Boosting vision transformers for image retrieval

Chull Hwan Song[1], Jooyoung Yoon[1], Shunghyun Choi[1], Yannis Avrithis[2,3]

[1]Dealicious, INC [2]Institute of Advanced Research on Artificial Intelligence (IARAI) [3]Athena RC

Introduction

- In instance-level image retrieval, vision transformers have not yet shown good performance compared to convolutional networks
- **Goal:** Improve their performance, without introducing a new architecture
- We show that a hybrid architecture is more effective than plain transformers
- We build a global representation by an advanced pooling mechanism over token embeddings

Contributions

- Collect global & local features from `[CLS]` & patch tokens respectively of multiple layers
- Dynamic position embedding (DPE) to handle dynamic image size at training
- Enhanced locality module (ELM) to investigate inductive bias in the deeper layers
- Training on all common datasets: NC-clean, SfM-120k, GLDv1-noisy, GLDv2-clean
- State of the art on image retrieval using vision transformers for the first time

Deep Token Pooling (DToP)

- Transformer encoder with L layers, each of $M = w \times h$ patch tokens
- Mapping of layer ℓ for $\ell = 1, \dots, L$

$$Z^\ell = f^\ell(Z^{\ell-1}) = [\mathbf{z}^\ell_{\texttt{[CLS]}}; \mathbf{z}^\ell_1; \dots; \mathbf{z}^\ell_M] \in \mathbb{R}^{(M+1)\times D}$$

- $A^\ell \in \mathbb{R}^{w\times h\times D}$: sequence $\mathbf{z}^\ell_1, \dots, \mathbf{z}^\ell_M$ of patch token embeddings of layer ℓ, unfolded into $w \times h \times D$ tensor
- Given $k \in \{1, \dots, L\}$, collect multi-layer `[CLS]` and patch features from the last k layers

$$F_c = [\mathbf{z}^{L-k+1}_{\texttt{[CLS]}}; \dots; \mathbf{z}^L_{\texttt{[CLS]}}] \in \mathbb{R}^{k\times D}$$
$$F_p = [A^{L-k+1}; \dots; A^L] \in \mathbb{R}^{k\times w\times h\times D}$$

- Global branch: multi-layer `[CLS]` features F_c mapped to N-dimensional space

$$\mathbf{u}_c = \mathrm{FC}(F_c) \in \mathbb{R}^N$$

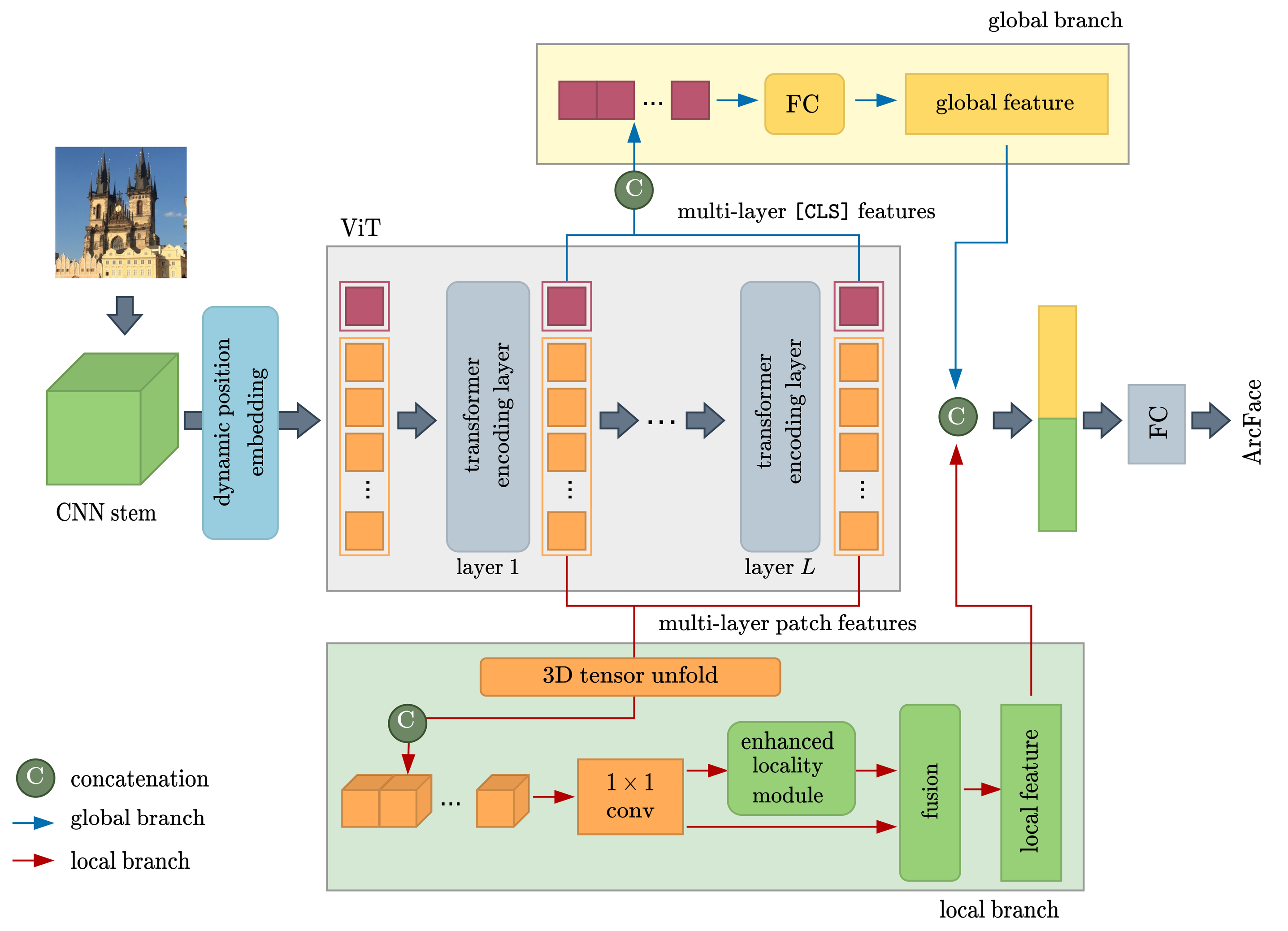
- Local branch: multi-layer patch features F_p processed by convolution operations across layers to enhance locality of interactions, followed by global average pooling

