Segmenter: Transformer for Semantic Segmentation

Robin Strudel
Inria

Ricardo Garcia
Inria

Ivan Laptev
Inria

Cordelia Schmid
Inria

Abstrak

Segmentasi semantik merupakan salah satu bidang pada visi komputer yang dapat membantu manusia dalam menyelesaikan berbagai permasalahan. Segmentasi semantik mampu mengelompokkan bagian dari suatu citra ke dalam kelas tertentu. Sudah banyak model yang dibuat dalam ranah ini, namun muncul metode baru yang mengandalkan metode self-attention dalam ekstraksi fitur. Metode baru bernama Segmenter ini dibangun murni menggunakan transformer, dan memberikan hasil akurasi dan waktu training yang menyaingi metode-metode tercanggih pada dataset umum yaitu ADE20K, Pascal Context, dan Cityscapes. Maka dari itu, akan dilakukan percobaan model Segmenter ini menggunakan dataset yang berbeda dari dataset yang digunakan pada umumnya untuk melihat performanya. Hasil yang didapat menunjukkan akurasi yang cukup baik dan waktu running yang cukup sebentar.

1. Pendahuluan

Segmentasi semantik merupakan permasalahan visi komputer yang cukup menantang dengan aplikasi yang sangat luas meliputi *autonomous driving*, robotika, *augmented reality*, rekayasa citra, pencitraan medis, dan lain-lain. Tujuan utama dari segmentasi citra adalah memasangkan setiap piksel pada citra ke label kategori dari objek dasarnya untuk memberikan representasi citra level-tinggi untuk target pekerjaan yang dibutuhkan, seperti mendeteksi batas antara orang dan pakaian mereka pada aplikasi coba pakaian. Meskipun sudah banyak usaha dan kemajuan dalam beberapa tahun terakhir, segmentasi citra masih menjadi masalah yang menantang karena variasi dalam-kelas yang banyak, variasi konteks, dan ambiguitas yang berasal dari *occlusion* dan resolusi citra yang rendah.

Pendekatan segmentasi semantik akhir-akhir ini cenderung mengandalkan arsiteksur konvolusi *encoder-decoder* tempat dimana

encoder membuat citra fitur dengan resolusi rendah dan *decoder* melakukan *upsampling* pada fitur ke peta segmentasi dengan *pixel-level scores*. Metode tercanggih menggunakan *Fully Convolutional Network* (FCN) dan menggapai hasil yang impresif pada *benchmark* segmentasi yang menantang. Metode-metode ini mengandalkan tumpukan konvolusi yang dapat belajar untuk menangkap informasi semantic yang kaya dan sudah sukses dalam dunia visi komputer. Namun, sifat lokal dari filter konvolusi membatasi akses pada informasi global dari sebuah citra. Padahal, informasi tersebut sangatlah penting saat melakukan *labeling* pada bagian lokal cenderung bergantung pada konteks citra secara keseluruhan. Untuk menangani hal tersebut, beberapa metode DeepLab menggunakan fitur agregasi menggunakan konvolusi yang didilasi dan *spatial pyramid pooling*. Penerapan ini dapat memperbesar area reseptif dari jaringan konvolusi dan mendapatkan fitur dengan berbagai skala. Dengan mengikuti beberapa perkembangan pada NLP, beberapa metode segmentasi mengeksplorasi skema-skema agregasi yang berbasis pada kanal ataupun atensi spasial dan atensi per-titik untuk dapat menangkap konteks informasi dengan lebih baik. Metode-metode tersebut namun masih mengandalkan *backbone* konvolusi, yang masih bias terhadap interaksi secara lokal. Penggunaan lapisan-lapisan ini secara ekstensif untuk menangani bias ini menunjukkan batasan dari arsitektur konvolusi dalam segmentasi.

