dapat ditambahkan ke berbagai model *image captioning*, sehingga membuat metode tersebut menjadi cukup fleksibel untuk digunakan.

Secara umum arsitektur DaE seperti yang terlihat pada Gambar II.6. *Caption* dari data latih digunakan untuk mengekstraksi informasi semantik menggunakan tf-idf. Hasil dari proses ekstraksi tersebut berupa vektor berukuran *vocabulary* yang masing-masing nilai elemennya berupa nilai tf-idf dari kata pada *caption* tersebut. Informasi semantik tersebut digunakan sebagai *ground-truth* dari gambar terkait. Model *distinctive-attribute prediction* kemudian dibangun menggunakan pasangan gambar dan atribut yang sudah diperoleh. Atribut yang diprediksi oleh model diteruskan ke model *image captioning* yang digunakan. Adapun pada eksperimennya, Kim dkk. (2018) menggunakan model *image captioning* SCN-LSTM yang diusulkan oleh Gan dkk. (2017).

Tabel II.6 Skor DaE pada dataset COCO (Kim dkk., 2018)

	B-1	B-2	B-3	B-4	M	R	CIDEr
5-refs							
SCN	0.729	0.563	0.426	0.324	0.253	0.537	0.967
DaE+SCN-LSTM	0.734	0.568	0.429	0.324	0.255	0.538	0.981
40-refs							
SCN	0.910	0.829	0.727	0.619	0.344	0.690	0.971
DaE+SCN-LSTM	0.916	0.836	0.734	0.625	0.348	0.694	0.990

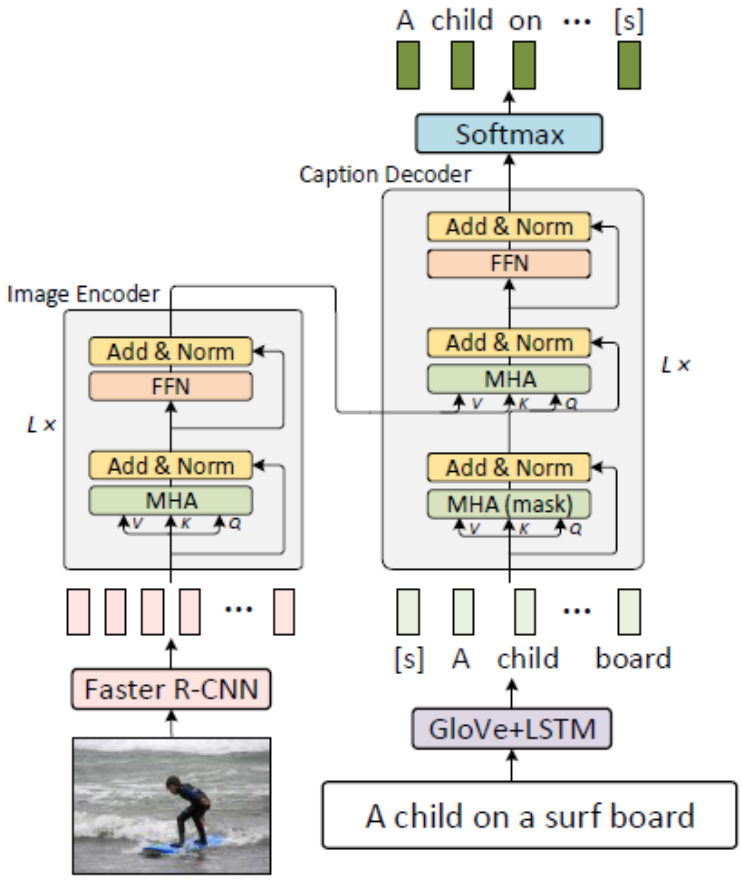
Arsitektur model *distinctive-attribute prediction* yang diusulkan oleh Kim dkk. (2018) terdiri atas suatu CNN dan empat FC *layer*. CNN yang digunakan pada penelitian tersebut yaitu ResNet152. Keluaran dari CNN berupa vektor berukuran 2048 hasil *global pooling* di *layer* konvolusi terakhir. Vektor tersebut diteruskan ke FC *layer*. Tiga FC *layer* pertama mempunyai 2048 simpul masing-masing, sementara di *layer* terakhir mempunyai simpul sejumlah atribut yang terdefinisi. Fungsi aktivasi untuk setiap FC *layer* menggunakan *relu*.