$$Y = \mathrm{conv}_{1\times1}(F_p) \in \mathbb{R}^{w\times h\times D}$$
$$Y' = \mathrm{FUSE}(Y, \mathrm{ELM}(Y)) \in \mathbb{R}^{w\times h\times D}$$
$$\mathbf{u}_p = \mathrm{FC}(\mathrm{GAP}(Y')) \in \mathbb{R}^N$$

- Image representation: concatenated global and local features $\mathbf{u}_c, \mathbf{u}_p$ mapped to N-dimensional space

$$\mathbf{u} = \mathrm{BN}(\mathrm{FC}(\mathrm{DROPOUT}([\mathbf{u}_c; \mathbf{u}_p]))) \in \mathbb{R}^N$$

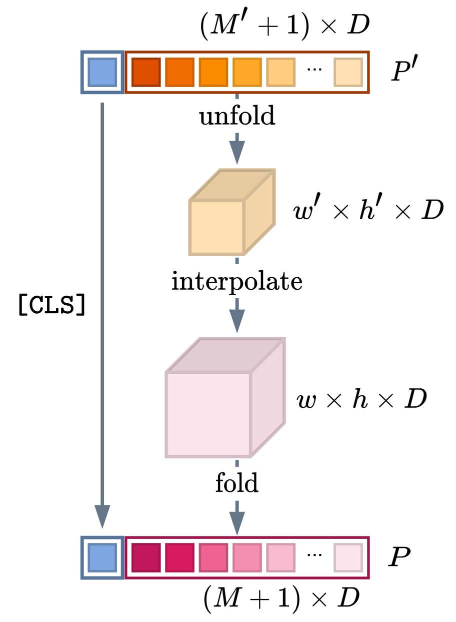
DToP architecture



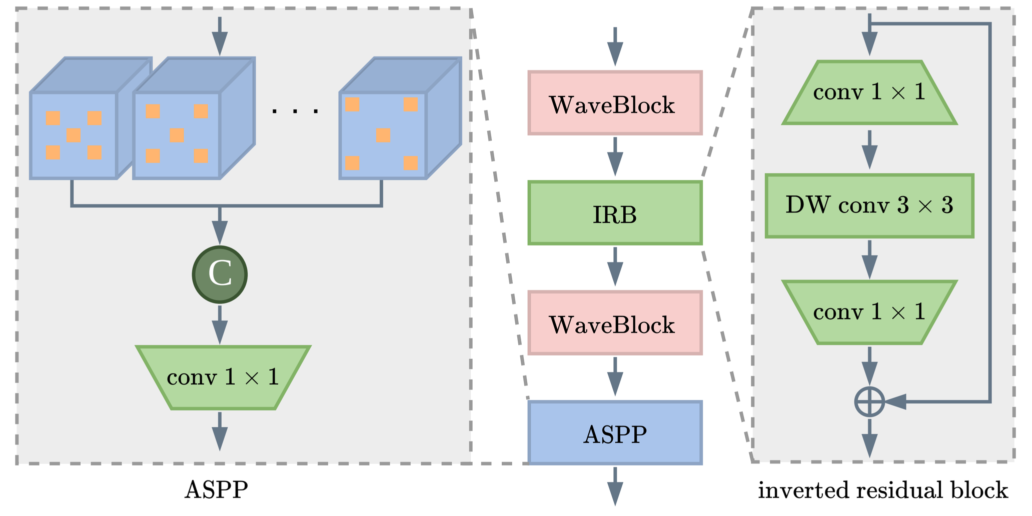
State of the art comparisons

Method	Medium ROxf mAP	Medium ROxf mP@10	Medium ROxf+R1M mAP	Medium ROxf+R1M mP@10	Medium RPar mAP	Medium RPar mP@10	Medium RPar+R1M mAP	Medium RPar+R1M mP@10	Hard ROxf mAP	Hard ROxf mP@10	Hard ROxf+R1M mAP	Hard ROxf+R1M mP@10	Hard RPar mAP	Hard RPar mP@10	Hard RPar+R1M mAP	Hard RPar+R1M mP@10
Global Descriptors (SfM-120k)																
RMAC-R101 ‡	53.5	76.9	–	–	68.3	97.7	–	–	25.5	42.0	–	–	42.4	83.6	–	–
GeM-R101	64.7	84.7	45.2	**71.7**	77.2	**98.1**	52.3	**95.3**	38.5	53.0	19.9	34.9	56.3	**89.1**	24.7	**73.3**
AGeM-R101	67.0	–	–	–	78.1	–	–	–	40.7	–	–	–	57.3	–	–	–
SOLAR-R101 †	52.5	73.6	–	–	70.9	**98.1**	–	–	27.1	41.4	–	–	46.7	83.6	–	–
