Jurnal JAMIKA Manajemen Informatika

Jurnal Manajemen Informatika (JAMIKA)
Volume 10 Nomor 2 Edisi Oktober 2020
E ISSN: 2655-6960 | P ISSN: 2088-4125
OJS: https://ojs.unikom.ac.id/index.php/jamika

Image Captioning menurut *Scientific Revolution* Kuhn dan Popper

Agus Nursikuwagus[1], **Rinaldi Munir**[2], **Masayu Layla Khodra**[3]
[1,2,3]Institut Teknologi Bandung
Sekolah Teknik Elektro dan Informatika, ITB, Bandung, Indonesia
e-mail: [1]agusnursikuwagus@email.unikom.ac.id, [2]rinaldi@informatika.org, [3]masayu@stei.itb.ac.id,

Abstrak

Perkembangan untuk memberikan caption pada suatu gambar merupakan suatu ranah perkembangan baru dalam bidang intelejensia buatan. Image captioning merupakan penggabungan dari beberapa bidang seperti computer vision, natural language, dan pembelajaran mesin. Aspek yang menjadi perhatian dalam bidang image captioning ini adalah ketepatan arsitektur neural network yang dimodelkan untuk mendapatkan hasil yang sedekat mungkin dengan ground-thruth yang disampaikan oleh person. Beberapa kajian yang sudah diteliti masih mendapatkan kalimat yang masih jauh dari ground-thruth tersebut. Permasalahan yang dibahas pada umumnya mengenai image captioning adalah image generator dan text generator yaitu penggunaan deep learning seperti CNN dan LSTM untuk menyelesaikan masalah captioning. Hal ini menjadi dasar permasalahan untuk memberikan kontribusi baru dalam bidang image captioning yang meliputi image extractor, text generator, dan evaluator yang bisa digunakan pada model yang diusulkan. Perspektif Kuhn dan Popper dalam hal image captioning, diperoleh bahwa caption dalam bidang geologi sangat diperlukan dan mencapai tahap krisis. Perlu adanya metode usulan baru untuk menyajikan caption untuk citra geologi.

Kata kunci: *Scientific, Revolution, Image, Captioning, Convolution Neural Network (CNN), Long short-term memory (LSTM)*

Abstract

Image captioning is one area in artificial intelligence that elaborates between computer vision and natural language processing. The focus on this process is an architecture neural network that includes many layers to solve the identification object on the image and give the caption. This architecture has a task to display the caption from object detection on one image. This paper explains about the connection between scientific revolution and image captioning. We have conducted the methodology by Kuhn's scientific revolution and relate to Popper's philosophy of science. The result of this paper is that an image captioning is truly science because many improvements from many researchers to find an effective method on the deep learning process. On the philosophy of science, if the phenomena can be falsified, then an image captioning is the science.

Keywords: *Scientific, Revolution, Image, Captioning, Convolution Neural Network (CNN), Long short-term memory (LSTM)*

1. Pendahuluan

Captioning merupakan teks singkat yang diberikan pada citra disuatu buku, majalah, atau surat kabar mengenai deskripsi atau penjelasan mengenai citra tersebut. Deskripsi ini diberikan menurut visualiasi yang dilihat seseorang pada citra tersebut. Text *caption* dikembangkan untuk lebih memberikan pemahaman suatu informasi yang ingin disampaikan pada citra yang ditunjukkan. Pemahaman terhadap citra menjadi fokus utama pada setiap *caption*ing yang diberikan. Srihari (1994) menyatakan bahwa lokalisasi suatu

objek pada citra menjadi sumber informasi dalam pemahaman suatu citra. Pemberian teks dapat membantu lebih mengerti tentang ilustrasi objek-objek yang ada pada suatu citra [1].

Image *caption* generation merupakan masalah bidang komputer yang melibatkan *computer vision, natural language processing (NLP*), dan pembelajaran mesin. Pekerjaan yang dilakukan adalah melakukan translasi image menjadi suatu teks yang sesuai dengan objek termaksud. Pekerjaan *caption*ing merupakan hal ini mudah jika dilakukan oleh manusia, tetapi hal ini menjadi tantangan tersendiri apabila pekerjaan ini dibangun pada suatu mesin. Apalagi mesin yang dibangun bisa memahami isi dari image yang diamati serta hubungannya dengan bahasa sehari-hari. Wang dkk menggunakan CNN untuk pembangkitkan image dan LSTM sebagai model untuk pembangkit teks. Evaluasi dari model learning pembangkit teks, penelitian ini menggunakan BLEU-N dan METEOR. Dataset yang digunakan sebagai eksperimen yaitu Flickr30k, dengan hasil evaluasi 64.5%, 45.8%, 32.2%, 22.4%, dan 19.0% dengan model learning CNN + LSTM. Sedangkan menggunakan MSCOCO, hasil evaluasi 64.5%, 45.8%, 32.2%, 22.4%, dan 19.0% dengan model learning CNN + LSTM yaitu 72.1%, 54.6%, 40.9%, 30.4%, dan 25.1% [2].

