Bundle-Transformer: A Geometric Position Paper

Charles De Clercq*

September 24, 2026

Abstract

In this note we describe a novel geometric re-interpretation of the Transformer architecture, where tokens are not represented by vectors in a Euclidean space, but as elements of the total space of the tautological bundle over a Grassmannian (or a flag manifold). This geometric bias is intended to factorize the token into a *concept* (a subspace) and a *contextual instance* (a vector lying in that subspace). All operations of the Transformer (attention, feed-forward, residual connections) are redesigned as bundle morphisms, thereby preserving the geometric structure throughout the model. This position paper outlines the motivations, mathematical foundations, and key architectural choices to stimulate discussion and invite collaboration towards a full implementation and empirical validation.

1 Introduction

The Transformer architecture [3] has become ubiquitous in sequence processing, from natural language to vision and beyond. Its core mechanism, self-attention, computes pairwise similarities between tokens in a sequence, enabling each token to gather contextual information from all others. Despite its success, the representation of a token as a plain vector in $\mathbb{R}^d$ is geometrically unstructured: it encodes both the invariant identity of the token and its contextual variations in a flat space, which may hinder interpretability, generalization and data efficiency.

Recent works have started to inject geometric structure into Transformers. For instance, [1] projects attention features onto Grassmannian subspaces to enforce low-rank constraints, while [4] replaces attention entirely by flows on Grassmannians. However, these approaches still treat tokens as vectors and use the Grassmannians only as a tool for interactions or regularization.

In this position paper, we advocate a more radical step: make the token representation itself geometric. Each token is represented by a point of the total space of the tautological bundle over a Grassmannian, that is a pair (Z, v) where Z is a subspace representing an invariant concept, and v is a vector lying in that subspace representing the contextual instance. This factorization is enforced architecturally via a differentiable QR decomposition and all subsequent operations are designed as morphisms respecting the bundle structure, to keep track of the underlying geometry.

We call this proposal Bundle-Transformer. The paper is structured as follows. Section 2 introduces the geometric background of Grassmannian varieties, endowed with their Riemannian and chordal metric, as well as the tautological bundle and its total space. Section 3 details the architecture of Bundle-Transformer: how tokens are encoded, how attention is computed using a hybrid similarity (instance similarity + chordal distance between subspaces), how the feed-forward block is implemented as a bundle morphism composing a linear transport with a fibre-wise non-linear transformation. Section 4 discusses complexity and design summary of the architecture.

*Université Sorbonne Paris Nord, LAGA, Équipe Topologie Algébrique. Email: declercq@math.univ-paris13.fr

2 Geometric foundations: the tautological bundle

2.1 Grassmann varieties

The Grassmannian $\mathrm{Gr}(k,n)$ of k-dimensional subspaces of $\mathbb{R}^n$ is a smooth compact manifold. It admits several descriptions, the two most classical being:

- as a Riemannian homogeneous space $\mathrm{Gr}(k,n) \cong O(n)/(O(k)\times O(n-k))$, endowed with an intrinsic Riemannian metric;
- as a projective algebraic variety via the Plücker embedding $\mathrm{Gr}(k,n) \hookrightarrow \mathbb{P}(\bigwedge^k \mathbb{R}^n)$.

From the first viewpoint arises the geodesic Riemannian distance between points of $\mathrm{Gr}(k,n)$. However meaningful and natural, the geodesic distance is computationally challenging :

- The geodesic distance d_{geo} is given by principal angles $\theta_1,\dots,\theta_k$:

$$d_{\mathrm{geo}}(Z_1,Z_2) = \left(\theta_1^2+\cdots+\theta_k^2\right)^{1/2}.$$

On the other hand, looking at Grassmannians as algebraic subsets of symmetric matrices, another metric, the chordal distance d_{chord}, can be considered. Although the chordal distance is extrinsic, it can be used to approximate geodesics, as explained below, and is computationally much more manageable.