Hasil penelitian Kim dkk. (2018) dapat dilihat pada Tabel II.6. Pada tabel ditunjukkan bahwa penambahan DaE pada SCN-LSTM dapat meningkatkan

performa model. Khususnya peningkatan skor pada metrik CIDEr menunjukkan bahwa model berhasil memperoleh kata dengan semantik yang lebih tepat.

II.5.2 Yu dkk. (2019)

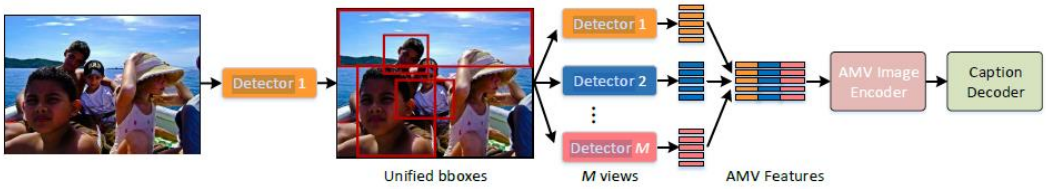
Penelitian Yu dkk. (2019) membahas penggunaan transformer dalam *task* image captioning yang disebut sebagai Multimodal Transformer. Sebagai tambahan, mekanisme *multi-view* terhadap masukan gambar juga dilakukan pada penelitian tersebut. Model *object detection* digunakan untuk menerima masukan gambar sehingga diperoleh objek-objek yang ada pada gambar. Objek tersebut yang menjadi masukan pada transformer untuk menghasilkan *caption* dari gambar.



Gambar II.7 Arsitektur transformer untuk *image captioning* (Yu dkk., 2019)

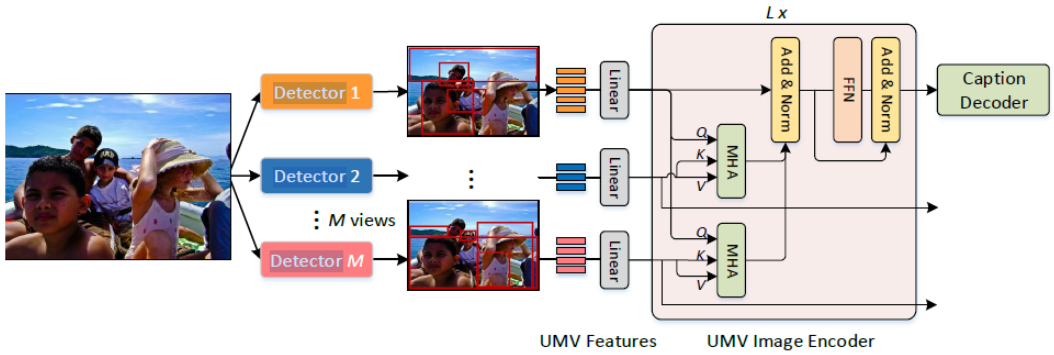
Secara umum arsitektur dari model yang diusulkan oleh Yu dkk. (2019) dapat dilihat pada Gambar II.7. Masukan gambar diproses terlebih dahulu oleh model *object detection*. Pada eksperimen, secara spesifik model *object detection* yang

dipilih yaitu Faster R-CNN. Keluaran berupa objek yang dideteksi berikut *confidence score* dari masing masing objek tersebut. Kemudian dipilih sejumlah objek dengan *confidence score* tertinggi. Region dari masing-masing objek tersebut kemudian melalui konvolusi dan *mean-pooling* sebelum akhirnya diperoleh matriks yang tersusun dari vektor objek tersebut. Matriks tersebut menjadi masukan pada transformer *encoder*. Pada transformer *decoder*, *ground-truth caption* direpresentasikan menggunakan *GloVe* dan *layer* LSTM untuk memberikan semantik urutan kata. *Sequence* yang diperoleh kemudian diteruskan ke transformer *decoder*.



Gambar II.8 Skema AMV (Yu dkk., 2019)

Pada *encoder* terdapat dua pendekatan *multi-view* dalam merepresentasikan objek ke dalam transformer yaitu *aligned multi-view* (AMV) dan *unaligned multi-view* (UMV). Pada AMV, objek dideteksi menggunakan suatu model *object detector*. Kemudian digunakan sejumlah model *object detector* termasuk model yang digunakan di awal untuk memperoleh fitur dari setiap objek yang dideteksi. Fitur yang diperoleh dari masing-masing detector kemudian disambungkan dan menjadi masukan pada *transformer encoder*. Proses tersebut dapat dilihat pada Gambar II.8.