Untuk menangani batasan-batasan ini, Strudel dkk. memformulasikan masalah segmentasi semantik ke dalam masalah *sequence-to-sequence* dan menggunakan arsitektur transformer untuk menangkap informasi kontekstual pada tahapan awal dalam model. Secara *by-design*, transformer dapat menangkap interaksi secara global antara elemen-elemen dalam suatu layer dan tidak memiliki induksi sebelumnya. Namun, modeing interaksi global memerlukan biaya komputasi yang kuadratik yang membuat metode yang mahal apabila dipakai pada citra mentah. Dengan mengikuti beberapa perkembanga pada Vision Transformer, gambar dipecah menjadi potongan-potongan dan *patch embedding* dijadikan sebagai token input untuk *encoder transformer*. *Sequence* token yang sudah terkontekstual lalu di-*upsampling* menggunakan *transformer decoder* ke kelas skor level per piksel. Untuk proses *decoding*, digunakan pemetaan linear per titik dari *patch embedding* ataupun skema yang berbasis transformer yang dapat mempelajari *class embedding* dan diproses bersama token *patch* untuk membuat mask kelas.

Adapun pada penelitian ini diimplementasikan Segmenter pada dataset yang berbeda. Seperti hasil yang didapat pada paper awal, pendekatan

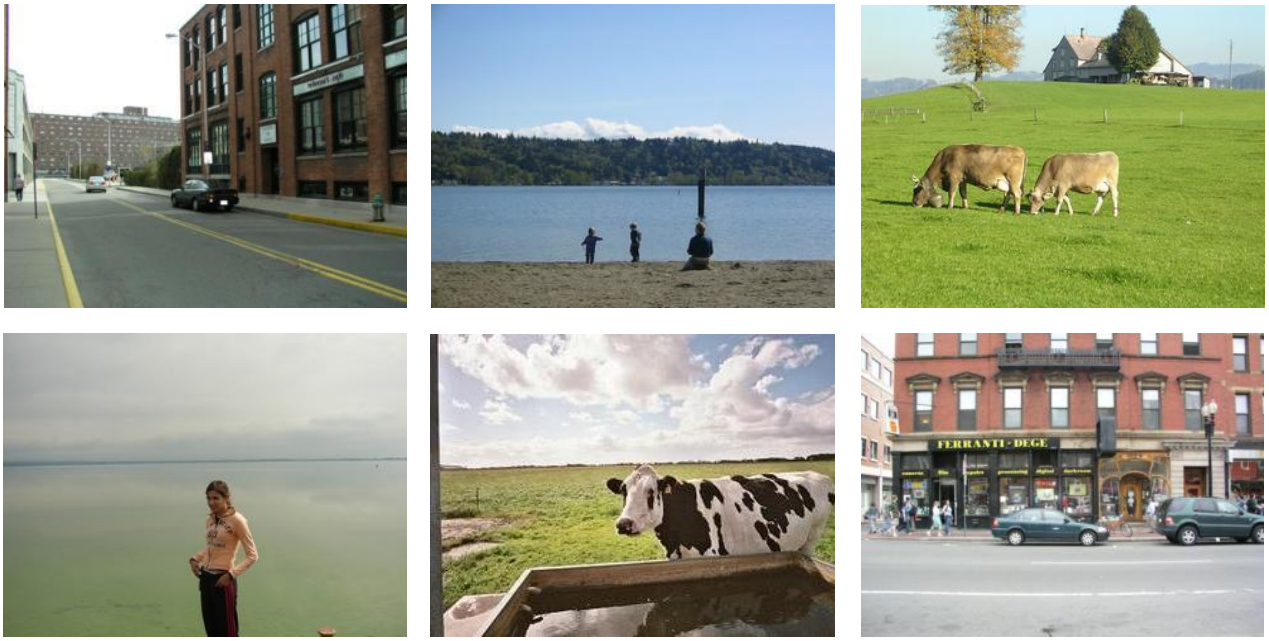
Segmenter ini memberikan hasil yang baik namun tetap simple, fleksibel, dan cepat. Salah satu varian model yang dicoba, varian *big*, mencapai nilai *mean Intersection over Union* (mIoU) sebesar 76,99% pada dataset yang digunakan. Varian ini memiliki keunggulan akurasi sekitar 6% dibandingkan varian yang lebih kecil.