Bila ditinjau dari segi objektivitas, tugas dari image *caption*ing dapat berkisar pada deskripsi faktual dari suatu image, termasuk objek yang diamati, pergerakan objek, dan hubungannya. Tetapi pada kebanyakan penelitian, subjek komponen non-factual terhadap penunjukan interpretasi dari image seringkali hilang pada kemunculan bentuk kata sifat dan kata keterangan. Objek ini hilang karena bukan objek utama yang diamati. Peristiwa ini, kemudian diperbaiki oleh [3] dengan membangkitkan kembali teks tersebut. Contoh penelitian dari Andrew Shin adalah mendefinisikan elemen subjektif sebagai suatu label sentimen dari image, dan memodifikasi istilah subjektif sebagai suatu istilah sentimen. Selain elemen faktual, istilah non-faktual sentimen juga ditambahkan sehingga akan lebih memperluas ekspresi, memperkaya estetika bahasa, dan lebih mendekati pernyataan manusia. Andrew Shin, menggunakan dataset Flickr dan DevianArt dengan menggunakan Bleu-1, Bleu-2, Bleu-3, Bleu-4, dan Meteor sebagai evaluasi modelnya. Hasil yang diperoleh adalah 56.5%, 35.9%, 21.7%, 13.0%, dan 11.6% [3].

Selain tugas yang sudah disebutkan, penelitian ini juga melibatkan penggunaan berbagai mesin learning seperti deep neural network (DNN). Penggunaan teknik ini untuk memproduksi kata yang tepat pada image yang sedang diamati. Teknik yang telah banyak digunakan untuk DNN seperti RNNs (recurrent neural nets), Long-Short-Term-Memory (LSTM) component [4]. Penelitian yang diusung oleh [5] juga menjelaskan penggunaan deep learning sebagai solusi untuk memproduksi teks yang tepat untuk suatu image. Dai menggunakan RNN dan LSTM untuk memproduksi teks untuk bisa menunjukan perbandingan dengan human description, walaupun hasilnya masih belum memuaskan. Pada studi yang telah dilakukan bahwa produksi *caption* dari image yang diamati masih jauh dari apa yang diharapkan. Jika menggunakan image lain sebagai pembanding, terkadang ada kemiripan walaupun ada perbedaan dari segi aspek pada setiap objek yang diamati [5]. Tujuan penelitian Dai adalah membuat *caption*ing dengan lebih memperjelas objek yang dideteksi dengan menerapkan metode yang dibuatnya yaitu Contrastive Learning. Dai sendiri berhasil memperjelas (distinctiveness) objek yang diamati dan berhasil membuat selisih evaluasi yang signifikan dibandingkan metode lainnya seperti Adaptive Attention (AA) dan Neuraltalk2. Penelitian image *caption*ing lainnya yang mencoba memberikan *caption*ing dari ungkapan kepribadian manusia telah dilakukan oleh [6]. Shuster menggunakan dataset baru personality-*caption*s, dengan 241,858 *caption*s, setiap *caption* memiliki 215 kemungkinan perbedaan mengenai kepribadian manusia. Shuster melakukan eksperimen *caption*ing dengan melibatkan berbagai *caption* dari segi

perseorangan. *Caption*ing yang diungkapkan adalah membuat kalimat berdasarkan perasaan dari seseorang ketika dalam keadaan seperti senang, marah, sedih, ataupun perasaan lainnya. Penggunaan model TransResNets untuk melakukan proses representasi kalimat dan representasi image dan mendapatkan evaluasi Bleu-1 = 79.8%, Bleu-4 = 36.3% [6].

Sebagai upaya perbaikan dari hasil text generator, peneliti banyak melakukan rekayasa terhadap CNN maupun LSTM [7]. Hal ini dilakukan karena belum mendapatkan hasil terbaik dari model yang dikembangkan. Pada sisi evaluasi, pengukuran yang dilakukan adalah dengan menggunakan BLEU-N yang telah diusulkan oleh Papineni [8]. Masalah lainnya juga adalah *caption*ing yang diproduksi ketika membangkitkan teks dari deteksi objek yang sudah dikenali, masih ada teks yang tidak berkesesuaian dengan objek yang dideteksinya. Disisi lain adalah akurasi image generator dan teks generator masih belum mencapai nilai yang optimal [9]–[12]. Bila disimpulkan beberapa tantangan dalam menyelesaikan image *caption*ing adalah ketika menentukan image generator yang tepat untuk domain *caption*ing yang diharapkan, kemudian teks generator yang tepat dalam domain yang ingin dicapai dalam menghasilkan *caption*ing yang diperoleh. Aspek lainnya yang masih dalam kajian adalah mendeteksi objek dengan menyebutkan *caption*ing yang tepat berdasarkan penampakan gambar. Kasus yang telah disebutkan ini belum termasuk deteksi hiden objek, objek bertumpuk, penambahan keterangan dari objek, dan penyusunan kalimat berdasarkan domain tertentu misalnya medical, biologi, dan seterusnya dari *caption*ing tersebut.

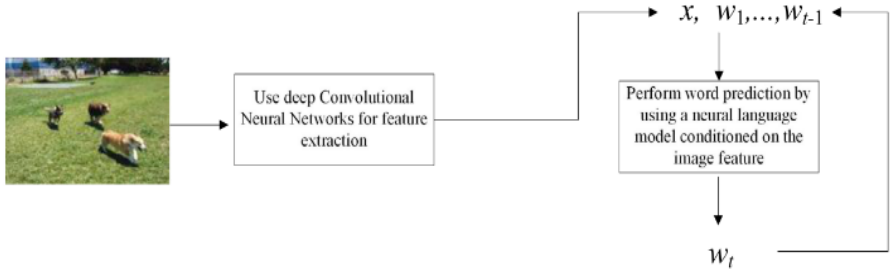
Berdasarkan penelahaan dari penelitian yang dilakukan mengenai image *caption*ing, seperti[3], [6], [13], telah membuktikan keberhasilan dalam menyelesaikan masalah image *caption*ing. Permasalahan yang masih dalam kajian yaitu bahwa deskripsi *caption*ing hanya bisa dilakukan pada satu input array dari pixel dengan resolusi yang sudah pasti [14]. Kemudian identifikasi objek untuk mengetahui *caption* yang tepat masih mendapatkan akurasi kecil [14]. Masalah dalam extraksi image untuk mendapatkan fitur yang direlasikan dengan word embedding yang perlu dikaji untuk menghindari overfitting pada konten *image* [15].

