Graph transformers

LECTURE 17

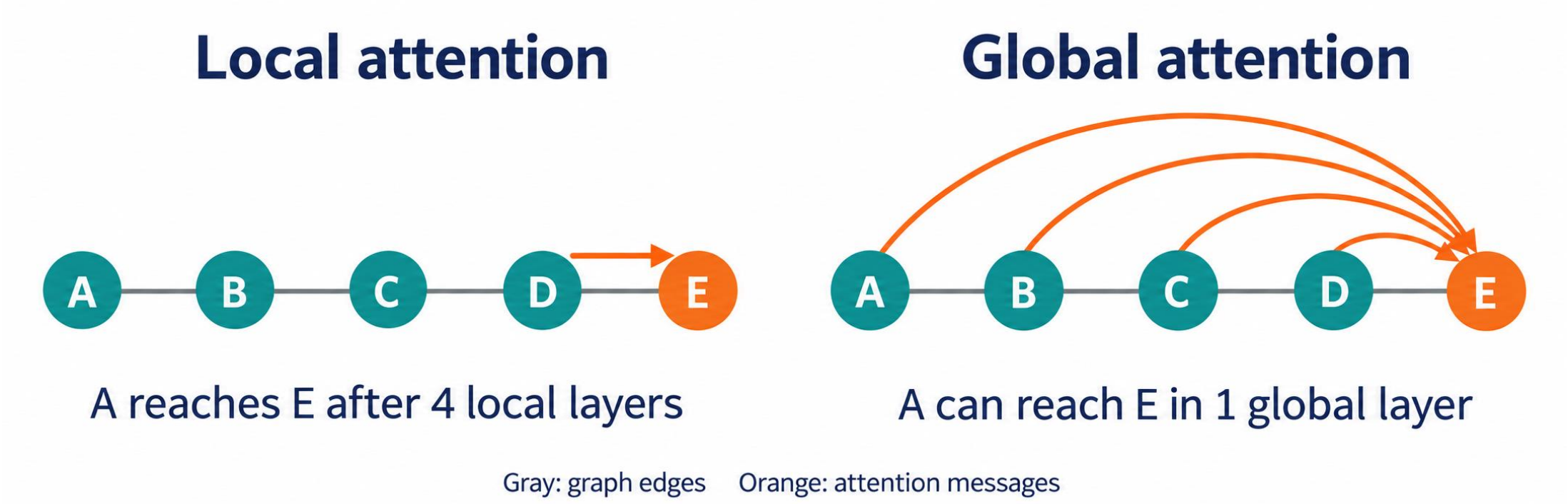
Full attention with graph structure

Local and global communication, structural encodings, and representative models

Attention combines node features with graph relationships

Dwivedi & Bresson (2020). A Generalization of Transformer Networks to Graphs.
Rampášek et al. (2022). Recipe for a General, Powerful, Scalable Graph Transformer.

Local and global communication



Global attention creates a communication route even when no original edge exists.

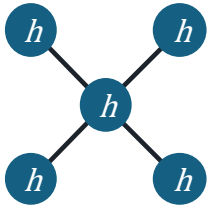
Graph transformers can use local, global, or hybrid attention

Dwivedi & Bresson (2020). A Generalization of Transformer Networks to Graphs.
Rampášek et al. (2022). Recipe for a General, Powerful, Scalable Graph Transformer.

Attention needs access to graph structure

A THOUGHT EXPERIMENT

Five-node path, 

Five-node star 

$$H_{path} = H_{star} \quad \Rightarrow \quad \text{Attention}_{\text{full}}(H_{path}) = \text{Attention}_{\text{full}}(H_{star})$$

If the model receives only H, it never sees the different edge sets.
Supply structural information through

- graph-dependent mask
- degree embeddings
- pairwise distances
- a local GNN

More communication alone does not tell the model which graph it received

Ying et al. (2021). Do Transformers Really Perform Bad for Graph Representation?
Kim et al. (2022). Pure Transformers are Powerful Graph Learners.

