VALUE-BYPASS ATTENTION
A SIMPLE VALUE-PATH VARIANT FOR BERT ON SQUAD V1

Shin D.B.
Simple Report on a Small Experiment

ABSTRACT

We investigated a simple modification of the BERT self-attention mechanism in which the value vectors are augmented with a shortcut from the input representation. Experiments on SQuAD v1 show that smaller and medium-sized BERT variants achieve higher mean F1 scores, while larger models exhibit negligible differences. In all cases, the variance between random seeds increases. These observations suggest that injecting a value-path bypass provides additional expressivity but reduces stability.

1 INTRODUCTION

The self-attention mechanism has been extensively studied, with many variants proposed to improve efficiency, incorporate positional information, or stabilize training. Almost all of these approaches focus on modifying queries and keys, since the attention weights depend directly on their dot product. In contrast, the value pathway, which carries the token information once the weights are determined, has remained largely unchanged. This report explores a minimal modification in the value path. Rather than altering the attention distribution, we add a positive shortcut to the input representation. The goal was to examine whether this change could improve information flow in smaller models and how it affects downstream performance.

2 METHOD

The proposed variant, denoted V', modifies the value vectors as follows:

$$V' = V + \alpha \cdot \mathrm{ReLU}(X)$$

where X is the input hidden state. The coefficient α was set arbitrarily at 0.5. Although there is a code for a learnable α, experiments were not conducted due to limited resources.

2.1 EXPERIMENT SETUP

2.1.1 PRETRAINING

- Corpus: English Wikipedia (0.6B tokens).
- Tokenizer: HuggingFace ‘bert-base-uncased‘.
- Objective: Masked Language Modeling
- Schedule:
- Warm stage: seq length 128, batch 48, 65k steps
- Long stage: seq length 512, batch 48, 15k steps
- Optimizer: AdamW with betas (0.9, 0.999), eps=1e-8, weight decay 0.01
- Scheduler: linear warmup and decay.

2.1.2 SQUAD V1

- Dataset: SQuAD v1, evaluation on dev set.
- Epochs: 3 (4.1k steps).

- Batch size: 64
- Learning rate: 5e-5
- Optimizer: Trainer default
- Evaluation: Trainer default
- Checkpoints: last step, no early stopping.
- Seeds:
- 4L-256: 5 runs (42, 2, 8, 9, 5)
- 6L-384 and 8L-512: 3 runs each (42, 2, 8)

Larger models have less runs due to resources being depleted.

3 RESULTS

The results are as follows

Table 1: Final F1 mean ± std

MODEL	BASE F1	OUR F1	Δ MEAN	Δ STD
4L-256-4d	55.51 ± 0.32	59.62 ± 3.32	+4.1	×10
6L-384-1d	63.15 ± 0.07	64.74 ± 1.47	+1.6	×20
8L-512-4d	73.37 ± 0.12	73.26 ± 0.50	≈0	×4

4 DISCUSSION

The value bypass can be interpreted as a residual shortcut of positive activations. For smaller models, this addition preserves more of the raw input information, improving expressivity and raising average scores. However, because the term is ungated, it bypasses attention selectivity, making training trajectories highly seed-sensitive. From a variance perspective:

$$Var[V + \alpha \cdot \mathrm{ReLU}(x)] = \mathrm{Var}[V] + \alpha^2 \mathrm{Var}[\mathrm{ReLU}(x)] + 2\,\mathrm{Cov}(V, \mathrm{ReLU}(x))$$

which is always $\geq$ Var[V]. Thus the output variance cannot decrease, consistent with observed instability.
Large models, already sufficient in capacity, show little gain. The bypass introduces noise without benefit, hence variance increases but mean unchanged.

4.1 LIMITATIONS

The pretraining corpus used in this work was considerably smaller in both size and scope (0.6B tokens) compared to BERT-base (3.3B). Evaluation was restricted to a single downstream task, SQuAD v1, and the number of random seeds was limited to five for the 4-layer model and three for the 6- and 8-layer models. The hyperparameter α was heuristically fixed at 0.5 without tuning. Checkpoints were taken only at the final training step, and several checkpoints for the 6- and 8-layer models could not be retained due to resource limitations. Finally, the implementation relied on a custom skeleton with weight loading; while the bypass path was not explicitly unit-tested, the observed results suggest that it was functioning as intended.

4.2 FURTHER RESEARCH

4.2.1 ALPHA

Alpha is fixed as of now, mostly due to it having no impact once is set to be trainable from the start. Therefore, different setups could be tested: per-head, per-layer, global, warm-up/decay, etc.

4.2.2 ACTIVATION

Only ReLU was tested. More variations such as GeLU, SiLU, etc. could be tested.

4.2.3 NORMALIZATION & INSERTION

Currently, only raw X is used. LN(X) could be substituted. Also, currently all layers have this bypass. Limiting the number of layers which has this variant might aid in reducing variance.

4.2.4 BENCHMARKS

Diverse downstream tasks should be tested. SQuAD v2 and GLUE tasks are coded, but could not be conducted. Other standard tasks, along with base model benchmarks should be provided.

4.2.5 SCALING

Larger dataset and models should be tested. Current benchmarks are not comparable to original paper's scores. Using a larger dataset might help to mitigate that issue.
Also, deeper layers with larger hidden dimensions should be tested, to see if this bypass loses effects.

5 CONCLUSION

We found that a simple value-path bypass can boost mean F1 in smaller BERT models, at the cost of increased variance. Larger models show negligible benefit. This trade-off implies that value-path modifications deserve further attention, especially when combined with gating or normalization mechanisms to stabilize training.

6 DISCLAIMER

Experiments were run on a single RTX 5090 GPU with PyTorch 2.8 and HuggingFace datasets 3.6.0. This report was drafted with the assistance of GPT for clarity.