Terkait dengan apa yang sudah dikemukakan oleh para peneliti, masih adanya masalah dalam bidang geologi terutama *caption* citra geologi. Beberapa kajian masih menghasilkan *caption* untuk objek-objek yang sudah dianotasi. Tetapi *caption* dari geologi bebatuan belum ada yang menggunakan teknik dari *image captioning*. Perspektif *scientific revolution* yang dikemukakan oleh Kuhn, bila dikaitkan dengan masalah *caption* pada citra geologi, maka masalah ini berada pada tahap krisis. Alasan berada pada tahapan ini, karena perjalanan pennyelesaian *captioning* belum mencapai tahap yang sustain untuk masalah seperti kedokteran, citra geologi, citra seni-rupa, citra iklan, citra yang mengenali kebencian, dan bentuk citra lainnya.

2. Kajian Pustaka

Image caption generation merupakan masalah bidang komputer yang melibatkan *computer vision, natural language processing (NLP),* dan pembelajaran mesin. Pekerjaan yang dilakukan adalah melakukan translasi *image* menjadi suatu teks yang sesuai dengan objek termaksud [2]. *Image captioning* bertujuan untuk memproduksi deskripsi kalimat yang selaras dengan image yang diberikan. *Image captioning* diinspirasi dari human *visual system* yang sudah beberapa tahun lalu dikenal dengan sebutan *visual attention*. *Visual attention* ini dibaurkan dengan berbagai model *image captioning* seperti pada penelitian yang telah disebutkan pada bagian pendahuluan. Mekanisme *attention* adalah mekanisme

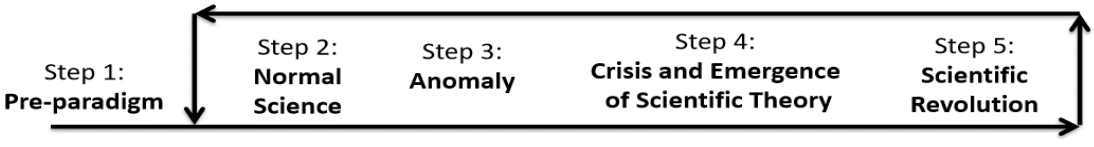
penuturan dalam membuat model untuk dapat memberi tanda spesifik pada suatu *region* ketika memproduksi *caption* pada pembacaan seluruh image [16]. Selain tugas yang sudah disebutkan, penelitian ini juga melibatkan penggunaan berbagai mesin learning seperti *deep neural network* (DNN) dan variasinya. Penggunaan teknik ini dapat memproduksi kata yang tepat pada *image* yang sedang diamati. Teknik yang telah banyak digunakan untuk DNN seperti RNNs (*recurrent neural nets*), *Long-Short-Term-Memory* (LSTM) *component* [17]. Dai menggunakan recurrent neural network (RNN) dan LSTM untuk memproduksi teks untuk bisa menunjukan perbandingan dengan human description, walaupun hasilnya masih belum memuaskan. Pada studi yang telah dilakukan bahwa produksi *caption* dari image yang diamati masih jauh dari apa yang diharapkan [5]. Terlepas dari kebutuhan untuk identifikasi terhadap objek yang berada dalam image, setiap image *caption*ing harus mampu menganalisis objek dalam keadaan apapun, memahami hubungan antar objek, serta mengekspresikan informasi semantik dalam bahasa natural [18]. Penelitian Gan (2017) melakukan *caption*ing dengan memproduksi teks yang dibaurkan dengan style dari gambar. Gan mengusulkan framework baru yang dinamakan StyleNet. Gan memberikan kontribusi untuk dataset set dengan *caption*ing style secara mandiri tanpa dipasangkan dengan image, dan juga pasangan *caption* dan image yang sesuai dengan keadaannya [19]. Secara menyeluruh, konsep kerja dari image *caption*ing memiliki beberapa tahapan seperti image generator, pembangkitan kata yang tepat dengan objek yang dibaca, dan pembangkitkan kalimat untuk image yang dimaksud. Setelah itu dilanjutkan dengan mengevaluasi mesin learning yang digunakan untuk melihat seberapa dekat kalimat yang dibangkitkan dengan image yang dimaksud [3].



Gambar 1. Struktur umum pemrosesan image *caption*ing [20]

Gambar 1 merupakan struktur umum untuk pembelajaran multimodel berbasiskan image *caption*ing. Kerja yang dilakukan adalah membuat image feature dengan cara mengektraksi gambar seperti menggunakan deep CNN (convolutional neural networks). Selanjutnya fitur image akan dilanjutkan ke proses model neural language yang memetakan fitur image kepada ruang fitur kata dan membentuk kondisi prediksi pada fitur image yang selaras dengan kata yang dimaksud [20].

