Aalto-yliopisto
Perustieteiden
korkeakoulu

Kahden optimointimenetelmän vertailu videoperusmallin koulutuksessa ja siirto-oppimisessa (valmiin työn esittely)

Johan Roos

06.11.2026

Ohjaaja: *Ahti Salo*
Valvoja: *Ahti Salo*

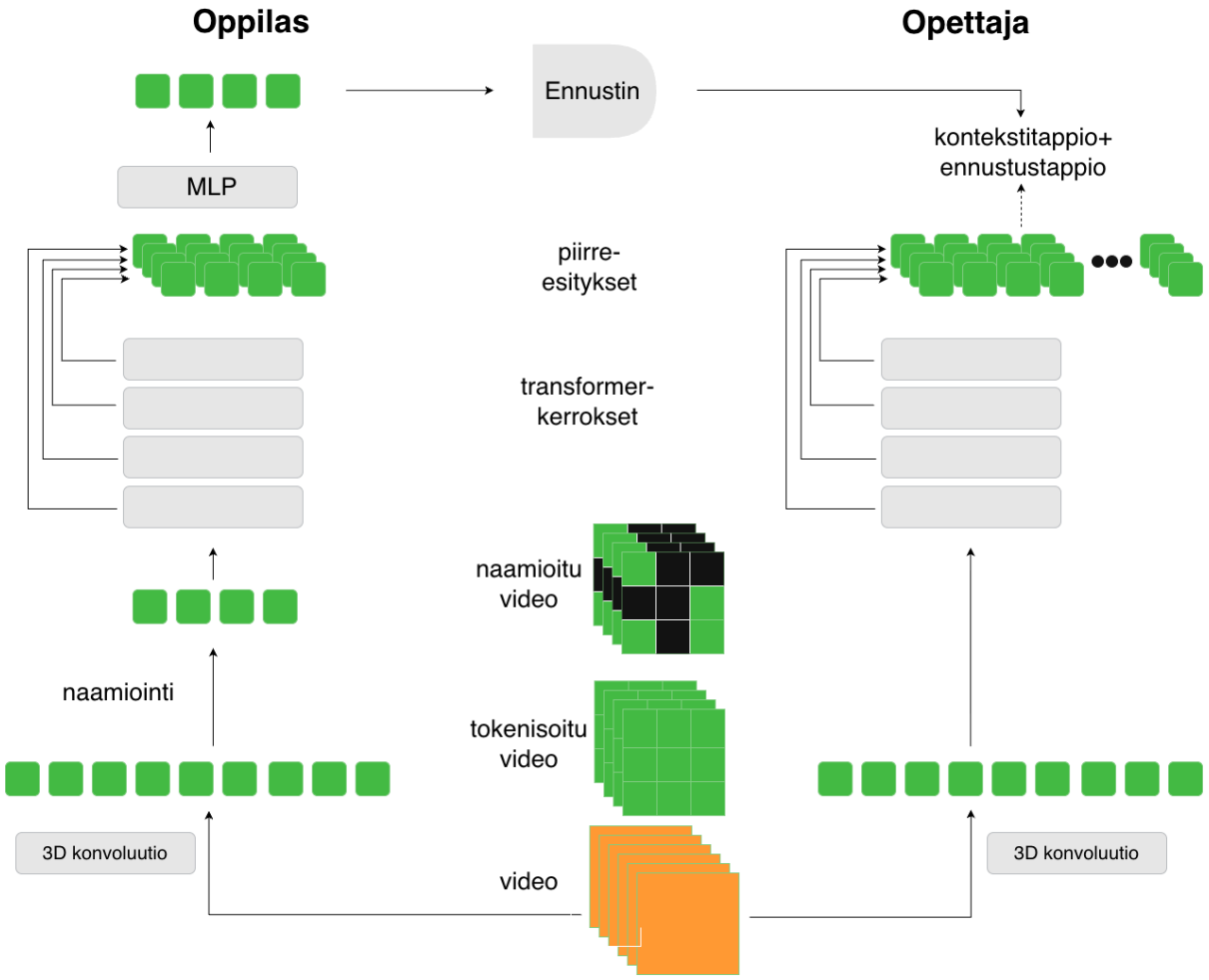
Työn saa tallentaa ja julkistaa Aalto-yliopiston avoimilla verkkosivuilla. Muilta osin kaikki oikeudet pidätetään.