2. Studi Terkait

Metode segmentasi semantik berbasis pada *Fully Convolutional Network* (FCN) yang digabungkan dengan arsitektur *encoder-decoder* telah menjadi pendekatan yang dominan. Pendekatan awal mengandalkan tumpukan dari konvolusi-konvolusi yang diikuti dengan *pooling* spasial untuk melakukan prediksi secara *dense*. Pendekatan selanjutnya melakukan *upsampling* peta fitur level tinggi dan menggabungkannya dengan peta fitur dengan level rendah dalam fase *decoding* untuk menangkap informasi global dan memulihkan batasan objek yang tajam. Untuk memperbesar *receptive field* dari beberapa layer konvolusi di awal, diusulkan pendekatan konvolusi yang didilasi atau *atrous convolution*. Untuk menangkap informasi global pada layer yang lebih tinggi, digunakan *spatial pyramid pooling* untuk menangkap informasi kontekstual dengan berbagai skala. Dengan menggabungkan beberapa pendekatan ini, DeepLabv3+ memberikan usulan sebuah arsitektur *encoder-decoder* berbasis FCN yang sederhana dan efektif. Beberapa penelitian mengganti layer *pooling* yang kasar dengan mekanisme atensi di atas *encoder* peta fitur untuk dapat menangkap ketergantungan jarak jauh dengan lebih baik.

Walau metode-metode segmentasi akhir-akhir ini lebih berfokus pada memperbaiki FCN, batasan pada operasi lokal yang dimiliki oleh konvolusi mungkin mengartikan bahwa konvolusi kurang efektif dalam mengekstraksi konteks citra secara global dan memberikan hasil segmentasi yang krang optimal. Maka dari itu, diusulkan arsitektur yang sepenuhnya berbasis transformer yang dapat menangkap konteks keseluruhan citra pada setiap layer dalam model baik dalam tahap *encoding* maupun *decoding*.

Transformer merupakan metode tercanggih dalam dunia *Natural Language Processing* (NLP). Model ini mengandalkan mekanisme *self-attention* dan menangkap keterkaitan jarak-jauh antara token/kata dalam sebuah kalimat. Transformer juga cocok paralelisasi yang dapat memfasilitasi ukuran dataset yang besar. Kesuksesan transformer pada NLP menginspirasi beberapa metode dalam visi komputer, dengan menggabungkan CNN dengan berbagai bentuk mekanisme *self-attention* dalam deteksi objek, segmentasi semantik, segmentasi panoptis, *video processing*, dan klasifikasi *few-shot*.



Gambar 1: Berbagai contoh citra pada dataset stanford background

Akhir-akhir ini, Vision Transformer (ViT) memperkenalkan arsitektur transformer tanpa konvolusi dalam melakukan klasifikasi citra tempat dimana citra input diproses dalam bentuk *sequence* dari token *patch*. Selagi ViT membutuhkan ukuran dataset yang besar untuk training, DeiT mengusulkan strategi distilasi berbasis token dan memberikan *vision transformer* yang kompetitif pada dataset ImageNet-1k menggunakan CNN sebagai rujukan. Penelitian selanjutnya mengembangkan persoalan ini ke klasifikasi video dan segmentasi semantik. SETR menggunakan *backbone* ViT dan *decoder* CNN standar. Swin Transformer menggunakan varian dari ViT dengan *local window* yang bergeser tiap layer dan Upper-Net sebagai *decoder* berbentuk pyramid.

3. Data

Data yang digunakan berasal dari *stanford background dataset*. Data citra mentah dalam bentuk .jpg, dan untuk label berbentuk .txt. Jumlah citra total dalam dataset ini adalah 715 yang berasal dari pilihan beberapa dataset publik seperti *LabelMe*, *MSRC*, *PASCAL*, *VOC*, dan *Geometric Context*. Contoh citra pada dataset dapat dilihat pada Gambar 1.