- The chordal distance d_{chord} is defined (taking orthonormal basis for each subspaces) by

$$d_{\mathrm{chord}}(Z_1,Z_2) = \frac{1}{\sqrt{2}}\|\Pi_{Z_1}-\Pi_{Z_2}\|_F = \sqrt{k-\|Z_1^\top Z_2\|_F^2}.$$

  where Π_* are the respective orthogonal projectors and $\|\cdot\|_F$ denotes the Frobenius norm.

These two metrics are equivalent in the sense that they induce the same topology and lead to the same infinitesimal structure. For our purposes of grouping tokens with respect to concept similarity and performing locality-sensitive hashing, both are thus equivalent. We choose to work with the chordal distance, which is computationally convenient and for which the $\mathrm{Gr}(k,n)$ is bounded by $\sqrt{k}$.

2.2 Tautological bundle on Grassmannians

Let $\mathrm{Gr}(k,n)$ be the Grassmannian of k-dimensional subspaces of $\mathbb{R}^n$. A point $Z \in \mathrm{Gr}(k,n)$ can be represented by an orthonormal matrix $Z \in \mathbb{R}^{n\times k}$ with $Z^\top Z = I_k$. The (total space of the) tautological bundle[1] $\mathcal{T}$ is defined as

$$\mathcal{T} = \{(Z,v) \mid Z \in \mathrm{Gr}(k,n),\, v \in Z \subset \mathbb{R}^n\}.$$

The fibre of the tautological bundle over the point $Z \in \mathrm{Gr}(k,n)$ is the subspace Z itself.

2.3 Relation with classical transformers

In the sequel, we interpret (Z,v) as a token: while Z encodes concepts (meaningful directions), the vector v is meant to encode its contextual instance. Note that in the trivial case $k=n$, the Grassmannian $\mathrm{Gr}(k,n)$ boils down to the ambient space $\mathbb{R}^n$ and the bundle representation (Z,v) becomes equivalent to a plain vector in $\mathbb{R}^n$.

[1] In the sequel we will often use "bundle morphism" in a loose sense for any map of the total space that respects the geometric context, even if it is not strictly a vector-bundle morphism.

In this $k = n$ context, the geometric attention mechanism built in this note projects the aggregated context onto $\mathbb{R}^n$ (via the identity) and updates v accordingly, matching the standard self-attention. The feed-forward network provided in the sequel reduces to the standard two-layer MLP applied to v with a residual connection (see Section 3.4). Thus, the Bundle-Transformer is a strict geometric generalization of classical transformers, imposing structure and geometric bias if $k < n$. The point of Bundle-transformers is to bring a geometrically meaningful architecture to open a door to more interpretable and potentially more efficient representations.

3 Bundle-Transformer architecture

3.1 Token encoding

Given an input token, we first compute a hidden representation $h \in \mathbb{R}^{d_h}$ via a shallow MLP. We then use two linear heads:

- a head producing a rank k matrix $M \in \mathbb{R}^{n\times k}$,[2]
- a head producing a vector $\alpha \in \mathbb{R}^k$ of fibre coordinates.

Let M be a matrix of rank k (attached to a point $Z \in \mathrm{Gr}(k, n)$). Writing $M = QR$ the QR decomposition of M, set $Z := \mathrm{QR}(M) = Q$, so that Z is represented by an orthonormal basis, and set $v = Z\alpha \in Z$. The vector v belongs to the subspace $Z \in \mathrm{Gr}(k, n)$, by construction.

3.2 Geometric attention

We aim at implementing an hybrid geometric attention, with respect to both Z (concept) and v (instance). Given two tokens (Z_i, v_i) and (Z_j, v_j), we introduce the two following similarity scores:

- **Instance similarity**: as for classical transformer architecture, given usual query, key, value matrices, set
$$s_{ij}^{\text{inst}} = \frac{(W^Q v_i)^\top (W^K v_j)}{\sqrt{d}}$$
where d denotes the query/key dimension ($W^Q \in \mathbb{R}^{d\times n}$, $W^K \in \mathbb{R}^{d\times n}$, $W^V \in \mathbb{R}^{n\times n}$).
- **Concept similarity**: we use chordal distance to detect similarity between subspaces
$$s_{ij}^{\text{concept}} = \gamma \exp\big(\frac{-d_{\text{chord}}(Z_i, Z_j)^2}{\sigma}\big)$$
where γ and σ are hyperparameters (roughly speaking, weight and temperature).