Pada *scientific revolution Kuhn [21]*, mengemukakan beberapa langkah untuk mengidentifikasi suatu metode berada pada suatu tahapa. Kuhn menggunakan beberapa tahapan untuk mengenali keberadaan metode.



Gambar 2. Tahapan perkembangan Science menurut Thomas Kuhn [21]

Tahapan perkembangan science yang dikemukakan Thomas Kuhn dimulai pada tahap *pre-paradigm, normal science, anomaly, crisis*, dan *scientific revolution*. Pre-paradigm merupakan fenomena yang muncul mengenai science. Tahap *normal science* merupakan tahap dimanan segala bentuk metode yang ada untuk memecahkan permasalahan science yang muncul. Tahap anomaly merupakan tahap dimana metode yang ada mulai ditemukan adanya penyelesaian yang kurang memuaskan bahkan tidak sesuai dengan tujuan science tersebut. Tahap crisis merupakan tahap bahwa metode penyelesaian untuk science diperlukan metode baru. Metode ini untuk menjawab metode sebelumnya yang dinilai tidak lagi sesuai untuk masalah *science* yang berkembang. Tahap *scientific revolution* merupakan tahap dimana adanya metode baru yang merubah metode lama dan meninggalkan metode lama.

3. *Image Captioning* menurut *Science revolution*

Image captioning merupakan perkembangan metode dari *text captioning* yang diberikan pada suatu citra. merupakan teks singkat yang diberikan pada citra disuatu buku, majalah, atau *Captioning* surat kabar mengenai deskripsi atau penjelasan mengenai citra tersebut. Deskripsi ini diberikan menurut visualiasi yang dilihat seseorang pada citra tersebut. Tahapan science revolution menurut Kuhn pada image *caption*ing akan dijelaskan berikut [22].

3.1 Tahapan Pre-Paradigm

Asal mula dari perkembangan *image captioning* berawal dari *text caption*. *Text caption* dikembangkan untuk lebih memberikan pemahaman suatu informasi yang ingin disampaikan pada citra yang ditunjukkan. Pemahaman terhadap citra menjadi fokus utama pada setiap *captioning* yang diberikan. Srihari (1994) menyatakan bahwa lokalisasi suatu objek pada citra menjadi sumber informasi dalam pemahaman suatu citra. Pemberian teks dapat membantu lebih mengerti tentang ilustrasi objek-objek yang ada pada suatu citra [1]. Berawal dari pemahaman ini perkembangan *caption*, menjadikan aspek science untuk dikembangkan dalam bidang *computer vision*. Srihari menganggap penting adanya pembangkitkan suatu teks secara otomatis dari citra yang diamati. Pada metode yang dikembangkan Srihari, bahwa suatu *caption*ing bisa dikembangkan dari penggabungan bahasa dan vision. Implikasi dari metode ini menghasilkan suatu metode yang mengembangkan suatu input teks dengan citra untuk mendapatkan suatu *caption [22]*.

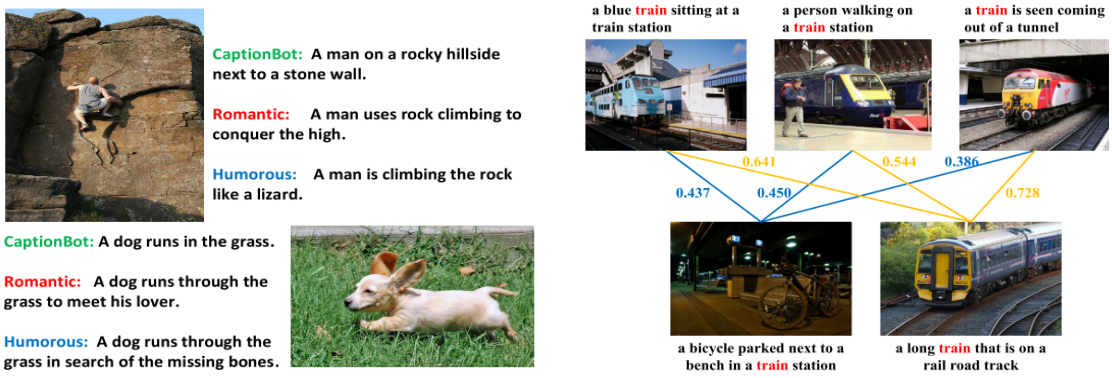
3.2. Tahap Normal Science

Perkembangan dari metode *captioning*, sempat terhambat, karena pembuatan teks otomatis dari pembacaan suatu citra masih belum dihasilkan sesuatu yang sesuai dengan teks hasil dari manusia. Terlebih lagi, mesin komputasi yang masih terbatas untuk menghasilkan *caption* yang sesuai dengan ground-truthnya masih terkendala. Diusulkan suatu teknik baru berbasiskan neural network yang dikenal dengan deep learning, memberikan tantangan baru untuk bisa menghasilkan suatu *captioning*. Terlebih dalam bidang pembangkitan teks, ditemukannya model LSTM pada tahun 1997 oleh Sepp Hochreiter dan Jürgen Schmidhuber. Kemudian diusulkannya model *captioning* pertama kali dari arsitektur yang dikenal ImageNet oleh Fei-Fei pada tahun 2009 [23].

