CodedVTR: **Code**book-based Sparse **V**oxel **Tr**ansformer with Geometric Guidance

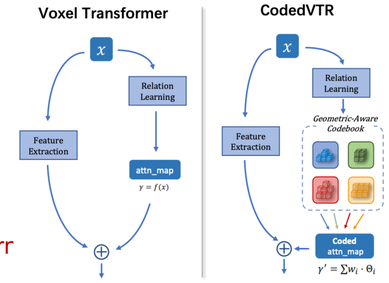
Zhao Tianchen[1], Zhang Niansong[1], Ning Xuefei[1], Wang He[2], Yi Li[1,3]*, Wang Yu[1]
1 Tsinghua University 2 Peking University

Project Page

CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

Motivation

- ❓ : How to Adapt Transformer to 3D Domain?
- **Transformer's Property?**
  - ✅: Less **inductive bias**
  - ✅: Better representation power
  - ❌: Harder to generalize
- **3D Data's Property?**
  - **Irregular** data structure
  - **Limited** data scale
  - Varying density & sparse patterr



🔑 **Key:** Alleviate the aggravated generalization issue with domain-specific inductive bias

- **Generalization Issue of Transformer**
- Transformer relies on large-scale pretraining / additional inductive bias to outperform CNN. Recent studies attribute it to the **Generalization Issue.**

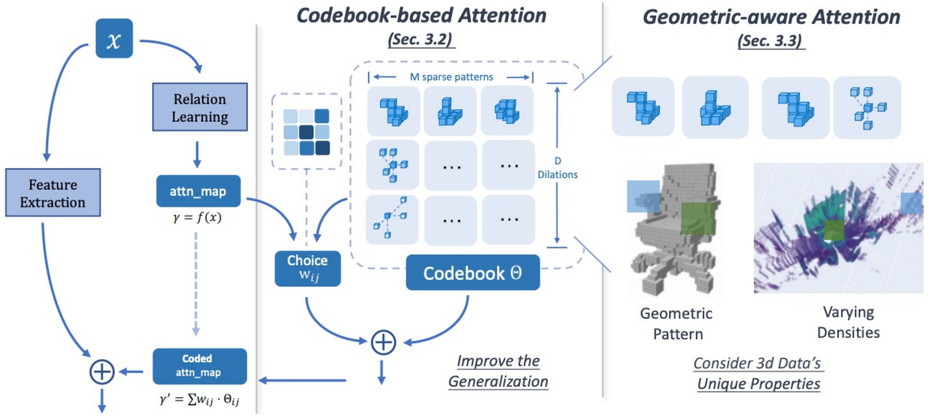
> *" When directly trained on the ImageNet, ViT yields modest accuracies of a few points below ResNets of comparable size"*[1]

Dataset	Method (Model)		Params	mIOU
ScanNet	Convolution	Minkowski-M	7M	67.3%
		Minkowski-L	11M	72.4%
	Transformer	PointTransformer	6M	58.6% (-8.7%)
		VoTR (Mink-M)	7M	62.5%(-4.8%)
		VoTR (Mink-L)	11M	66.1%(-6.3%)
SemanticKITTI	Convolution	Minkowski-M	7M	58.9%
		Minkowsk-L	11M	61.1%
	Transformer	VoTR (Mink-M) †	7M	56.5%(-2.4%)
		VoTR (Mink-L)	11M	58.2%(-2.9%)

(Plain 3D Transformer fails to outperform Convolution)

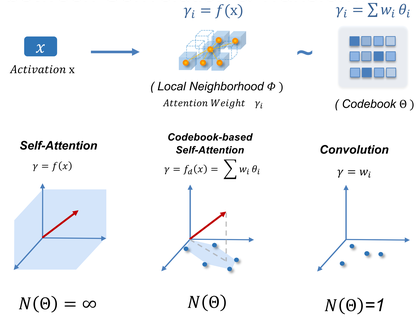
[1] Dosovitskiy, Alexey et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." *ArXiv* abs/2010.11929 (2021): n. pag.

Methodology



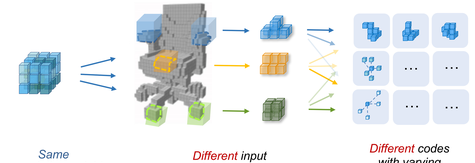
Codebook-based Self-Attention

- Project the attention space into a subspace spanned by codebook "prototypes"
- Work as Regularization for better Generalization
- Could be viewed as an intermediate state between Convolution & Transformer



Geometric-aware Self-Attention

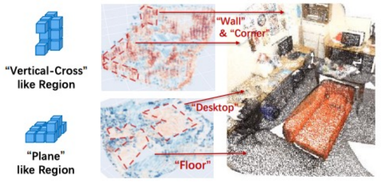
- Consider 3D data's unique properties:
  - Unique geometric shape
  - Varying densities
- Assign different geometric shape as attention span for codebook elements
- Guide the attention learning by encouraging the choice of codes with matching sparse pattern



Performances

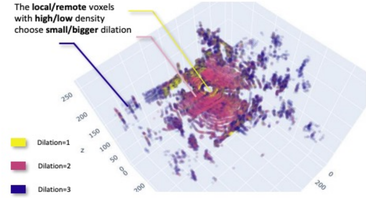
Dataset	Method (Model)		Param	mIOU
ScanNet	Convolution	Minkowski-M [18]	7M	67.3%
		Minkowski-L [3]	11M	72.4%
	Transformer	PointTransformer [30]†	6M	58.6%
		VoTR (Mink-M) [17]†	7M	62.5%
		VoTR (Mink-L) [17]†	11M	66.1%
		CodedVTR (Mink-M)	7M	**68.8%**
		CodedVTR (Mink-L)	11M	**73.0%**
SemanticKITTI	Convolution	Minkowski-M [21]	7M	58.9%
		Minkowsk-L [21]	11M	61.1%
		SPVCNN [21]	8M	60.7%
	Transformer	VoTR (Mink-M) [17] †	7M	56.5%
		VoTR (Mink-L) [17]†	11M	58.2%
		CodedVTR (Mink-M)	7M	**60.4%**
		CodedVTR (Mink-L)	11M	**63.2%**
		CodedVTR (SPVCNN)	8M	**61.8%**
Nuscenes	Convolution	Minkowski-M [21]	7M	66.5%
		Minkowsk-L [21]	11M	69.4%
	Transformer	CodedVTR (Mink-M)	7M	**69.9%**
		CodedVTR (Mink-L)	11M	**72.5%**

CodedVTR consistently outperform Conv & Transformer
(CodedVTR is compatible with current sparse conv methods e.g., SPVCNN)

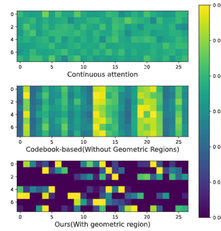


Codebook Design		mIoU
D	M	
1	1	57.1%
3	1	58.5%
4	1	55.3%
3 (RS)	8 (RS)	54.2%
3	8	**60.4%**
3	16	58.1%

Ablation study: codebook size



CodedVTR learns to choose codes with matching **sparse pattern**



CodedVTR prevents the **"attention collapse"**