The total score is
$$s_{ij} = \lambda_{\text{inst}}\, s_{ij}^{\text{inst}} + \lambda_{\text{concept}}\, s_{ij}^{\text{concept}},$$
and attention weights a_{ij} are obtained by softmax. The weights λ_{inst} and λ_{concept} control the relative importance of instance and concept similarities. They may be set as fixed hyperparameters, bearing in mind that s_{ij}^{concept} is bounded (the chordal distance on $\mathrm{Gr}(k, n)$ is at most $\sqrt{k}$) whereas s_{ij}^{inst} is not, so the relative scaling requires care. Alternatively, they may be learnt per layer and per attention head; in that case, since the subspace component is meant to stabilise over the course of training, the evolution of the ratio $\lambda_{\text{concept}}/\lambda_{\text{inst}}$ across layers provides a useful

[2] In practice, matrices produced by trained linear layers are almost surely full rank; any degeneracy is easily detectable through the QR decomposition and is inconsequential for training.

interpretability diagnostic. More expressive variants are also conceivable: for instance, decoupling the two scores entirely by computing separate softmax distributions and combining the resulting attention matrices, or using a partially decoupled scheme in which concept similarity s_{ij}^{concept} is used alone for one set of heads and the combined score $s_{ij}^{\text{inst}} + s_{ij}^{\text{concept}}$ for another. The optimal design will depend on the task and must be determined empirically.

The context for token i, given by

$$c_i = \sum_j a_{ij} W^V v_j \in \mathbb{R}^n,$$

does not necessarily land in the subspace Z_i. We thus consider its projection

$$\tilde{c}_i = \Pi_{Z_i}(c_i) = Z_i Z_i^\top c_i \in Z_i$$

and update v_i through

$$v_i \leftarrow v_i + \tilde{c}_i.$$

This concludes the residual part of the attention process[3].

3.3 Positional encoding

Rotary Position Embedding (RoPE) [2] is perfectly suited to our context, we thus adopt a variant, which respects the bundle structure. For each token (Z, v) we first extract the vector $\alpha := Z^\top v \in \mathbb{R}^k$ of v_i's coordinates in the fixed orthonormal basis of the subspace Z and then apply a position-dependent rotation $R \in SO(k)$ to α:

$$\alpha \leftarrow R\alpha, \qquad v \leftarrow Z\alpha.$$

This corresponds to rotating the instance vector v within its fiber subspace Z. As Z remains unchanged, the operation is a fibre-preserving bundle morphism: it injects relative positional information exclusively into the instance component, leaving the invariant concept untouched. The rotation R is constructed in the same way as standard RoPE, but acting on the k-dimensional coordinate space.

3.4 Geometric feed-forward

The standard Transformer's feed-forward network (FFN) applies the same two-layer MLP independently to each token vector. In our framework, the FFN must respect the bundle structure while fulfilling two requirements: (i) it must be able to *update the concept*, i.e. move the subspace Z_i, since the attention mechanism leaves it unchanged; (ii) when $k = n$, it must reduce to the standard FFN. We achieve both by composing a shared linear transport with a fibre-wise non-linear transformation, and placing the residual connection at the layer level.

Step 1: Linear transport. A shared matrix $A \in \mathbb{R}^{n \times n}$ is applied simultaneously to the subspace and the instance:

$$Z_i^{\text{new}} = \text{QR}(A\, Z_i), \qquad \tilde{v}_i = A\, v_i.$$

Because $v_i \in Z_i$, the transported vector satisfies $\tilde{v}_i \in Z_i^{\text{new}}$: the bundle constraint is preserved without any projection.

[3]Note that the subspace Z_i remains unchanged during the attention mechanism.