Ways to encode graph structure

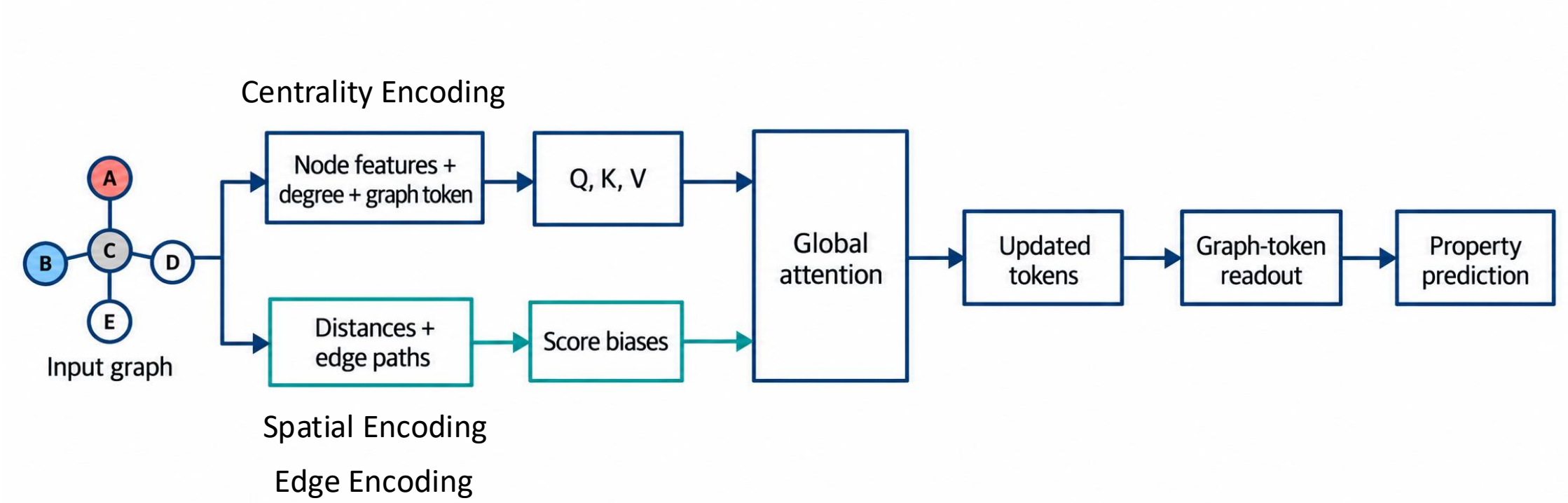
Where it enters	Example	What it tells the model
Node input	Degree embedding	How many neighbors a node has
Node input	Laplacian or random-walk encoding	Position or structural role
Attention score	Shortest-path distance, edge path	How a pair of nodes relates
Local branch	GNN neighborhood aggregation	Which local patterns surround a node

Node indices are arbitrary. Encodings must respect graph relabeling.

Laplacian eigenvectors need care with signs and repeated-eigenvalue bases

Kreuzer et al. (2021). Rethinking Graph Transformers with Spectral Attention.
Rampášek et al. (2022). Recipe for a General, Powerful, Scalable Graph Transformer.

Graphormer: structure in inputs and scores



Graphormer

Node input

$$h_i^{(0)} = x_i + z_{deg(i)}$$

A learned degree embedding adds the node's structural role – how important a node is