Gambar II.9 Skema UMV (Yu dkk., 2019)

Adapun pada UMV, masukan gambar dideteksi objeknya oleh sejumlah *object detector* berbeda. Masing-masing *detector* melakukan ekstraksi fitur untuk objek yang berhasil dideteksi. Fitur yang berhasil dideteksi oleh masing-masing *detector* kemudian diproyeksi menjadi suatu vektor dengan dimensi yang sama. Vektor-vektor tersebut kemudian menjadi masukan pada komponen *multi-head attention* di transformer *encoder*. Proses tersebut dapat dilihat pada Gambar II.9.

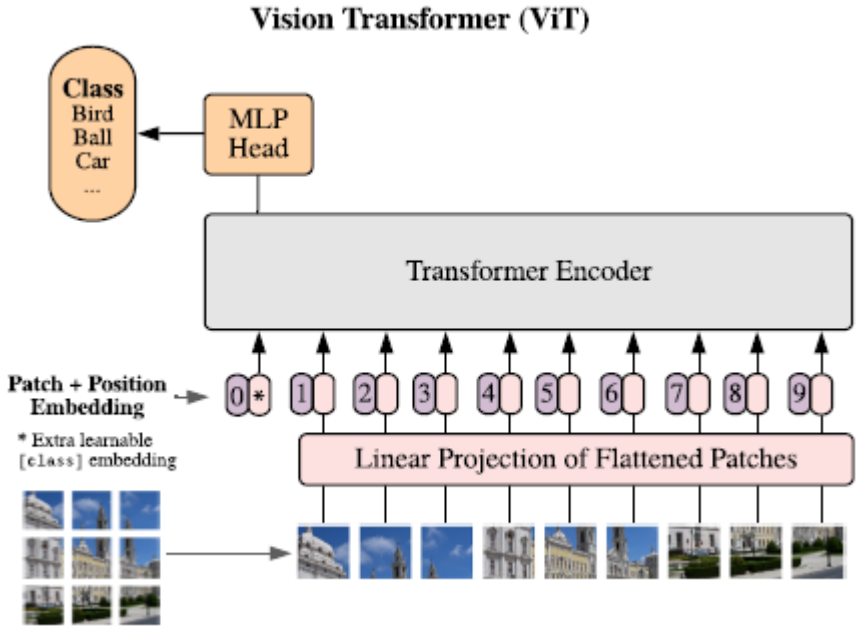
Tabel II.7 Skor MT pada dataset Flickr30k (Yu dkk., 2019)

Metode	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
NIC	66.3	42.3	27.7	18.3	-
Soft-Att	66.7	43.4	28.8	19.1	-
Hard-Att	66.9	43.9	29.6	19.9	-
SAS-RE	66.3	44.3	30.5	21.1	18.6
NeuralTalk2-T-oe	64.6	43.8	31.9	22.4	19.2
Att-RegionCNN+LSTM	73.0	55.0	40.0	28.0	-
MT_{sv} (R-101)	74.4	57.5	43.4	32.5	23.6
MT_{sv} (R-152)	74.6	57.7	43.6	32.8	24.1
MT_{amv} (R-101, R-152)	75.5	58.5	44.0	33.2	24.2
MT_{umv} (R-101, R-152)	75.6	58.6	44.3	33.3	24.3
MT_{umv} (R-101, R-152, X-101)	75.8	58.7	44.5	33.3	24.5

Hasil skor model yang diusulkan oleh Yu dkk., (2019) dapat dilihat pada Tabel II.7. Berdasarkan hasil tersebut, ditunjuuakn bahwa penggunaan transformer untuk *task image captioning* dapat menghasilkan *caption* yang lebih baik secara general. Kemudian penggunaan UMV memberikan hasil yang lebih baik dibandingkan dengan AMV maupun SV. Selain itu, penggunaan jumlah *detector* yang lebih banyak dapat memberikan hasil yang lebih optimal.

II.5.3 Dosovitskiy dkk. (2020)

Penelitian Dosovitskiy dkk. (2020) mengenai penggunaan transformer untuk *task image recognition*. Teknik yang diusulkan mengenai perepresentasian masukan gambar agar dapat diterima oleh arsitektur transformer. Adapun arsitektur transformer yang digunakan hanya bagian *encoder* saja. Model yang diusulkan diberi nama Vision Transformer (ViT).