Step 2: Fibre-wise MLP. We extract the coordinates of $\tilde{v}_i$ in the new orthonormal basis and apply a two-layer MLP:

$$\alpha_i = (Z_i^{\text{new}})^\top \tilde{v}_i \in \mathbb{R}^k, \qquad \alpha_i' = W_2\,\sigma\big(W_1\,\alpha_i\big),$$

where $W_1 \in \mathbb{R}^{r\times k}$, $W_2 \in \mathbb{R}^{k\times r}$ are learnable, σ is a non-linearity (e.g. GeLU), and r is the intermediate dimension. The FFN output is then reconstructed as

$$v_{\text{ffn}} = Z_i^{\text{new}}\,\alpha_i'.$$

Step 3: Residual connection. The residual connection is placed at the layer level, outside the FFN. Because the subspace may have changed from Z_i to Z_i^{new}, the previous instance vector v_i does not in general belong to Z_i^{new}. We therefore project it onto the new subspace and add it to the FFN output:

$$v_i^{\text{new}} = v_{\text{ffn}} + \Pi_{Z_i^{\text{new}}}(v_i), \qquad \text{where} \quad \Pi_{Z_i^{\text{new}}}(v_i) = Z_i^{\text{new}}(Z_i^{\text{new}})^\top v_i.$$

The projection discards the components of v_i that lie outside the new subspace. This selective filtering is by design: it ensures that only the information compatible with the updated concept is carried forward through the residual path.

Remark : The projection step can be bypassed entirely by placing the residual connection inside the fibre. Since the transported vector $\tilde{v}_i$ already belongs to Z_i^{new}, one may replace Step 2 by $\alpha_i \leftarrow \alpha_i + W_2\,\sigma(W_1\,\alpha_i)$ and omit Step 3.

Recovery of the standard Transformer ($k = n$). When $k = n$, the Grassmannian $\mathrm{Gr}(n, n)$ is a single point and $Z = I_n$. Setting $A = I_n$ (or any orthogonal matrix, which is absorbed by subsequent layers), we get $Z^{\text{new}} = I_n$, $\alpha = \tilde{v} = v$, and the projection reduces to the identity. The update becomes $v \leftarrow W_2\,\sigma(W_1\,v) + v$, which is exactly the standard two-layer FFN with residual connection. The Bundle-Transformer FFN is therefore a strict geometric generalisation of the classical one.

Design properties. This construction fulfils the following desiderata:

- **Concept update**: the subspace moves from Z_i to $Z_i^{\text{new}} = \mathrm{QR}(AZ_i)$, allowing the model to refine concepts across layers.
- **Non-linear expressivity**: the fibre-wise MLP introduces a non-linearity acting on instance coordinates, matching the role of the FFN in standard Transformers.
- **Geometric consistency**: $v_i^{\text{new}} \in Z_i^{\text{new}}$ by construction; the layer-level projection ensures the residual also respects the bundle constraint.
- **Information selection**: the projection in the residual path acts as a geometric gate, retaining only the components of the previous instance that are compatible with the refined concept.

Practical implementation. A full $n \times n$ matrix A introduces $O(n^2)$ parameters and costs $O(n^2k)$ per token. To keep the model lightweight, we factorise $A = I_n + UV^\top$ with $U, V \in \mathbb{R}^{n\times r_A}$ and $r_A \ll n$, reducing the parameter count to $O(nr_A)$ and the per-token cost of applying A to Z_i to $O(nr_Ak)$. The identity term ensures that A starts close to the identity at initialisation, providing stable early training.

3.5 Layer normalization

We adopt the pre-LN convention: before each sub-layer (attention and feed-forward) we normalize the instance vector v_i. To keep v_i inside its subspace Z_i, we perform the normalization inside the fibre subspace. Since $v_i = Z_i\alpha_i$ with $\alpha_i \in \mathbb{R}^k$ (the coordinates of v_i in the orthonormal basis given by the columns of Z_i), we apply standard layer normalization to α_i and then reconstruct:

$$\alpha_i \leftarrow \mathrm{LayerNorm}(\alpha_i), \qquad v_i \leftarrow Z_i\alpha_i.$$

The operation is applied per token independently, uses learnable parameters that are shared across tokens (but differ per layer), and adds negligible computational cost. This fiber-wise normalization ensures that the instance vectors remain in their respective subspaces while benefiting from the training stability of layer normalization.