Attention score

$$e_{ij} = \frac{q_i^T k_j}{\sqrt{d_k}} + b_{dist(i,j)} + c_{ij}$$

Distance bias b and edge encoding c modify feature compatibility

Value aggregation

$$o_i = \sum_{j=1}^{n} \alpha_{ij} v_j, \qquad \alpha_{ij} = softmax_j(e_{ij})$$

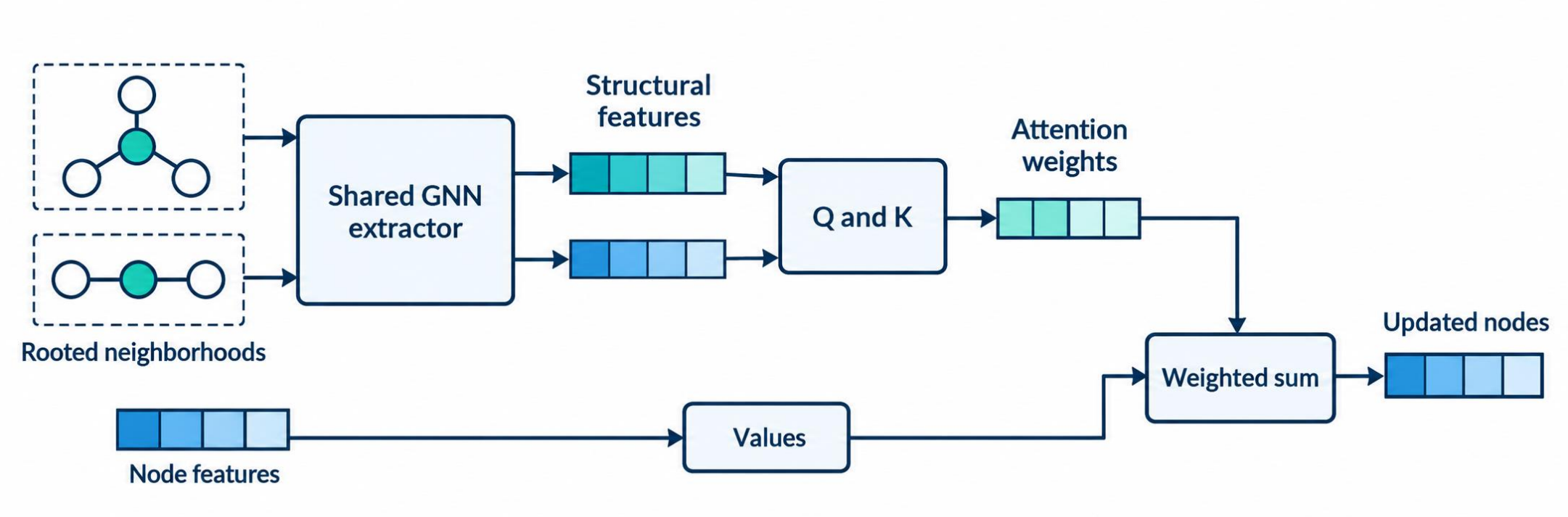
Queries, keys, and values are learned projections of the node features:

$$q_i = W_Q h_i, k_j = W_K h_j, v_j = W_V h_j.$$

Distance and edge information affect which node values receive weight.

Ying et al. (2021). Do Transformers Really Perform Bad for Graph Representation? Eqs. (5)–(7).

Structure-Aware Transformers (SAT) attention compares local structures



SAT enriches the comparison itself with subgraph information

SAT

Structural representation

$$s_i = \varphi(i, G; X)$$

The shared extractor summarizes the attributed neighborhood rooted at i.

Attention score

$$e_{ij} = \frac{(W_Q s_i)^T (W_K s_j)}{\sqrt{d_k}}$$

Queries and keys compare structural representations of two roots.