Gambar II.10 Arsitektur ViT (Dosovitsky dkk., 2020)

Secara umum arsitektur ViT dapat dilihat pada Gambar II.10. Masukan gambar dipecah menjadi beberapa bagian terlebih dahulu. Kemudian masing-masing bagian melalui proses *flatten* dan diproyeksi menggunakan suatu fungsi linear yang dapat dilatih menjadi suatu vektor dengan ukuran tertentu. Vektor tersebut bersama dengan suatu *classification token* sebagai masukan posisi nol yang dapat dipelajari model menjadi masukan untuk transformer encoder. Proses berikutnya sebagaimana *encoder* pada transformer umumnya bekerja. Keluaran dari *encoder* transformer diteruskan menuju sebuah MLP dengan satu *hidden layer* pada saat *pre-train* dan menuju sebuah *linear layer* pada saat *fine-tuning*.

Pada penelitiannya ViT dibangun menjadi beberapa varian berdasarkan ukuran model dan ukuran *patch* gambar yang digunakan. Selain itu dilakukan juga beberapa eksperimen *pre-training* dengan menggunakan dataset berbeda yaitu JFT-300M dan ImageNet-21k. Kemudian performa model diukur pada dataset lain dengan *fine-tuning*. Hasil pengukuran performa model dapat dilihat pada Tabel II.8.

Tabel II.8 Skor akurasi ViT pada berbagai dataset (Dosovitsky dkk., 2020)

	JFT (ViT-H/14)	JFT (ViT-L/16)	ImgNet-21k (ViT-L/16)	BiT-L (ResNet152×4)
ImageNet	88.55 ± 0.04	87.76 ± 0.03	85.30 ± 0.02	87.54 ± 0.02
ImageNet ReaL	90.72 ± 0.05	90.54 ± 0.03	88.62 ± 0.05	90.54
CIFAR-10	99.50 ± 0.06	99.42 ± 0.03	99.15 ± 0.03	99.37 ± 0.06
CIFAR-100	94.55 ± 0.04	93.90 ± 0.05	93.25 ± 0.05	93.51 ± 0.08
Oxford-IIIT Pets	97.56 ± 0.03	97.32 ± 0.11	94.67 ± 0.15	96.62 ± 0.23
Oxford Flowers-102	99.68 ± 0.02	99.74 ± 0.00	99.61 ± 0.02	99.63 ± 0.03
VTAB (19 tasks)	77.63 ± 0.23	76.28 ± 0.46	72.72 ± 0.21	76.29 ± 1.70

ViT sebagai model *image recognition* yang mempunyai performa sangat baik mempunyai potensi untuk diimplementasikan pada *task* terkait lainnya. Salah satunya yaitu *image captioning* yang memerlukan komponen *visual understanding*. ViT dapat diimplementasikan pada *task image captioning* dengan cara menghilangkan bagian klasifikasi sehingga hanya terdiri dari bagian proyeksi *patch* gambar dan *encoder* transformer. Secara intuitif, komponen *caption generation* dapat menggunakan *decoder* transformer untuk menerima keluaran dari ViT.

Bab III Analisis Masalah dan Rancangan Solusi

III.1 Analisis Masalah

Masalah utama dari penelitian tesis ini yaitu *image captioning* pada domain produk *e-commerce*. Secara lebih spesifik, masalah yang diselesaikan yaitu mengenai bagaimana memperoleh judul produk yang detail hingga merek, variasi, atau model.

NIC
lois jeans original celana panjang wanita skinny fsw 219 a

Ground Truth
MKY Clothing Elastic Waist Culotte Jeans in Blue

Gambar III.1 Judul produk hasil NIC

Pada penelitian sebelumnya, NIC terbukti dapat digunakan pada domain produk *e-commerce* untuk menghasilkan judul produk. Namun seperti yang terlihat pada Gambar III.1, judul yang dihasilkan oleh NIC masih kurang baik. Secara umum judul yang dihasilkan sudah sesuai dengan produknya, namun untuk detail spesifik seperti merek dan model masih belum tepat.

Hal tersebut disebabkan karena NIC hanya menggunakan informasi berupa fitur yang merepresentasikan gambar secara utuh sehingga judul yang dihasilkan cenderung fokus terhadap objek apa yang ada pada gambar. Pendekatan yang

dilakukan oleh NIC tidak mampu menangkap informasi detail pada gambar. Kata pada judul yang dihasilkan hanya bergantung pada kata sebelumnya dan masukan gambar di awal saja.

