ブルーベリー栽培支援de

江﨑郁磨

Esaki Ikuma

(20xx 年度入学, xxxxxxxx)

指導教員: 清水郁子 准教授

東京農工大学 工学部 知能情報工学科

20xx 年度卒業論文

(20xx 年 x 月 xx 日 提出)

東京農工大学 工学部 知能情報工学科 20xx 年度 卒業論文 要旨

題目　タイトル

英文タイトル

学籍番号 xxxxxxxx　　氏名 名前 (英名)

提出日 20xx 年 x 月 xx 日

概要

目次

図目次

表目次

第 1 章 はじめに

1.1 研究の背景

1.2 本論文の構成

第 2 章　先行研究

研究の背景でもある Vision Transformer (ViT) [Dosovitskiy et al. 2021] は, 画像のクラス分類問題解決のために実装されたモデルである. モデルの設計は図 2.1 に示すとおりであり, Transformer Encoder [Vaswani et al. 2017] に BERT [Devlin et al. 2019] を適用した “Transformer” に加えて, 画像のトークン化を行う “Tokenizer” を構成要素に持つ.

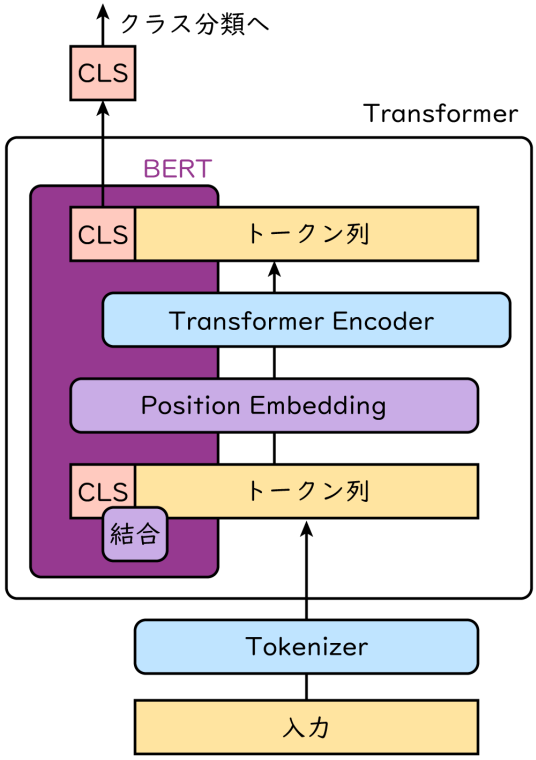

図 2.1　Vision Transformer の構成

中略

2.0.1 BERT

BERT [Devlin et al. 2019] とは Bidirectional Encoder Representations from Transformer の短縮表現であり, 自然言語処理領域にて発表されたモデルである. ViT と同様, Transformer Encoder を基にしたモデルであり, ViT の設計においても, BERT の要素技術が反映されている.

まず, ViT モデルのネットワーク規模は BERT の表現を踏まえたものとなっている. 具体的には, トークンの次元と Encoder Block の数, 次節で述べる MSA のヘッド数が表 2.1 のとおり定められている. ViT では Base(B) のネットワーク規模であることを, ViT-B と表記することで明示している.

中略

表 2.1　ViT（BERT）のネットワーク規模

ネットワーク規模	ViT の表記	トークンの次元 (d_t)	Encoder Block の数	MSA のヘッド数
Base(B)	ViT-B	768	12	12
Large(L)	ViT-L	1024	24	16