4 Complexity and design summary

The chordal distance computation adds a constant factor $O(nk^2)$ per pair during the attention process, but the overall asymptotic complexity remains $O(L^2)$, where L is the number of tokens (same as in standard Transformers). Hence, the need for efficient attention is not specific to our model; it is a well-known challenge for long sequences. Among all the solutions introduced to address this issue for classical Transformers, we are particularly interested in adapting Reformer hashing: our geometric structure built on compact Grassmannians is suited to make locality-sensitive hashing particularly stable and effective.

4.1 Reformer-style hashing

Locality-sensitive hashing (LSH) groups tokens whose representations are similar, and attention is computed only within each bucket. In our framework, we can define hash functions that are sensitive to the chordal distance between subspaces on the Grassmannian, reducing complexity while preserving the most relevant interactions. This approach is particularly natural with our geometric structure, as the concept similarity s_{ij}^{concept} already measures proximity.

The compactness of the Grassmannian is here a major advantage, in contrast to hashing raw token vectors in $\mathbb{R}^d$, which may suffer from unbounded norms and scale sensitivity. Indeed, all subspaces lie within a bounded set of diameter $\sqrt{k}$. This boundedness makes locality-sensitive hashing particularly stable, producing well-separated buckets without pathological edge cases. Moreover, the uniform distribution of subspaces on the compact manifold (under the Haar measure) facilitates even partitioning, which helps control bucket sizes and reduces the risk of worst-case collisions.

4.2 Design summary and trade-offs

The Bundle-Transformer is designed with computational feasibility as a primary concern. All operations (QR decomposition, low-rank factorisation of the shared FFN matrix, chordal distance computation) are efficient and differentiable. The shared, low-rank FFN keeps the per-token cost modest, while the attention mechanism can be made scalable via locality-sensitive hashing, as in classical Transformers. These choices give us a baseline that is both mathematically meaningful and practically implementable.

Of course, expressivity may be increased at the cost of more computation: for instance, using token-specific FFN matrices, adding concept-dynamic attention, including more sophisticated geodesics or simply by increasing the ambient dimension n or the concept dimension k. The optimal configuration will ultimately depend on the task, the dataset and will be determined empirically. The hashing strategy, in particular, is expected to play a crucial role in scaling

to long sequences without sacrificing the geometric benefits. We view the current design as a starting point for exploration, and we invite collaboration to refine these choices.

5 Conclusion and invitation

Bundle-Transformer aims to build a fundamental rethinking of token representations in the Transformer architecture. By embedding tokens into the tautological bundle we aim at implementing in an effective way deep geometric properties, in order to factorize the invariant concept and the contextual instance. Following this philosophy, geometric bias is achieved by constraining all operations to be bundle morphisms. This geometric prior promises improved interpretability, stability, and possibly data efficiency.

In order to improve expressivity, our architecture can naturally be generalized to more general contexts (flag varieties, or considering Riemannian geodesics, for instance) but we restricted this note to a computationally-controlled setting. A full implementation requiring substantial engineering effort to fix definitive architectural choices and hyperparameters is needed. This position paper serves as an invitation to join us in exploring this promising direction. We welcome feedback, discussions, and potential collaborations.

References

[1] Kang, H. et al. GrFormer: A novel Transformer on Grassmann manifold for infrared and visible image fusion. Information Fusion. 125, 2025.

[2] Su, J. et al. RoFormer: Enhanced Transformer with Rotary Position Embedding, *arXiv:2104.09864*, 2021.

[3] Vaswani, A. et al. Attention is all you need, Advances in Neural Information Processing Systems 30, 2017.

[4] Zhang, C. Attention Is Not What You Need. *arXiv:2512.19428*, 2025.