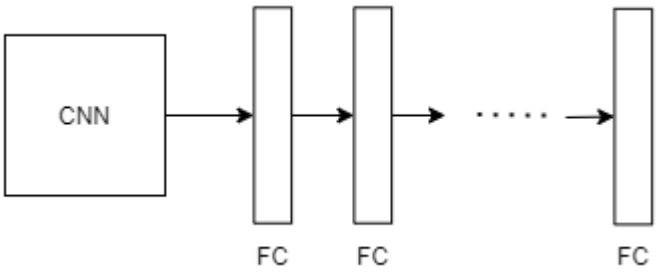
Selain pada pendekatan yang digunakan, dataset yang digunakan pada penelitian sebelumnya masih kurang baik. Dataset diperoleh melalui *e-commerce* dengan syarat penjual yang cukup fleksibel sehingga judul yang diberikan pun relatif tidak terstruktur dan bertele-tele. Dataset yang kurang baik menghasilkan model yang kurang baik pula sehingga diperlukan dataset dengan data yang lebih baik terutama judul produk yang lebih jelas.

Pada penelitian sebelumnya, fokus terkait detail spesifik dari judul produk tidak begitu diperhatikan sehingga evaluasi hanya berdasarkan analisis pada skor BLEU saja. Secara umum, BLEU hanya memberikan gambaran tentang seberapa sama *caption* yang dihasilkan model dengan *ground-truth*. Namun, pada penelitian tesis ini diperlukan metrik yang dapat mengevaluasi detail spesifik pada judul produk. Berdasarkan pertimbangan tersebut, maka metrik yang digunakan pada penelitian tesis ini yaitu BLEU untuk mengevaluasi judul secara umum dan metrik lain yang dapat memberikan bobot lebih pada istilah spesifik yang terdapat pada judul.

III.2 Rancangan Solusi

Dataset diperoleh dengan cara melakukan *crawling* data pada *website e-commerce* Zalora berbeda dengan peneletian sebelumnya yang memperoleh data dari *website* Bukalapak. Zalora dipilih karena beberapa alasan. Pertama, Zalora merupakan *e-commerce* khusus produk *fashion* sehingga sesuai dengan lingkup tesis. Spesialisasi yang dimiliki Zalora tersebut membuat lingkup produk lebih terbatas sehingga tidak akan ada produk di luar *fashion* yang terselip masuk ke dalam dataset. Kedua, persyaratan untuk menjadi penjual di Zalora cukup ketat jika dibandingkan dengan Bukalapak. Pada Bukalapak semua yang mempunyai akun bisa berjualan, sementara pada Zalora ada persayaratan terkait *brand,* produksi, toko, hingga jumlah penjualan per bulan. Penjual resmi yang terdapat pada Zalora membuat

kualitas informasi dari produk yang dijual baik gambar maupun judul menjadi lebih terjamin karena dilakukan secara profesional.