Penelitian Gan dkk (2017) melakukan *captioning* dengan memproduksi teks yang dibaurkan dengan *style* dari gambar. Gan mengusulkan *framework* baru yang dinamakan *StyleNet*. Gan memberikan kontribusi, Gambar 3a, untuk dataset set dengan *captioning style* secara mandiri tanpa dipasangkan dengan *image*, dan juga pasangan *caption* dan

Jurnal Manajemen Informatika (JAMIKA)
Volume 10 Nomor 2 Edisi Oktober 2020
E ISSN: 2655-6960 | P ISSN: 2088-4125
OJS: https://ojs.unikom.ac.id/index.php/jamika

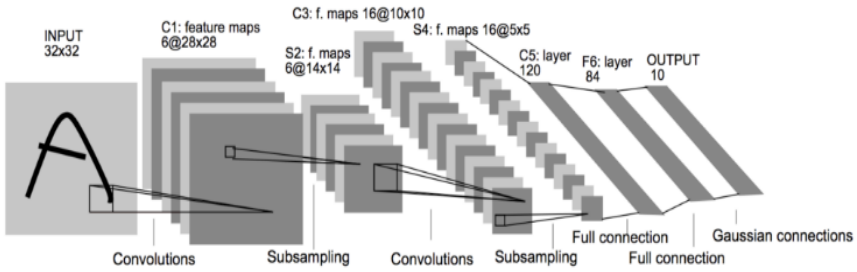
image yang sesuai dengan keadaannya [13]. Terlepas dari kebutuhan untuk identifikasi terhadap objek yang berada dalam image, setiap image *caption*ing harus mampu menganalisis objek dalam keadaan apapun, memahami hubungan antar objek, serta mengekspresikan informasi semantik, Gambar 3b, dalam bahasa natural [24].



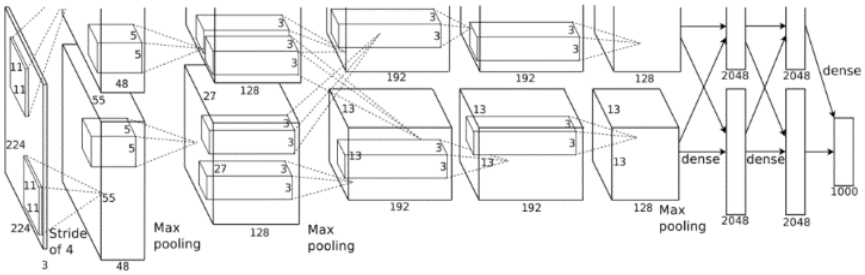
(a) (b)

Gambar 3. Berbagai *normal science image captioning [18]* [19]

Menurut tahapan *normal science Kuhn*, bahwa sudah ditemukan metode yang bisa memberikan teks otomatis pada gambar yang dibaca. Bila diamati pada Gambar 3, disajikan berbagai hasil *caption* dengan menggunakan LSTM dan CNN. Kedua metode ini sudah sanggup menyajikan hal tersebut. Metode yang sudah ditemukan dengan menggunakan CNN dan LSTM bisa membangkitkan teks yang diinginkan yang sesuai dengan bahasa manusia. Pada arsitektur CNN, telah dibuat beberapa model CNN yaitu: LeNet-5 [25], Gambar 4, AlexNet Gambar 5 [26], VGG [27], Inception and GoogLeNet [28], dan Residual Network atau ResNet [28].



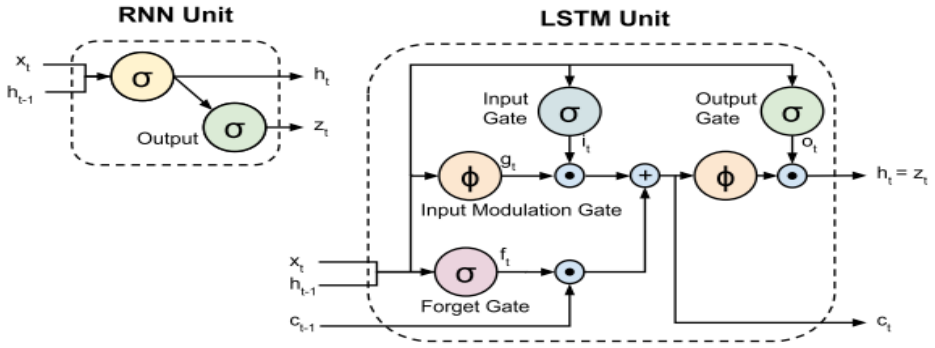
Gambar 4. Arsitektur CNN LeNet-5 [25]



Gambar 5. Arsitektur CNN AlexNet [26]

Jurnal Manajemen Informatika (JAMIKA)
Volume 10 Nomor 2 Edisi Oktober 2020
E ISSN: 2655-6960 | P ISSN: 2088-4125
OJS: https://ojs.unikom.ac.id/index.php/jamika

Long Short-Term Memory (LSTM) merupakan modul recurrent yang memungkinkan pembelajaran dalam jangka waktu yang panjang. Unit LSTM memiliki state tersembunyi tambahan sebagai mekanisme nonlinear yang mengijinkan suatu state melakukan backpropagasi tanpa adanya modifikasi, perubahan, atau reset. Pembelajaran pada LSTM menggunakan gerbang fungsi sederhana yang memiliki kemampuan untuk pembelajaran speech recognition dan language translation models.



