論文紹介: Image as Set of Points (ICLR'23)

2023-05-31 17:00 Hachiuma Ryo
AI, DataAnalytics

目次 [非表示]

はじめに

この記事では、International Conference on Learning Representations (ICLR) 2023のOral presentationに採択されたImage as Set of Points [1]の論文を整理に解説いたします。 なお、特に断りのない限り、記事中の画像は論文のものを引用しています。また、コードはこちらに公開されています。

概要

深層学習を用いた画像認識タスクは、畳み込みニューラルネットワーク(CNN)を用いた技術が多く提案され、近年ではTransformerが広く用いられ、MLP-Mixer[2]などAttentionを用いない手法など多く提案されています。

その中で、この論文では、2次元画像を点群データをみなし3次元点群深層学習技術を応用したネットワークを提案しています。その結果として、画像認識の複数タスク(Image classification, semantic segmentation, object detection)において、同等のパラメータ数を持つCNNやTransformerベースの手法と比較し優れていることを検証していました。また、それだけではなく3次元点群物体の識別タスクに応用しても従来手法より優れた認識精度であることを検証しています。

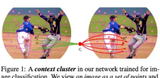

Figure 1: A *context cluster* in our network trained for image classification. We view *an image as a set of points* and sample c centers for points clustering. Point features are aggregated and then dispatched within a cluster. For cluster center C_i, we first aggregated all points $\{x_1^E, x_2^E, \cdots, x_i^E\}$ in ith cluster, then the aggregated result is distributed to all points in the clusters dynamically. See § 3 for details.

手法

提案手法の概要が下図に示されています。ここで特徴的なのは、入力が3次元ベクトルのn個の集合で表されているということです。3次元は、各画素のx, y座標、R, G, B値から構成されています。そのため、点群は画像の縦・横の画素数分となっています。 この3次元点群データから段階的に特徴抽出と集約を繰り返すことで、画像からの効率的な特徴抽出を実現しています。

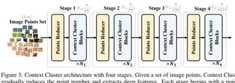

Figure 3: Context Cluster architecture with four stages. Given a set of image points, Context Cluster gradually reduces the point number and extracts deep features. Each stage begins with a points reducer, after which a succession of context cluster blocks is used to extract features.

このネットワークでは、Point Reducerと呼ばれる点群を削減するBlockとContext Cluster Blockという特徴点をクラスタリングし、そのクラスタ毎に特徴計算をするBlockから成っています。 このPoint Reducerは点群からアンカーを均等に定め、近傍点をlinear embeddingする処理となっており、画像の場合はVision TransformerのPatch embeddingと同等の処理が行われます。

Context Cluster Blockは下図のようになっています。

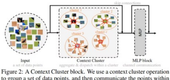

Figure 2: A Context Cluster block. We use a context cluster operation to group a set of data points, and then communicate the points within clusters. An MLP block is applied later.

2つ目のMLP BlockはTransformer等同様、点毎に特徴抽出をする処理であり、1つ目のContext Clusterが主な提案となっています。
Context Clusterでは、特徴点の集合を入力とし、点間の類似度を元にクラスタリングし、そのクラスタリングされた領域毎に各点の特徴を更新するような処理を行っています。

その際のクラスタリングでは、画像上に均等に配置されたc個の中心点の特徴ベクトルをc近傍点の特徴ベクトルの平均を取り計算し、各中心点と各点の特徴ベクトルの類似度を計算した行列 $\mathbf{S} \in \mathbb{R}^{n \times c}$を作成し、各画素で最も類似度が高くなる 中心点にアサインすることで実現しています。

実験結果

実験では、画像認識の様々なタスクや3次元点群識別タスクで従来手法と比較をしています。実験結果を抜粋して紹介します。

下表では、ImageNet-1Kのベンチマーク結果となっています。 一番下が提案手法の結果となっています。 パラメータ数がおおよそ同等のResNet50、PVT-Smallと比較し、提案手法のMediumは認識性能が優れていることがわかりました。

Table 1: Comparison with representative backbones on ImageNet-1k benchmark. Throughput (images / s) is measured on a single V100 GPU with a batch size of 128, and is averaged by the last 500 iterations. All models are trained and tested at 224×224 resolution, except ViT-B and ViT-L.