Aalto-yliopisto
Perustieteiden
korkeakoulu

Taustaa

- Itseohjatun oppimisen tavoitteena on oppia raakadatasta
- Siirto-oppimisessa perusmallin oppimia piirteitä hyödynnetään eri tehtävässä kuin mihin perusmalli on koulutettu
- Perusmallien koulutus vaatii paljon laskentaa
- Tässä työssä vertaillaan AdamW- ja Muon-optimointimenetelmää videoperusmallin kouluttamisessa
  - AdamW on todettu yleisesti toimivaksi, joten sitä käytetään lähtökohtaisesti kun uusia kuva- ja videomalleja kehitetään
  - Muonilla on saatu hyviä tuloksia suurten kielimallien koulutuksessa, mutta sen toimivuutta videoista oppimisessa on tutkittu kirjallisuudessa vähemmän

Perusmallin koulutusvaihe



Optimointimenetelmät

1. AdamW on adaptiivinen optimointimenetelmä
   - Lasketaan gradientin avulla sopivan kokoiset päivitykset jokaiselle parametrille erikseen
   - Nimessä W (weight decay) viittaa siihen, että parametreja vaimennetaan kohti nollaa tehden optimoitavasta funktiosta paremmin käyttäytyvän
2. Muon on ortogonalisoiva optimointimenetelmä
   - Jokaisessa neuroverkon kerroksessa Muon ortogonalisoi parametrien päivitysmatriisin
   - Gradientin singulaariarvot asetetaan ykkösiksi, jolloin päivitys keskittyy tasaisemmin useaan pääakselisuuntaan
     $W \leftarrow W - \eta UV^{\mathrm{T}}$, kun $G = U\Sigma V^{\mathrm{T}}$

Siirto-oppimisvaihe

- Mallin tuottamien piirteiden hyödyntäminen on mahdollista kahdella tapaa
  1. Jäädytetään mallin parametrit ja koulutetaan sen päälle kevyt luokitin
  2. Mallin parametrejä hienosäädetään
- Tässä työssä käytetään lineaarista luokitinta
  - Perusmalli tuottaa N tokenia, jotka sijaitsevat d-ulotteisessa avaruudessa
  - Lineaariselle luokittimelle annetaan syötteenä vektori $v \in \mathbb{R}^d$, joka on N tokenin keskiarvo
    - Luokitin antaa todennäköisyyden kullekin tehtävässä määritellylle luokalle

Aineistot



- Something-Something v2 (174 luokkaa)
  - Yli 200 tuhatta videopätkää
  - Sisältää ajallista hahmottamista parantavia piirteitä
- ImageNet-1k (1000 luokkaa)
  - Yli miljoona kuvaa
  - Sisältää laajasti erilaisia visuaalisia piirteitä
- Diving48 (48 luokkaa)
  - Yli 15 tuhatta videopätkää
  - Tausta on samankaltainen jokaisessa videossa, joten liike on keskeisemmässä roolissa luokittelutehtävässä