Gambar 7. Arsitektur RNN dan LSTM [29]

Metode image *caption*ing yang dituangkan dalam bentuk algoritma sebagai berikut:

1. *Setting* CNN, dengan meniadakan *layer* paling akhir dan *softmax layer*, sehingga mendapatkan 4096 *dimensional vector* yang mendeskripsikan global konten dari image
2. Mentransfer bobot dari VGG16 yang telah dipersiapkan pada *IMAGENET* ke dalam model Xianwei.
3. Untuk *decoder* (*word generation*):
   a. Melakukan pemetaaan objek satu per satu kepada teks dengan process *recurrent*.
   b. Secara matematik mendefinisikan suatu distribusi probabilitas P(S|I), dimana S adalah kumpulan kata <w1, w2, …, wn> pada suatu *image* I.
   c. Model dilatih untuk memaksimalkan *posterior* pada suatu *training dataset*. Probabibilitas distribusi dapat dituliskan sebagai berikut:

$$P\left(w_1, w_2, \ldots, w_{|s|} \middle| I ¿= \prod_{t=1}^{|s|} P\left(w_t \vee I, w_{1:t-1}\right)\right. \quad (1)$$

   d. Asumsi yang dibuat adalah pembangkitkan untuk kata bergantung pada *image* I dan kata yang dibangkitkan sebelumnya $w_{1:t-1}$.
   e. Pengulangan pada distribusi probabilitas digunakan untuk RNN.
   f. Saat diberikan input dari hasil RNN pada waktu t, LSTM menerima *hidden state* h_{t-1} pada saat *cell* state c_{t-1}.

$$i_t = \sigma\left(W_{xi} x_t + W_{hi} h_{t-1} + b_i\right) \quad (2)$$

$$f_t = \sigma\left(W_{xf} x_t + W_{hf} h_{t-1} + b_f\right) \quad (3)$$

$$o_t = \sigma\left(W_{xo} x_t + W_{ho} h_{t-1} + b_o\right) \quad (4)$$

$$g_t = \tanh\left(W_{xc} x_t + W_{hc} h_{t-1} + b_c\right) \quad (5)$$

$$c_t = f_t \circ c_{t-1} + i_t \circ g_t \quad (6)$$

$$h_t = o_t \circ \tanh\left(c_t\right) \quad (7)$$

Untuk proses *training* dan inferensi, yaitu memaksimasi *log-likelihood* dengan fungsi *loss* =

Jurnal Manajemen Informatika (JAMIKA)
Volume 10 Nomor 2 Edisi Oktober 2020
E ISSN: 2655-6960 | P ISSN: 2088-4125
OJS: https://ojs.unikom.ac.id/index.php/jamika

$$L = \sum_{I,S \epsilon X} \log P(S \vee I;\theta) \quad = \quad \text{¿} \sum_{t=1}^{\text{¿}s \vee \text{¿} \log P(w_t \vee w_{1:t-1}, I;\theta)} \text{¿} \sum_{I,S \epsilon X} \text{¿} \qquad (8)$$

3.3. Tahap *Anomaly*

Pada tahap ini, setiap *science* yang telah ditemukan, maka akan dicari *falsification* dari metode tersebut. *Falsification* ini diajukan oleh Karl Popper yang menyatakan bahwa setiap *science* yang ditemukan harus bisa difalsifikasi agar bisa tetap dikatakan *science* dan bukan *pseudo-science*. Beberapa masalah yang belum bisa diselesaikan dengan model CNN dan LSTM, kemudian dijadikan peluang sebagai tantangan pada riset berikutnya. Beberapa penelitian yang membangun metode baru untuk memperbaiki metode CNN dan LSTM yang belum bisa menangani masalahnya.

Beberap contoh falsifikasi dari image *caption*ing seperti masalah deskripsi faktual dari gambar, termasuk objek, gerakan, dan hubungannya, memberikan *captioning* berupa *caption* sentimen dari gambar yang dideteksi berupa sifat [3]. Andrew menyatakan bahwa model CNN dan LSTM, tidak bisa menangani masalah yang ditanganinya [3]. *Caption*ing dengan teknik frekuensi kemunculan kata dan menyediakan batasan performansi rendah. Masalah lainnya adalah dengan memanfaatkan image retrieval dan asumsi bahwa image serupa memiliki diagnosis serupa juga [30]. Memberikan teks dengan gaya bahasa yang diinginkan seperti kalimat *romantic*, lucu, pernyataan senang [2].

(a) Caption: Sad, Gorgeous, scary, lovely, cute, creepy

(b) Caption: a man riding skis down a snow covered slope a slope

Gambar 8. Berbagai contoh falsifikasi *captioning* untuk model *CNN* dan *LSTM*

Permasalahan *caption* pada citra, Gambar 9, lainnya adalah pada citra geologi bebatuan, dimana pemberian *caption* berdasarkan singkapan dari bebatuan yang terdiri dari beberapa jenis bebatuan. *Caption*ing yang ada pada saat ini adalah merupakan pemberian ahli geologi terhadap bebatuan yang ditemukannya. Proses identifikasi dilakukan dengan cara melakukan penyingkapan dan uji kimia bebatuan untuk menentukan jenis bebatuan yang ditemukan. Sehingga pada akhirnya, setelah beberapa kali pengamatan dan pengalaman seorang geologi, bisa langsung menentukan nama bebatuannya dan posisi bebatuan yang dimaksud dengan hanya visual. Contoh Gambar 8 dan Gambar 9, merupakan anomaly yang terjadi pada metode CNN dan LSTM yang ada. Berdasarkan hal ini maka diperlukan perbaikan metode untuk mendapatkan *caption* yang diberikan oleh manusia.