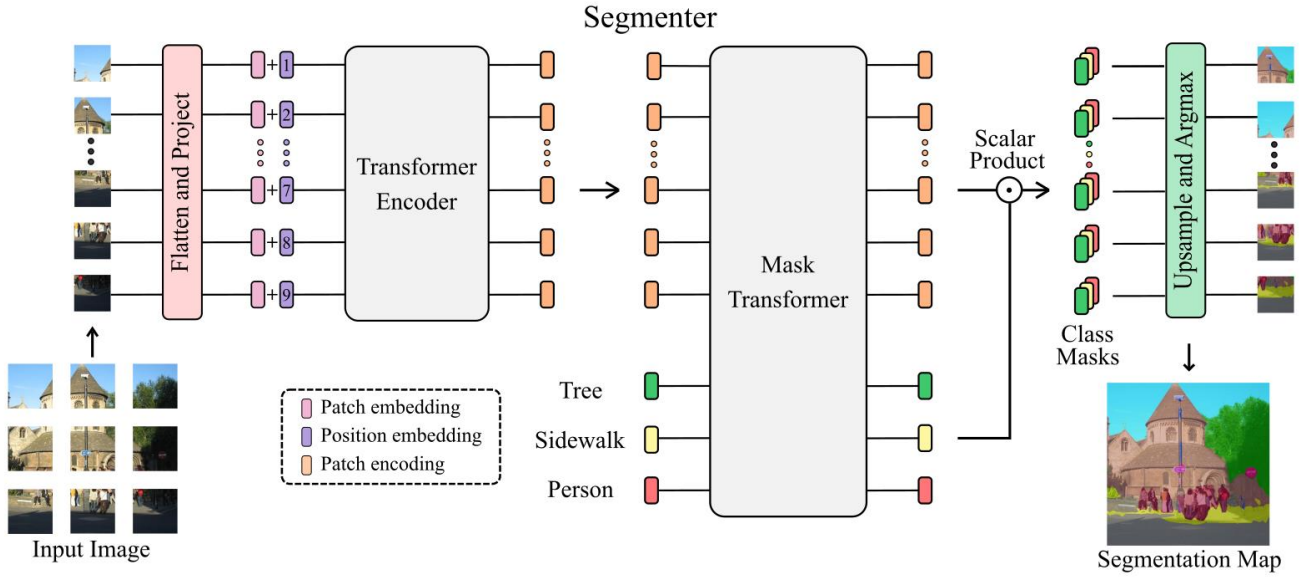
File label yang digunakan terdiri dari empat jenis file. File “horizons.txt” merepresentasikan dimensi dari citra dan letak horizon. File “*.regions.txt” merepresentasikan kelas

semantik dari setiap piksel, dengan kelas berupa; *sky*, *tree*, *road*, *grass*, *water*, *building*, *mountain*, dan *foreground*. Angka negatif merepresentasikan kelas *unknown*. File "*.surfaces.txt" merepresentasikan kelas geometri dari setiap piksel, dengan kelas berupa; *sky*, *horizontal*, dan *vertical*. File "*.layers.txt" merepresentasikan region yang berbeda pada citra.

4. Metode

Metode yang digunakan dalam melakukan segmentasi semantik mengacu pada Segmenter: Transformer untuk Segmentasi Semantik. Metode ini menggunakan pendekatan transformer dengan mekanisme *self-attention* dalam mencari atensi dari suatu citra. Kelebihan dari metode ini adalah kemampuan mekanisme tersebut dalam mencari konteks global dalam setiap iterasinya dan sistemnya yang bekerja secara paralel. Alur proses dari model Segmenter dapat dilihat pada Gambar 2.

Metode ini bekerja dengan membagi citra masukan menjadi potongan-potongan kecil dengan istilah *patch*. Potongan ini lalu diproyeksikan secara linear menjadi sebuah *patch embedding*, yang lalu disematkan dengan *positional emdedding* untuk menyimpan informasi posisinya. *Sequence* ini lalu dilewatkan ke *transformer encoder* yang tersusun atas beberapa blok *multi-headed self-attention* (MSA), *multilayer perceptron (MLP)*, dan *layer norm* (LN). Mekanisme *self-attention* dibentuk dengan tiga *point-wise linear layer* yang memetakan token ke representasi *intermediate*, *query* ***Q***, *key* ***K***, dan *value* ***V***. *Self-*



Gambar 2: Alur proses dari model Segmenter.

attention lalu dihitung berdasarkan rumus berikut.

$$MSA(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d}})V$$

Hasil keluarannya adalah sebuah *sequence encoding* yang kaya akan informasi semantik. Hasil ini lalu dilanjutkan ke *decoder* untuk membuat peta segmentasi dengan memetakan *patch-level encoding* ke *patch-level* skor kelas. Skor kelas ini lalu di-*upsampling* menggunakan interpolasi bilinear ke *pixel-level* skor untuk menghasilkan citra yang sudah tersegmentasi.