2.0.2 計算量

トークン数 l とその次元 d_t による影響を明確にするため, $T(d_{vec})$ は変数を用いた表現へと改める. 内積における計算処理が, 要素積に対して総和を求めていることを考慮すると時間計算量 $T(d_{vec})$ は d_{vec} の比例関数とみなせるため, $\alpha d_{vec} + \beta$ によって近似する. 加えて, ViT-B16 においては, $d_t = hd_m = 768$ かつ $l = 256$ であるため, ld_t および l^2h は ld_t^2 や l^2d_t に対して, 十分小さいものとする. 以上から, 全体の時間計算量 T_{msa} は式 (2.1) のとおり推定する. 時間計算量の推定値において, トークン数 l による時間計算量のオーダーは $O(n^2)$ であり, トークン数の増加は実行時間に大きく影響することが分かる.

$$T_{msa} = 3ld_t(\alpha d_t + \beta) + ld_t(\alpha d_t + \beta) + l^2h(\alpha d_m + \beta) + ld_t(\alpha l + \beta) = 2ld_t(2d_t + l)\alpha \tag{2.1}$$

謝辞

第 A 章 補足

参考文献

[Krizhevsky et al. 2012] Krizhevsky Alex, Sutskever Ilya, and Hinton Geoffrey E. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, volume 25, pages 1097–1105, 2012.

[Radford et al. 2015] Radford Alec, Metz Luke, and Chintala Soumith. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In ICLR, 2016.

[He et al. 2016] He Kaiming, Zhang Xiangyu, Ren Shaoqing, and Sun Jian Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[Zhang et al. 2019] Zhang Han, Goodfellow Ian, Metaxas Dimitris, and Odena Augustus. Self-attention generative adversarial networks. In International conference on machine learning, pages 7354–7363, PMLR.

[Vaswani et al. 2017] Vaswani Ashish, Shazeer Noam, Parmar Niki, Uszkoreit Jakob, Jones Llion, Gomez Aidan N, Kaiser, Łukasz, and Polosukhin Illia. Attention is all you need. In Advances in Neural Information Processing Systems, pages 5998-6008, 2017.

[Wang et al. 2018] Wang Xiaolong, Girshick Ross, Gupta Abhinav, and He Kaiming. Non-local neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7794–7803, 2018.

[Parmar et al. 2018] Parmar Niki, Vaswani Ashish, Uszkoreit Jakob, Kaiser Lukasz, Shazeer Noam, Ku Alexander, and Tran Dustin. Image Transformer. In Proceedings of the 35th International Conference on Machine Learning, pages 4055–4064. PMLR, Jul, 2018.

[Child et al. 2019] Child Rewon, Gray Scott, Radford Alec, and Sutskever Ilya. Generating Long Sequences with Sparse Transformers. In CoRR, May, 2019.

[Dosovitskiy et al. 2021] Dosovitskiy Alexey, Beyer Lucas, Kolesnikov Alexander, Weissenborn Dirk, Zhai Xiaohua, Unterthiner Thomas, Dehghani Mostafa, Minderer Matthias, Heigold Georg, Gelly Sylvain, Uszkoreit Jakob, and Houlsby Neil. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale In International Conference on Learning Representations, 2021

[Olga et al. 2015] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. In International Journal of Computer Vision (IJCV) 3, volume 3, pages 4171–4186, 2015.

[Devlin et al. 2019] Devlin Jacob, Chang Ming-Wei, Lee Kenton, and Toutanova Kristina. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, volume 1, pages 4171–4186, Jun, 2019.

[Wu et al. 2020] Wu Bichen, Xu Chenfeng, Dai Xiaoliang, Wan Alvin, Zhang Peizhao, Tomizuka Masayoshi, Keutzer Kurt, and Vajda Péter In ArXiv, volume abs/2006.03677, 2020. volume=abs/2006.03677

[Yuan et at. 2021] Yuan Li, Chen Yunpeng, Wang Tao, Yu Weihao, Shi Yujun, Jiang Zi-Hang, Tay Francis E.H., Feng Jiashi, and Yan Shuicheng. Tokens-to-Token ViT: Training Vision Transformers From Scratch on ImageNet. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 558–567, October, 2021.

[Kingma and Ba. 2015] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In 3rd International Conference on Learning Representations, ICLR, May, 2015.