3.4. Tahap Crisis

Pada setiap tahap yang telah dipaparkan, pada tahap ini Kuhn menyampaikan perlu adanya penggalian metode baru untuk menanggulangi metode lama yang tidak bisa memberikan *caption* sesuai dengan pemberian manusia. Perbaikan dan pengembangan metode diperlukan untuk menganggulangi masalah ini seperti pada *CNN*. Apakah parameter seperti *convolusi, filter, stride*, dan *pooling* yang harus di-tune. Perlakukan ini dilakukan secara terus menerus untuk mendapatkan *caption* yang benar-benar mendekati pemberian manusia. Masalah lainnya yang perlu ditemukan metode baru pada penggunaan CNN dan LSTM seperti : *agreement with semantics, robustness to noise (invariant to perturbations), computational efficiency (ability to work in real time and large scale), invariance to background (allowing region-based querying), local linearity (*contoh *following triangle inequality in a neighborhood).* Sedangkan kebutuhan akan metode seperti untuk*: treating features as vectors, non-vector representations, or ensembles, using region-based similarity, global similarity, or a combination of both, computing similarities over linear space or nonlinear manifold, considering the role played by image segments in similarity computation, using stochastic, fuzzy, or deterministic similarity measures, use of supervised, semi-supervised, or unsupervised learning*. Masalah dan metode yang disebutkan sebelumnya menjadikan metode *caption*ing menjadi pada tahap crisis, dan perlu ditemukan metode-metode baru guna menangani masalah ini.

(a) (b) (c) (d)

Gambar 9. (a) ***caption*** : runtunan batupasir bersisipan batulumpur, (b) ***caption*** : singkapan batugamping kristalin , (c) ***caption*** : batuan batupasir konglomeratik (d) ***caption*** : singkapan batugamping terumbu.

4. Pembahasan

Pada bab 3 telah dipaparkan serangkaian kajian science image *caption*ing menurut Kuhn. Perkembangan dari image *caption*ing dimulai dari pemberian teks pada foto, koran, atau gambar apapun sehingga bisa dimengerti arti dari foto tersebut. Secara kronologis bahwa perkembangan image *caption*ing berawal pada tahun 1994 yang hanya memberikan teks pada foto yang ada. Kemudian berkembang menjadi suatu teknik yang dikenal dengan computer vision. Dimana metode computer vision ini mencoba memberikan hasil berdasarkan citra yang diamati. Seperti image processing, image recognition, image detection, image classification, image *caption*ing. Pada metode image *caption*ing inilah adanya penggabungan metode dengan natural language processing yang mengupayakan adanya *caption* terhadap citra yang dibaca. Seperti pada penelitian [14], [15], [17], [31],

[32], perubahan suatu metode menyebabkan adanya perbaikan pada metode CNN dan LSTM. Walaupun sampai saat ini, CNN dan LSTM, masih bisa membangkitkan teks dan representasi gambar serta membangkitkan *captioning*. Hasil resume falsifikasi terhadap metode CNN dan LSTM, bisa dilihat pada Tabel 1.

Tabel 1. Kesimpulan metode CNN dan LSTM yang bisa difalsifikasi

Aspek	Metode	Falsifikasi
Generator representasi fitur citra	CNN	Masih memerlukan ketelitian identifikasi objek
		Performansi jumlah konvolusi masih bisa dituning
		Dimensi konvolusi, filter, stride, dan pooling masih bisa diperbaiki
		Deteksi edge antar objek
		Kesamaan antar objek
		Reduksi representasi dimensi untuk objek yang sama
Generator Teks	LSTM	Membangkitkan teks berdasarkan relasinya
		Membangkitkan teks yang seseuai dengan ground-truthnya
		Tuning parameter

Berdasarkan hasil pada Tabel 1, dapat dikatakan bahwa penelitian mengenai image *caption*ing masih terbuka lebar. Terkait objek penelitian yang masih dapat digali diberbagai bidang seperti geologi, kedokteran, seni rupa, video *captioning*, dan berbagai objek penelitian lainnya.

5. Kesimpulan

Image captioning merupakan gabungan beberapa bidang seperti *computer vision*, pemrograman bahasa natural, dan mesin learning. Sebagai tahapan tugas dari *image captioning* terdiri dari beberapa tahapan seperti *image* representasi yaitu tugas untuk mendapatkan ektraksi fitur dari objek yang dideteksi dengan problem domain yang diajukan. Pada tahapan *image* representasi selain ekstraksi fitur yang membentuk vektor setiap objek yang dideteksi, tugas lainnya adalah membuat label pada setiap objek yang dideteksi. Selanjutnya yaitu *sentence representation*, yaitu tahapan untuk mendapatkan kalimat yang sesuai dengan objek yang diamati. Pada bagian ini hasil yang dikerjakan pada bagian *image* representation harus mampu memberikan *caption* yang mendekati dengan *ground-truth* yang diberikan. Selain kedua tahapan tersebut, ada bagian yang dikenal dengan *aligment objective*, yang bertugas menjaga kemiripan teks yang diproduksi dengan *ground-truth* yang sudah ada. Setiap mesin pembelajaran yang menjadi kontribusi penelitian ini, akan dievaluasi hasilnya dengan menggunakan BLEU, METEOR, dan CIDEr. Tahapan *scientific revolution* menurut Kuhn untuk *image captioning* berada pada tahapan *crisis*. Dimana masih dibuka peluang perbaikan metode terhadap objek citra dari berbagai bidang. Sedangkan dari filsafat ilmu bahwa *image captioning* masih bisa difalsifikasi, sehingga masih bisa disebut *science*.