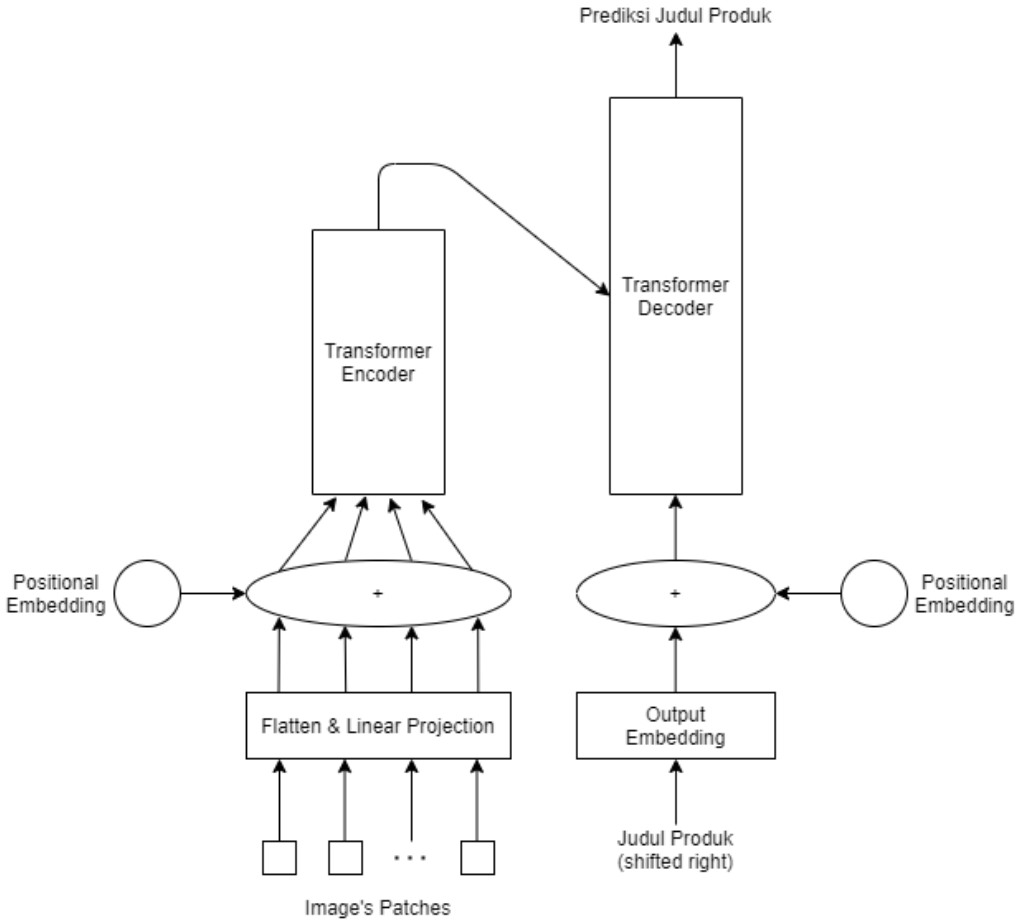
Gambar III.2 Arsitektur model klasifikasi atribut

Pendekatan yang digunakan pada penelitian tesis ini yaitu menggunakan DaE (Kim dkk., 2018). Berdasarkan hal tersebut, solusi terdiri dari dua modul utama. Modul pertama terkait ekstraksi atribut dari gambar. Pada modul tersebut dilatih suatu model untuk melakukan prediksi atribut dari masukan gambar. Kelas atribut pertama-tama diperoleh terlebih dahulu dengan cara melakukan *stemming* pada seluruh kata yang ada pada dataset. Kemudian dipilih kata dengan nilai idf yang lebih besar dari suatu *threshold* tertentu. *Ground-truth* untuk setiap gambar diperoleh dengan memberikan nilai untuk setiap kelas atribut berdasarkan nilai tf-idf kata pada judul terkait. Model terdiri dari suatu jaringan CNN dengan *classification layer* yang terdiri dari bebebrapa FC *layer* seperti pada Gambar III.2.

Modul kedua terkait pembangkitan judul produk berdasasarkan masukan gambar dan atribut yang diperoleh dari modul pertama. Pada modul tersebut model dilatih menggunakan data latih yang terdiri dari gambar produk, judul produk, dan atribut yang diperoleh oleh modul pertama. Secara umum arsitektur model menggunakan transformer seperti pada Gambar III.3. Secara khusus *encoder* menggunakan arsitektur ViT tanpa bagian klasifikasi menggunakan MLP untuk menerima masukan gambar.

Evaluasi dilakukan secara kuantitatif menggunakan metrik *image captioning* yaitu BLEU dan CIDER. BLEU digunakan untuk mengevaluasi seberapa mirip judul