GeM-R101 †	54.0	72.5	–	–	64.3	92.6	–	–	25.8	42.2	–	–	36.6	67.6	–	–
GLAM-R101 ‡	66.2	–	–	–	77.5	–	–	–	39.5	–	–	–	54.3	–	–	–
DOLG-R101 †	46.4	66.8	–	–	56.6	91.1	–	–	18.1	27.9	–	–	26.6	62.6	–	–
IRT-DeiT-B	55.1	–	–	–	72.7	–	–	–	28.3	–	–	–	49.6	–	–	–
DToP-R50+ViT-B	**68.5**	**85.4**	**48.9**	**71.7**	**83.1**	96.4	**56.5**	94.0	**43.0**	**56.9**	**24.7**	**38.9**	**65.8**	**89.1**	**30.3**	69.6
Global Descriptors (GLDv2-clean)																
GeM-R101	76.2	–	–	–	87.3	–	–	–	55.6	–	–	–	74.2	–	–	–
GLAM-R101	78.6	88.2	68.0	82.4	88.5	**97.0**	73.5	94.9	60.2	72.9	43.5	62.1	76.8	93.4	53.1	84.0
DELG-GeM-R50	73.6	–	60.6	–	85.7	–	68.6		51.0	–	32.7	–	71.5	–	44.4	–
DELG-GeM-R101	76.3	–	63.7	–	86.6	–	70.6		55.6	–	37.5	–	72.4	–	46.9	–
DOLG-R50	80.5	–	76.6	–	89.8	–	80.8	–	58.8	–	52.2	–	77.7	–	62.8	–
DOLG-R101	81.5	–	**77.4**	–	91.0	–	**83.3**	–	61.1	–	**54.8**	–	80.3	–	**66.7**	–
DOLG-R101 ∃	78.8	91.6	64.2	82.1	87.8	96.6	68.7	94.1	58.0	74.8	37.3	57.7	74.1	91.1	45.1	80.0
DToP-R50+ViT-B	**82.1**	**91.7**	70.9	**83.9**	**92.0**	96.6	81.9	**96.4**	**64.5**	**77.4**	49.0	**66.6**	**82.9**	**94.3**	64.0	**90.6**
DOLG-R101 ∃□	79.3	93.2	71.3	89.1	89.2	98.9	74.7	97.7	57.2	73.0	43.4	62.6	76.6	94.1	53.6	89.7
DToP-R50+ViT-B□	84.4	94.1	78.9	91.3	92.3	97.1	85.4	96.9	64.8	76.7	57.1	72.1	84.6	95.4	71.2	94.6