Ucapan Terima Kasih

Terima kasih kepada **Dr. Dimitri Mahayana** yang telah memberikan pengetahuan berharga tentang filsafat ilmu dan *scientific revolution*. Metode-metode yang diajarkan sangat membantu dalam penulisan dan penelitian dari *image captioning*.

Daftar Pustaka

[1] R. K. Srihari, "Use of Captions and Other Collateral Text in Understanding Photos," 1994, pp. 1–32.
[2] J. Gu, G. Wang, J. Cai, and T. Chen, "An Empirical Study of Language CNN for Image Captioning," in *IEEE International Conference on Computer Vision An*, 2017, pp. 1231–1240.
[3] Y. U. and T. H. Andrew Shin, "Image Captioning with Sentiment Terms via Weakly-Supervised Sentiment Dataset," in *British Machine Vision Conference*, 2016, p. 53.1-53.1.
[4] J. Aneja and A. G. Schwing, "Convolutional Image Captioning."
[5] B. Dai, "Contrastive Learning for Image Captioning," in *Advances in Neural Information Processing Systems Conferece*, 2017, no. 30, pp. 898–907.
[6] K. Shuster, S. Humeau, H. Hu, A. Bordes, and J. Weston, "Engaging Image Captioning via Personality," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12516–12526.
[7] B. Dai and S. Fidler, "A Neural Compositional Paradigm for Image Captioning," no. NeurIPS, pp. 1–11, 2018.
[8] K. Papineni, S. Roukos, T. Ward, and Z. Wei-Jing, "BLEU: a Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002, pp. 311–318.
[9] J. Devlin *et al.*, "Language Models for Image Captioning : The Quirks and What Works," *Comput. Lang.*, pp. 100–105, 2015.
[10] V. Kougia, J. Pavlopoulos, and I. Androutsopoulos, "A Survey on Biomedical Image Captioning," 2016.
[11] X. Li, X. Song, L. Herranz, Y. Zhu, and S. Jiang, "Image Captioning with both Object and Scene Information," in *24th ACM international conference on Multimedia*, 2016, pp. 1107–1110.
[12] D. Shin and I. Kim, "Deep Image Understanding Using Multilayered Contexts," *Math. Probl. Eng.*, vol. 2018, pp. 1–11, 2018.
[13] Z. Gan *et al.*, "Semantic compositional networks for visual captioning," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017–Janua, pp. 1141–1150, 2017.
[14] A. Karpathy and L. Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 664–676, 2017.
[15] X. He, B. Shi, X. Bai, G. Xia, and Z. Zhang, "Image Caption Generation with Part of Speech Guidance," *Pattern Recognit. Lett.*, vol. 0, pp. 1–9, 2017.
[16] J. Mun, L. Yang, Z. Ren, N. Xu, B. Han, and A. Go, "Streamlined Dense Video Captioning."
[17] J. Aneja and A. G. Schwing, "Convolutional Image Captioning," *Comput. Vis. Pattern Recognit.*, pp. 5561–5570, 2017.
[18] G. Ding, M. Chen, S. Zhao, H. Chen, and J. Han, "Neural Image Caption Generation with Weighted Training and Reference," *Cognit. Comput.*, p. 10.1007, 2018.
[19] C. Gan, Z. Gan, X. He, and J. Gao, "StyleNet : Generating Attractive Visual Captions with Styles," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 955–964.
[20] S. Bai and S. An, "A survey on automatic image caption generation Shuang,"

Neurocomputing, vol. 311, pp. 291–304, 2018.
[21] T. Kuhn, *The Structure of Scientific Revolutions, 4 (2012)*. University of Chicago Press, 1962.
[22] D. Mahayana, *Filsafat Ilmu Pengetahuan*. Bandung, Indonesia: ITB Press, 2018.
[23] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-fei, "ImageNet : A Large-Scale Hierarchical Image Database," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.
[24] "F. Serafino, G. Pio and M. Ceci, 'Ensemble Learning for Multi-Type Classification in Heterogeneous Networks,' in," vol. 30, no. 12, p. 8525379, 2018.
[25] H. Kaiming, Z. Xiangyu, R. Shaoqing, and S. Jian, "Deep Residual Learning for Image Recognition," *Comput. Vis.*, pp. 1–9, 2016.
[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Adv. Neural Inf. Process. Syst. 25 (NIPS 2012)*, vol. 25, pp. 1–9, 2012.
[27] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-Based Learning Applied to Document Recognition," in *IEEE*, 1998.
[28] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arxiv*, pp. 1–14, 2015.
[29] J. Donahue *et al.*, "Long-term Recurrent Convolutional Networks for Visual Recognition and Description," pp. 1–14, 2015.
[30] V. Kougia, J. Pavlopoulos, and I. Androutsopoulos, "A Survey on Biomedical Image Captioning," in *Proceedings of the Second Workshop on Shortcomings in Vision and Langua*, 2016, pp. 26–36.
[31] H. Agrawal *et al.*, "nocaps : novel object captioning at scale," in *IEEE International Conference on Computer Vision (ICCV) 2019*, 2019.
[32] V. Batra, Y. He, and G. Vogiatzis, "Neural Caption Generation for News Images," pp. 1726–1733, 2016.