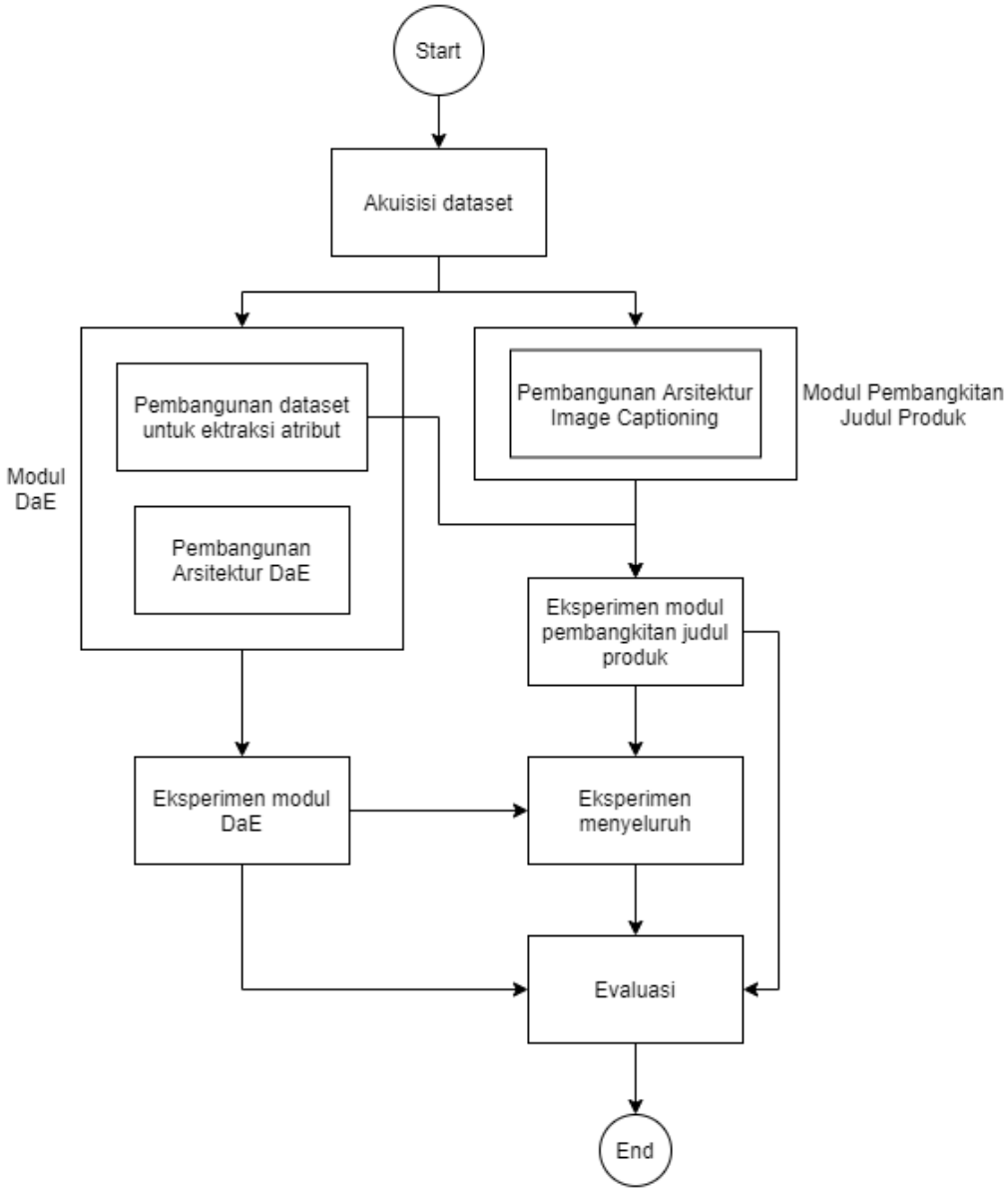
yang dihasilkan model dengan *ground-truth.* CIDER digunakan untuk mengevaluasi apakah konsep spesifik seperti merek, variasi, atau model berhasil dibangkitkan secara tepat atau tidak. CIDER menghitung skor dengan mempertimbangkan nilai idf kata sehingga tepat untuk mengevaluasi konsep spesifik tersebut.



Gambar III.3 Atsitektur model *image captioning*

III.3 Alur Pembangunan Solusi

Secara umum alur pembangunan solusi seperti yang terlihat pada Gambar III.4. Pembangunan solusi diawali dengan akuisisi dataset, dilanjutkan dengan pembangunan modul DaE, pembangunan modul pembangkitan judul produk, eksperimen, dan evaluasi.



Gambar III.4 Alur pembangunan solusi

III.3.1 Akuisisi Dataset

Dataset diperoleh dengan cara melakukan *crawling* dari *website* Zalora. *Crawling* dilakukan per sub-kategori *fashion* wanita yang disediakan oleh *website*. Data yang diperoleh yaitu gambar, merek, dan judul produk. Judul produk yang sebenarnya diperoleh dengan menambahkan merek di depan judul produk. Jika judul produk sudah mengandung merek sebelum merek ditambahkan, maka merek pada judul tersebut dihapuskan agar tidak mengulang merek yang akan ditambahkan di awal

judul produk. Dataset yang digunakan berupa pasangan gambar dan judul produk akhir.

III.3.2 Modul DaE

Terlebih dahulu dibuat dataset dengan judul produk hasil *stemming*. Kemudian dibangun kelas atribut berupa setiap kata pada dataset tersebut yang memiliki nilai idf melebihi threshold tertentu. Setiap data pada dataset kemudian diberi label berupa nilai tf-idf dari setiap atribut yang terdapat pada judul produk. Kemudian dilakukan proses pelatihan model menggunakan dataset berupa pasangan gambar dan label tersebut.