Dynamic Position Embedding & Enhanced Locality Module

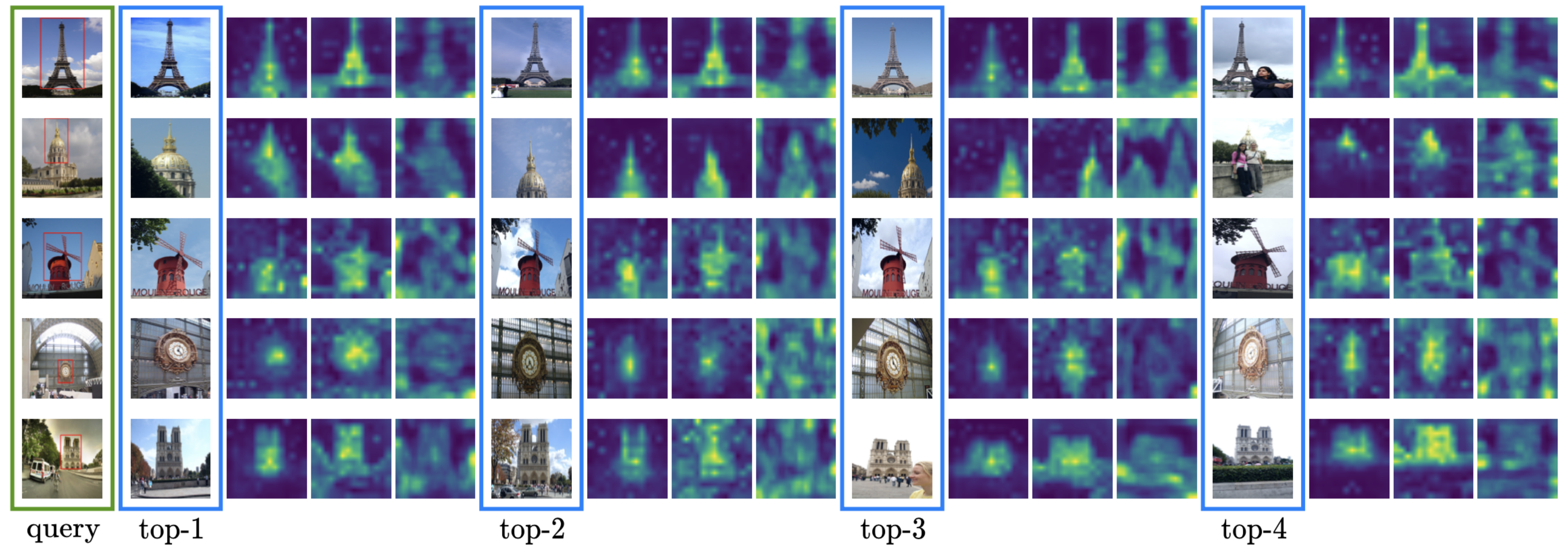


Dynamic Position Embedding (DPE)



Enhanced Locality Module (ELM)

Top-4 ranking and spatial attention



Ablation study on SfM-120k

PE Type	Oxf5k	Par6k	Medium ROxf	Medium RPar	Hard ROxf	Hard RPar
no PE	82.8	85.7	59.7	73.9	32.5	47.4
CPE [11]	85.9	88.8	62.6	77.9	37.1	58.2
DPE (bi-cubic)	87.6	91.0	65.2	82.2	38.3	64.6
DPE (bi-linear)	**89.7**	**92.7**	**68.5**	**83.1**	**43.0**	**65.8**

mAP: position embedding

CNN stem	Global branch	Local branch	ELM	Oxf5k	Par6k	Medium ROxf	Medium RPar	Hard ROxf	Hard RPar
				77.7	85.9	52.6	76.0	26.6	52.0
	✓			76.6	87.3	54.7	77.0	27.7	54.8
	✓	✓		78.3	89.7	57.9	78.2	24.2	54.4
	✓	✓	✓	81.5	89.8	61.4	79.7	32.5	57.4
✓				81.2	86.4	55.5	76.2	31.4	52.1
✓	✓			88.3	91.9	66.6	**83.6**	41.9	**67.8**
✓	✓	✓		**89.8**	91.2	67.6	81.1	40.7	62.5
✓	✓	✓	✓	89.7	**92.7**	**68.5**	83.1	**43.0**	65.8

mAP: algorithmic components