	Method	Param.	GFLOPs	Top-1	Throughput (images/s)
MLP	♠ ResMLP-12 (Touvron et al., 2022)	15.0	3.0	76.6	631.4
	♠ ResMLP-24 (Touvron et al., 2022)	30.0	6.0	79.4	509.1
	♠ ResMLP-36 (Touvron et al., 2022)	45.0	8.9	79.7	452.9
	♠ MLP-Mixer-B/16 (Tolstikhin et al., 2021)	59.0	12.7	76.4	400.8
	♠ MLP-Mixer-L/16 (Tolstikhin et al., 2021)	207.0	44.8	71.8	125.2
	♠ gMLP-Ti (Liu et al., 2021a)	6.0	1.4	72.3	511.6
	♠ gMLP-S (Liu et al., 2021a)	20.0	4.5	79.6	509.4
Attention	♥ ViT-B/16 (Dosovitskiy et al., 2020)	86.0	55.5	77.9	202.0
	♥ ViT-L/16 (Dosovitskiy et al., 2020)	307.0	190.7	76.5	92.8
	♥ DeiT-Ti (Touvron et al., 2021)	5.7	1.3	72.2	-
	♥ PVT-Small (Wang et al., 2021)	24.5	3.8	79.8	-
	♥ T2T-ViT-7 (Yuan et al., 2021a)	4.3	1.1	71.7	-
	♥ DeiT-Tiny/16 (Touvron et al., 2021)	5.7	1.3	72.2	523.8
	♥ DeiT-Small/16 (Touvron et al., 2021)	22.1	4.6	79.8	521.3
	♥ Swin-T (Liu et al., 2021b)	29	4.5	81.3	-
Context Modulation	♣ ResNet18 (He et al., 2016)	12	1.8	69.8	584.9
	♣ ResNet50 (He et al., 2016)	26	4.1	79.8	524.8
	♣ ConvMixer-512/16 (Trockman et al., 2022)	21.1	-	73.8	-
	♣ ConvMixer-768/32 (Trockman et al., 2022)	14.6	-	75.8	-
	♣ ConvMixer-768/32 (Trockman et al., 2022)	21.1	-	80.16	142.9
	♣ Context-Cluster-Ti (ours)	5.3	1.0	71.8	518.4
	♣ Context-Cluster-Til (ours)	5.3	1.0	71.7	510.8
	♣ Context-Cluster-Small (ours)	14.0	2.6	77.5	513.0
	♣ Context-Cluster-Medium (ours)	27.9	5.5	81.0	325.2

また、入力のposition information, context cluster block, multi-headに対するAblation studyの結果が以下となっています。 position情報がない場合はそもそも学習ができなかったようで、context cluster blockが3.3%の精度向上に貢献していることがわかります。

Table 2: Component ablation studies of Context-Cluster-Small on ImageNet-1k.

position info.	context cluster	multi head	top-1 acc.
✓	✗	✗	74.2 (3.3)
✓	✓	✗	76.6 (0.9)
✓	✓	✓	**77.5**

さいごに

本記事では、画像を点群データとみなし特徴抽出をする新しい論文について紹介させていただきました。 Transformer, MLP-Mixer, Vision GNN[3]など様々なアーキテクチャが提案されていますが、本記事のように点群データとみなす手法も1つの選択となるのでしょうか。

本記事が皆様にとって有益な情報であれば、幸いです。

今後もコニカミノルタAI技術開発部では社会実装まで見据えた技術選定を意識していくことで、価値のあるサービスを提供していけるよう心がけていきます。

引用

[1] Xu Ma., et al. Image as Set of Points, ICLR 2023.

[2] Tolstikhin, Ilya O., et al. Mlp-Mixer: An All-mlp Architecture for Vision., NeurIPS 2021.

[3] Han, Kai, et al. Vision GNN: An Image is Worth Graph of Nodes. NeurIPS 2022.

コニカミノルタは画像IoTプラットフォームFORXAIを通じて、お客様やパートナー様との共創を加速させ、技術・ソリューションの提供により人間社会の進化に貢献してまいります。

中途採用に関する情報については以下の採用情報ページをご覧ください。

Hachiuma Ryo
FORXAI事業開発部 AI技術開発部 所属 入力動画を中心とする強化学習の光学レ開発などを行っています

前の記事
人の感性を見える化する！EX感性を使ってみた
次の記事
論文紹介：Reinforcement Learning from Human Feedback