III.3.3 Modul Pembangkitan Judul Produk

Vocabulary terlebih dahulu dibangun menggunakan kata-kata yang terdapat pada seluruh judul produk. Modul ini membutuhkan modul DaE yang sudah dilatih sebelumnya pada aplikasinya. Dataset yang digunakan berupa pasangan gambar, *ground truth* pada modul satu dan judul produk tanpa *stemming*. Model dilatih menggunakan dataset tersebut.

III.3.4 Eksperimen

Eksperimen dilakukan untuk setiap modul. Pada modul pertama, secara umum eksperimen dilakukan untuk menemukan *hyperparameter* terbaik. Secara khusus dilakukan eksperimen pemilihan arsitektur CNN terbaik untuk ekstraksi fitur dari gambar dan pemilihan jumlah *layer* dan simpul pada bagian FC *layer* untuk mendapatkan prediksi atribut.

Sama halnya dengan eksperimen pertama, pada modul kedua eksperimen dilakukan untuk menemukan *hyperparameter* terbaik. Adapun secara khusus dilakukan pemilihan mekanisme *decoding* pada *decoder* dan juga mengenai bagaimana menambahkan masukan atribut pada setiap proses pembangkitan satu kata. Selain itu juga dilakukan pemilihan arsitektur CNN terbaik untuk ekstraksi fitur dari gambar.

III.3.5 Evaluasi

Evaluasi dilakukan pada masing-masing modul dan pada usulan solusi secara menyeluruh. Pada modul pertama, evaluasi dilakukan dengan cara melakukan *testing* model pada data *test.* Masukan berupa gambar dan keluaran berupa nilai pada setiap atribut. Pengukuran dilakukan pada hasil prediksi terhadap *ground-truth* berupa nilai tf-idf setiap atribut pada judul terkait.

Pada modul kedua, evaluasi dilakukan menggunakan data *testing* yang merupakan partisi pada dataset. Masukan berupa gambar dan atribut berdasarkan *ground-truth* yang terdapat pada dataset di modul pertama untuk data terkait. Adapun keluaran dari model berupa judul produk untuk gambar masukan. Pengukuran dilakukan pada judul yang dihasilkan terhadap *ground-truth* berupa judul produk sebenarnya.

Evaluasi secara menyeluruh dilakukan dengan menggunakan kedua model secara terurut. Masukan gambar diprediksi terlebih dahulu atributnya menggunakan modul pertama. Keluaran dari modul pertama kemudian menjadi masukan untuk modul kedua bersamaan dengan gambar. Pengukuran dilakukan pada judul yang dihasilkan terhadap judul sebenarnya.

DAFTAR PUSTAKA

Biten, A. F., Gomez, L., Rusinol, M., dan Karatzas, D. (2019): Good News, Everyone! Context driven entity-aware captioning for news images, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, California, U.S., 12466-12475.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., dan Houlsby, N. (2020): An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Gan, Z., Gan, C., He, X., Pu, Y., Tran, K., Gao, J., Carin, L., dan Deng, L. (2017): Semantic compositional networks for visual captioning, *Proceedings of the IEEE conference on computer vision and pattern recognition*, Honolulu, U.S, 5630-5639.

Hoxha, G., Melgani, F., dan Demir, B. (2020): Toward remote sensing image retrieval under a deep image captioning perspective, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, **13**, 4462-4475.

Kim, B., Han Lee, Y., Jung, H., dan Cho, C. (2018): Distinctive-attribute extraction for image captioning, *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Jerman.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., dan Polosukhin, I. (2017): Attention is all you need, *Advances in neural information processing systems*, California, U.S., 5998-6008.

Vinyals, O., Toshev, A., Bengio, S., dan Erhan, D. (2015): Show and tell: A neural image caption generator, *Proceedings of the IEEE conference on computer vision and pattern recognition*, Boston, U.S., 3156-3164.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., dan Bengio, Y. (2015): Show, attend and tell: Neural image caption generation with visual attention, *International conference on machine learning*, Lille, Perancis, 2048-2057.

Yu, J., Li, J., Yu, Z., dan Huang, Q. (2019): Multimodal transformer with multi-view visual representation for image captioning, *IEEE Transactions on Circuits and Systems for Video Technology*.