DeepCrossAttention explained: replacing fixed residual connections with learned depth-wise routing

24 November 2026 · 8 min read · blogs.iictp.org/deepcrossattention-transformer-residual-connections-explained

Short answer: DeepCrossAttention (DCA) is a transformer architecture from Google Research that replaces the plain “add the input back” residual connection with learned weights and cross-attention across earlier layers, so each layer can choose which earlier representations to reuse. Google Research described it in a blog post on 23 November 2026, based on a paper presented at ICML 2025. It matters technically because the residual connection is in almost every large language model, and DCA changes it while adding very few parameters.

TL;DR The short version

- DeepCrossAttention replaces the fixed residual connection in transformers with learned weights and attention across earlier layers.
- Google Research says it reaches the same quality up to three times faster with very few extra parameters, based on an ICML 2025 paper.
- Results cover language-modelling perplexity and image classification, so compare against a fixed-budget baseline before adopting it.

Read online, listen to it, or change the text size: https://blogs.iictp.org/deepcrossattention-transformer-residual-connections-explained

What is DeepCrossAttention?

A residual (skip) connection adds a layer’s input to its output, which keeps the learning signal alive through dozens of layers. According to the Google Research post by Kyriakos Axiotis and Thomas Fu, that plain sum treats every earlier representation with equal weight, so detail from early layers can be averaged out in deep networks. DCA generalises the connection so a layer can pick and combine features from the history of earlier layers instead of receiving one blended sum.

How it works

- **Dynamic layer routing.** For each token, DCA computes mixing weights from a learnable static base plus an input-dependent projection. The post says this lets the network move between a purely sequential and a more parallel structure depending on the token.
- **Depth-wise cross-attention.** Normal attention runs across tokens in a sequence. DCA applies attention across layers: a token queries a weighted mix of earlier-layer embeddings, using a split into query, key and value inputs.

- **Retrofit-friendly start.** The post says DCA can be initialised to behave like a standard residual network, which is why it can be added to a model that is already pre-trained.

What is new compared with earlier approaches

The post places DCA in a wider family of “dynamic residual” designs. It names attention residuals from Moonshot AI, which use a learnable softmax attention with a layer-specific pseudo-query and a block variant to save memory, and hyper-connections. It also says the Inkling 2 model from Thinking Machines adopted attention residuals together with DCA’s query/key/value split. The shared idea is to replace a fixed sum with selective retrieval from earlier depth.

Results reported by the source

- The summary says DCA reaches the same quality “up to three times faster” with a negligible increase in parameters; later the post says it matched standard transformers in up to five times fewer training steps on the C4 and LM1B language-modelling datasets. These are Google’s figures from its own experiments.
- On LM1B, to reach a perplexity below 14 a standard transformer of fixed width needed 42 layers, while DCA reached lower perplexity with 30 layers, a 29% reduction in depth and parameter count, according to the post.
- DCA adds less than 0.2% to total model size, the post says, though it raises memory bandwidth because representations from different layers are presented intact.
- Retrofitting DCA onto an already pre-trained transformer and continuing training briefly lowered LM1B perplexity by 0.19, against 0.02 for continued standard training.

Limitations and open questions

- The blog describes a 2025 conference paper, so the peer-reviewed claim sits with the paper, not the blog; read the paper for the exact setups.
- The results named in the post are language-modelling perplexity and image classification. Whether the gains hold at frontier scale on downstream tasks is not shown in the post.
- Extra memory bandwidth is a real cost on accelerators where bandwidth, not arithmetic, is the bottleneck.

How to evaluate it for your own models

1. Pick a small baseline transformer and train it with a fixed token budget so you have a reference perplexity.
2. Swap the residual connections for a depth-wise routing module with the same width, and match the parameter count to within a fraction of a percent.
3. Compare perplexity at equal steps and at equal wall-clock time, because DCA’s extra bandwidth can change the second comparison.
4. Only then test a retrofit on a pre-trained checkpoint, and check downstream tasks, not perplexity alone.

What it means for engineers and researchers

For practitioners the point is architectural literacy. Residual connections look like settled plumbing, yet at least three groups are changing them. If you train or fine-tune models, expect checkpoints with non-standard residual paths, which affects conversion tools, quantisation and inference engines that assume the classic structure. See how to

evaluate an LLM application for testing a model on your own task.

Frequently asked questions

- **What is DeepCrossAttention?** A transformer design from Google Research that lets each layer weight and combine outputs of earlier layers with depth-wise cross-attention, instead of using a fixed residual sum.
- **Who published it and when?** Google Research described it on 23 November 2026; the underlying paper was presented at ICML 2025.
- **Does it need a model to be trained from scratch?** Not necessarily. The post says it can start out equivalent to a standard residual network, and reports a retrofit experiment on a pre-trained model.
- **How much does it add to a model?** Less than 0.2% of total size, according to Google Research.

Related reading

- Language models that rewrite their own weights
- Inkling 2 Mini: a looped transformer for local AI
- What is an LLM context window?
- AI Practitioner

Sources and currency

Sources were read on 24 November 2026. Products, rules and standards change, so check the linked originals before you rely on any detail.

> Related credential: AI Practitioner. Explore the exam facts and syllabus.

Sources and further reading

1. Kyriakos Axiotis and Thomas Fu, Google Research. Supercharging transformer residual connections. 23 November 2026. research.google/blog/supercharging-transformer-residual-connections.

Published by IICTP, an independent technical credential body, on 24 November 2026. The current version of this post is at https://blogs.iictp.org/deepcrossattention-transformer-residual-connections-explained. More guides: blogs.iictp.org.