

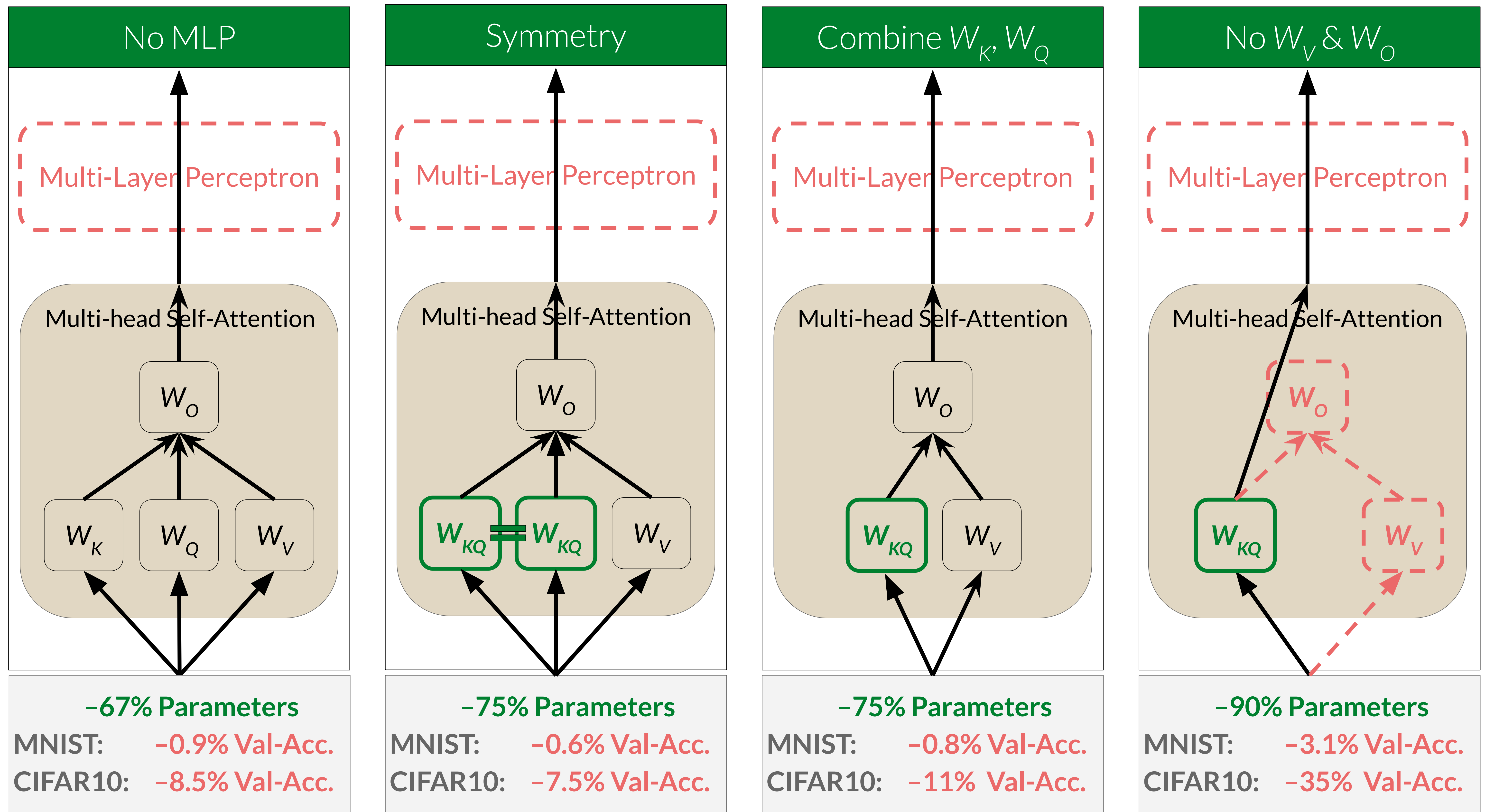
Reducing the Transformer Architecture to a Minimum

Bernhard Bermeitinger, Tomas Hrycej, Massimo Pavone, Julianus Kath, Siegfried Handschuh

University of St.Gallen (HSG), Switzerland

bernhard.bermeitinger@unisg.ch, tomas.hrycej@unisg.ch,

massimo.pavone@student.unisg.ch, julianus.kath@students.unisg.ch, siegfried.handschuh@unisg.ch



Methodology and Conclusion

MOTIVATION, CONCEPT & APPROACH

- Models chosen to be slightly over-determined:
 - $Q = (KM)/P$ ($K=\#TrainingSize$, $M=\#Classes$, $P=\#Parameters$)
 - Target models: $1 < Q < 5$
- No MLP**
 - Contributes most parameters (>60%)
 - Overwhelms importance of *attention*
 - Not necessary for *nonlinearity*
- Symmetry**
 - For tokens x, y : $x \sim y$ and $y \sim x$
 - Use same matrix (shared parameters) for W_K and W_Q
- Combine W_K and W_Q**
 - Two square matrices can be collapsed into one
- Drop W_V and W_O**
 - Without MLP, the matrices W_O and W_V collapse into one
 - All dimensions are equal, so this projection is not needed

IMPLEMENTATION & EXPERIMENTAL SETUP

- Adam with default parameters
- Seeds all fixed

LIMITATIONS

- Focus on Computer Vision: NLP difficult
- Only MNIST/CIFAR-10
- Models fairly small (60k-200k parameters)
- Most reductions work only in **single-head** configuration

FUTURE WORK

- Additional datasets
- Further decrease model parameters
- Use **Conjugate Gradient** as more efficient optimizer
- Extend idea to LLMs: GPT-like architectures