Value aggregation

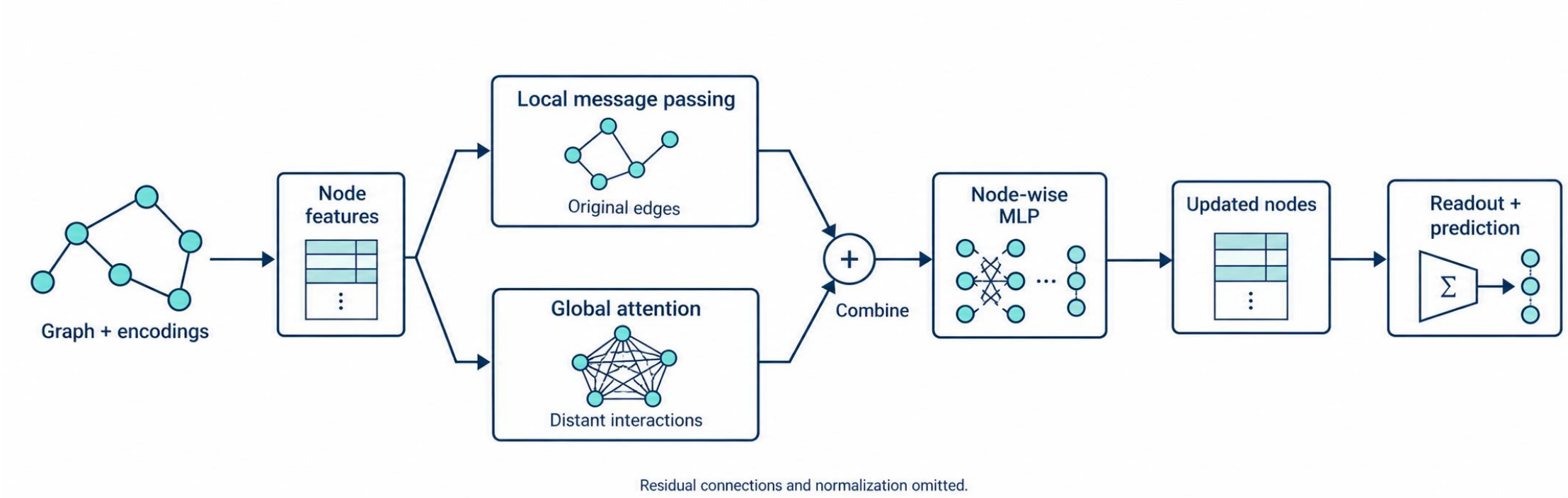
$$o_i = \sum_{j=1}^{n} \alpha_{ij} W_V x_j, \qquad \alpha_{ij} = softmax_j(e_{ij})$$

Node features provide values. Structural comparisons determine their weights.

The extractor changes the attention comparison before values are combined.

Chen, O'Bray & Borgwardt (2022). Structure-Aware Transformer for Graph Representation Learning. Eqs. (4)–(6).

GraphGPS: local and global branches



The local branch uses real edges. The global branch exchanges distant information.

GraphGPS is a modular recipe, with a choice of encodings and attention

GraphGPS

The paper's simplified layer update, with residuals, dropout, and normalization omitted

Local branch

$$X_M^{\ell+1}, E^{\ell+1} = MPNN_e^{\ell}(X^{\ell}, E^{\ell}, A)$$

A is adjacency. E contains edge features used by local message passing.

Global branch

$$X_T^{\ell+1} = GlobalAttn^{\ell}(X^{\ell})$$

The global branch starts from the same node features X at layer ℓ.

Combine branches

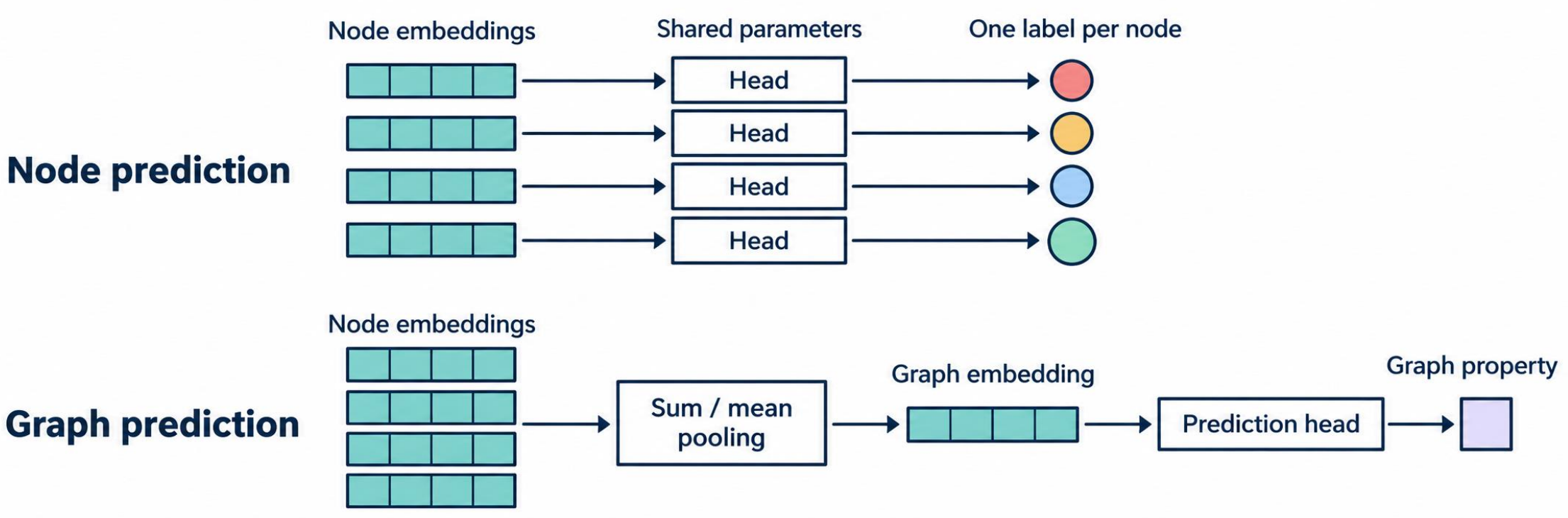
$$X^{\ell+1} = MLP^{\ell}(X_M^{\ell+1} + X_T^{\ell+1})$$

