CodedVTR: Codebook-based Sparse Voxel Transformer with Geometric Guidance

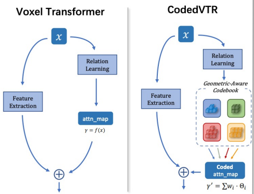
Zhao Tianchen, Zhang Niansong, Ning Xuefei, Wang He, Yi Li*, Wang Yu

*Corresponding Author

How to Adapt Transformer to 3D Domain?

Q: Transformer's Property?
- Less **inductive bias**
- Better representative power
- Harder generalization

Q: 3D Domain-Specific Problem?
- **Irregular** data structure
- **Limited** data Scale
- **Harder generalization**



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

Our Contributions: ***New Attention Scheme***
- Codebook: "Encode" attn-map to **regularize** attn learning space
- Geometric-aware: Utilize geometric-info to **guide** attn learning
- Could be embedded into sparse-conv-based methods

Generalization Issue of Transformer

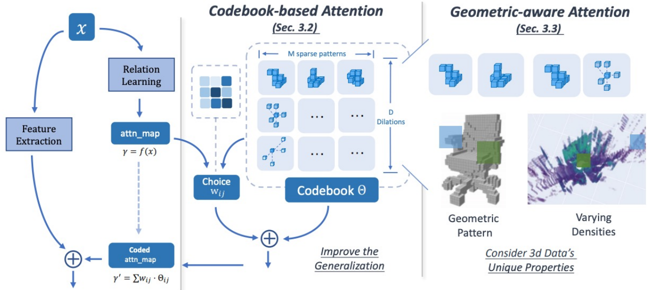
Transformer relies on large-scale pretraining / additional inductive bias to outperform CNN. Recent studies attribute it to the **Generalization Issue**.

> *" 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%)

(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.



Codebook-based Self-Attention

- **Project** the Attention Space into a **Subspace** spanned by Codebook "**Prototypes**"
- Work as **Regularization** for better **Generalization**

Dataset	Method (Model)		Params	mIOU
ScanNet	Convolution	Minkowski-M	7M	67.3%
		Minkowski-L	11M	72.4%
	Transformer	CodedVTR (Mink-M)	7M	**68.8%**(+1.5%)
		CodedVTR (Mink-L)	11M	**73.0%**(+0.6%)
SemanticKITTI	Convolution	Minkowski-M	7M	58.9%
		Minkowsk-L	11M	61.1%
		SPVCNN	8M	60.7%
	Transformer	CodedVTR (Mink-M)	7M	**60.4%** (+0.5)
		CodedVTR (Mink-L)	11M	**63.2%** (+2.1%)
		CodedVTR (SPVCNN)	8M	**61.8%**(+1.1%)
Nuscenes	Convolution	Minkowski-M	7M	66.5%
		Minkowsk-L	7M	69.4%
	Transformer	CodedVTR (Mink-M)	7M	**69.9%** (+3.4%)
		CodedVTR (Mink-L)	11M	**72.5%** (+3.1%)

(CodedVTR outperforms Convolution as Dataset-size Scale-up)

(CodedVTR Block could be embedded into current Sparse Convolution based Methods, e.g., SPVCNN)

Geometric-Aware Self-Attention

- Different **Shapes/Dilations** for Codebook Element
- Encourage the Attention to Choose "Prototype" that **matches the real sparse pattern**