5. Eksperimen

Eksperimen yang akan dilakukan adalah mengimplementasikan Segmenter pada dataset standford background dataset. Percobaan akan dilakukan pada media Google Colab, dengan *environment* yang dipakai mengacu pada *mmsegmentation*. Model yang dicoba juga akan dibagi menjadi 3 model; *tiny*, *small*, dan *big*, untuk melihat perbedaan pengaruh dari *hyperparameter* berupa ukuran *sequence* dan jumlah head pada MSA. Setiap model dilatih dengan *max_iteration* sejumlah 1000.

Model *tiny* dalam varian ini memiliki nilai *hyperparameter* berupa ukuran token *sequence* sebesar 192, dan jumlah head pada *multi-headed self-attention* yaitu 3. Model *small* memiliki ukuran token yang lebih besar, yaitu 384, dan jumlah head pada setiap blok adalah 6. Model varian terbesar yaitu *big* memiliki panjang token sebesar 768, dan jumlah head sebesar 12.

Metrik yang akan diukur dari setiap model adalah mIoU dan waktu latih. IoU atau *Intersection over Union* dapat merepresentasikan nilai akurasi segmentasi model. IoU merupakan perbandingan antara irisan dan gabungan dari 2 area yaitu area *ground truth* dan *mask prediction*. Model yang dipakai juga sudah dilakukan *pretraining*. Dataset yang dipakai juga dilakukan data augmentasi/*pipeline* terlebih dahulu untuk menambah variasi data pada pelatihan model. Di antara augmentasi yang dilakukan ada; pengubahan ukuran antara 50% sampai 200%, pemotongan citra, pembalikan citra dan normalisasi. Berikut hasil eksperimen yang dilakukan.

Varian	Token size	Head	mIoU	Train time
Tiny	192	3	70,65%	259 s
Small	384	6	74,84%	345 s
Big	768	12	76,99%	743 s

Tabel 1: Hasil percobaan

Dapat dilihat pada tabel 1, varian *tiny* memiliki nilai *mean Intersection over Union* (mIoU) sebesar 70,65% dengan waktu training total selama 259 detik. Varian *small* menyaingi akurasi varian *tiny* sebesar 4,19%, yaitu dengan mIoU sebesar 74,84% dan waktu training yang lebih lama yaitu 345 detik. Varian terakhir, *big*,

menyaingi varian sebeluknya sebesar 2,15%, yaitu dengan mIoU sebesar 76,99%, dan waktu running terlama yaitu 743 detik.

Dapat dilihat bahwa semakin besar ukuran token dan jumlah head pada MSA, semakin tinggi nilai mIoU. Hal ini disebabkan representasi yang dihasilkan oleh encoder lebih mendetail dan head yang melakukan pencarian atensi lebih banyak. Namun, tradeoff dari ukuran hyperparameter yang besar adalah waktu komputasi yang cenderung lebih lama.

6. Kesimpulan

Pendekatan baru untuk segmentasi semantik berupa Segmenter mampu menjalankan tugasnya dengan akurasi yang baik dan waktu training yang cukup singkat. Dengan menambahkan ukuran token dan jumlah head pada blok MSA, dapat ditingkatkan akurasi dengan penambahan waktu training. Untuk kasus-kasus tempat dimana tidak diperlukan akurasi yang cukup tinggi, Segmenter dapat diimplementasikan karena waktu trainingnya yang cukup singkat, namun dengan akurasi yang cukup baik. Kemampuan model transformer untuk mengambil konteks global memberikan akurasi yang baik, dan modelnya yang bekerja secara paralel memberikan waktu running yang cukup sebentar.

7. Referensi

[1] Strudel, Robin, et al. "Segmenter: Transformer for semantic segmentation." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[2] Gould, Stephen, Richard Fulton, and Daphne Koller. "Decomposing a scene into geometric and semantically consistent regions." 2009 IEEE 12th international conference on computer vision. IEEE, 2009.