Add corresponding node representations, then apply a node-wise MLP.

Local and global branches operate in parallel within each GPS layer.

Rampášek et al. (2022). Recipe for a General, Powerful, Scalable Graph Transformer. Eqs. (2)–(4).

From node representations to predictions



SAT and GraphGPS: choose a node head or a graph readout to match the task.

Graphormer: the final graph token supplies the graph embedding instead of pooling.

Ying et al. (2021). Do Transformers Really Perform Bad for Graph Representation?
Chen, O'Bray & Borgwardt (2022). Structure-Aware Transformer for Graph Representation Learning.
Rampášek et al. (2022). Recipe for a General, Powerful, Scalable Graph Transformer.

Communication range and computational cost

Attention choice	Pair interactions	Main tradeoff
Local / sparse	$O(\|E\|d)$	Long paths need more layers
Dense global	$O(n^2 d)$	Direct long-range access, quadratic work
Efficient global	Method-dependent	Approximation or restricted interactions
Hybrid GPS	Local cost + global cost	Cost depends on the global module

These costs describe attention mixing, excluding feature projections.
A stored dense attention matrix uses $O(n^2)$ memory per head.

GPS is linear only with suitable efficient attention and compatible encodings

Rampášek et al. (2022). Recipe for a General, Powerful, Scalable Graph Transformer.

Main points and worked review

A GRAPH TRANSFORMER DESIGN MUST SPECIFY

The communication pattern defines which nodes exchange information.
Encodings, score biases, or local layers provide graph structure.
The interaction pattern determines the pairwise computational cost.

PATH AND STAR: SAME FEATURES, DIFFERENT STRUCTURE

With identical features and no graph-dependent inputs or operations,
global attention gives the same output for a path and a star.

Shared parameters support new graph sizes, but do not guarantee transfer accuracy

Ying et al. (2021). Do Transformers Really Perform Bad for Graph Representation?
Rampášek et al. (2022). Recipe for a General, Powerful, Scalable Graph Transformer.

Representative works for further reading

A Generalization of Transformer Networks to Graphs
Dwivedi & Bresson (2020). Local attention with Laplacian encoding

Do Transformers Really Perform Bad for Graph Representation?
Ying et al. (2021) · Graphormer. Degree, shortest-path, and edge-path encodings

Rethinking Graph Transformers with Spectral Attention
Kreuzer et al. (2021) · SAN. Learned spectral positional encoding

Structure-Aware Transformer for Graph Representation Learning
Chen, O'Bray & Borgwardt (2022) · SAT. Subgraph-aware attention comparisons

Recipe for a General, Powerful, Scalable Graph Transformer
Rampášek et al. (2022) · GraphGPS. Local message passing with global attention

Pure Transformers are Powerful Graph Learners
Kim et al. (2022) · TokenGT. Node and edge tokens with identifier embeddings

Graphormer illustrates structural biases. GraphGPS combines local and global computation.