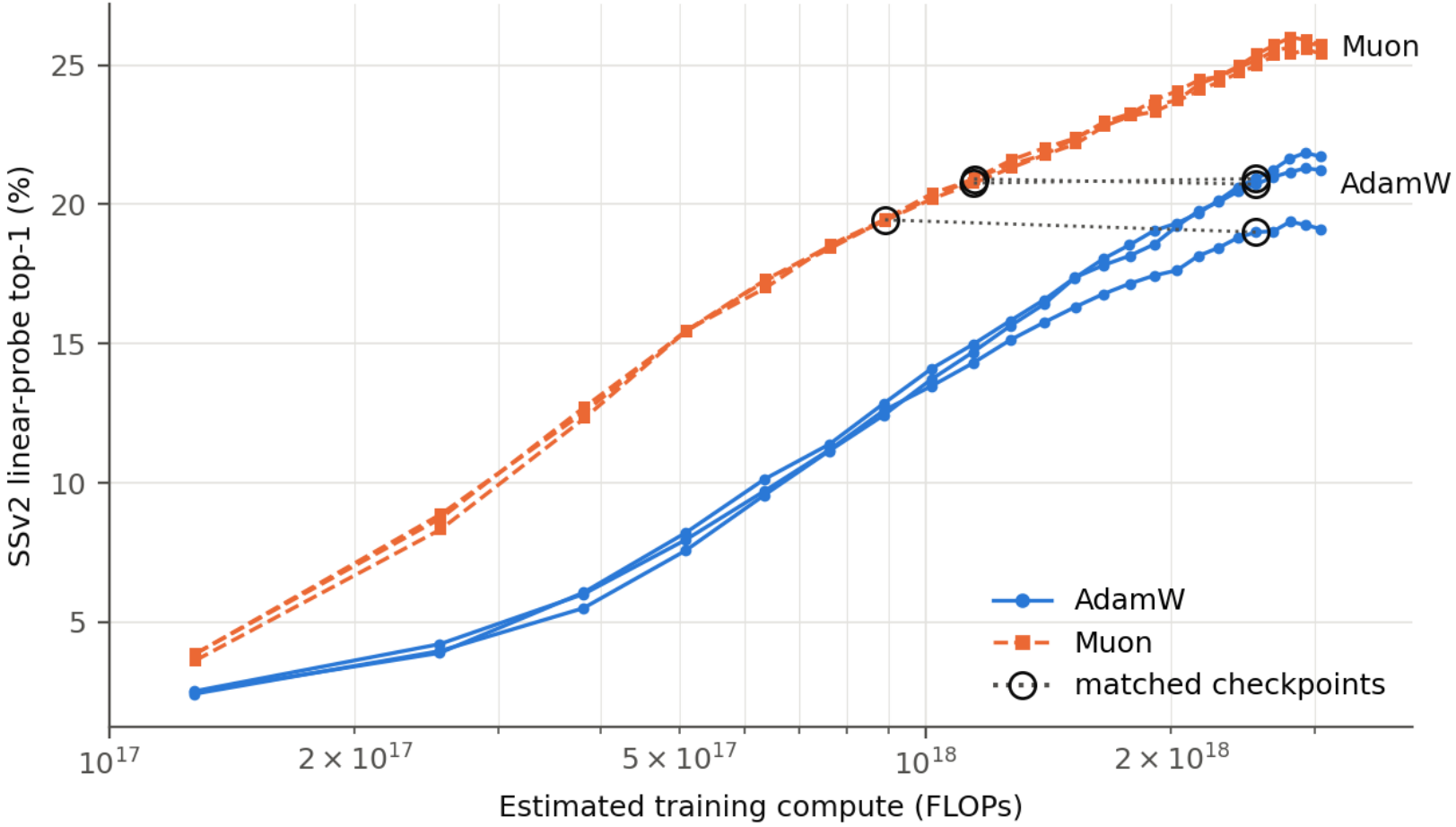




Menetelmät

- Muonin ja AdamW:n kokeet tehtiin pareittain kolmella satunnaissiemenellä ja 24 jakson koulutuksella
  - Koulutuksessa käytettiin SSv2 videoita ja ImageNet kuvia
- Mallien parametrit tallennettiin jokaisen jakson jälkeen
  - Laskenta mitattiin liukulukuoperaatioina (FLOPs)
  - Centered Kernel Alignmentilla (CKA) piirteiden vertailu
  - Suorituskykyä arvioitiin lineaarisen luokittimen avulla
- Videoista poimittiin mallin syötteeksi 8 kuvaa sekunnissa
  - Alkuperäinen resoluutio muunnettiin 224x224 pikseliksi

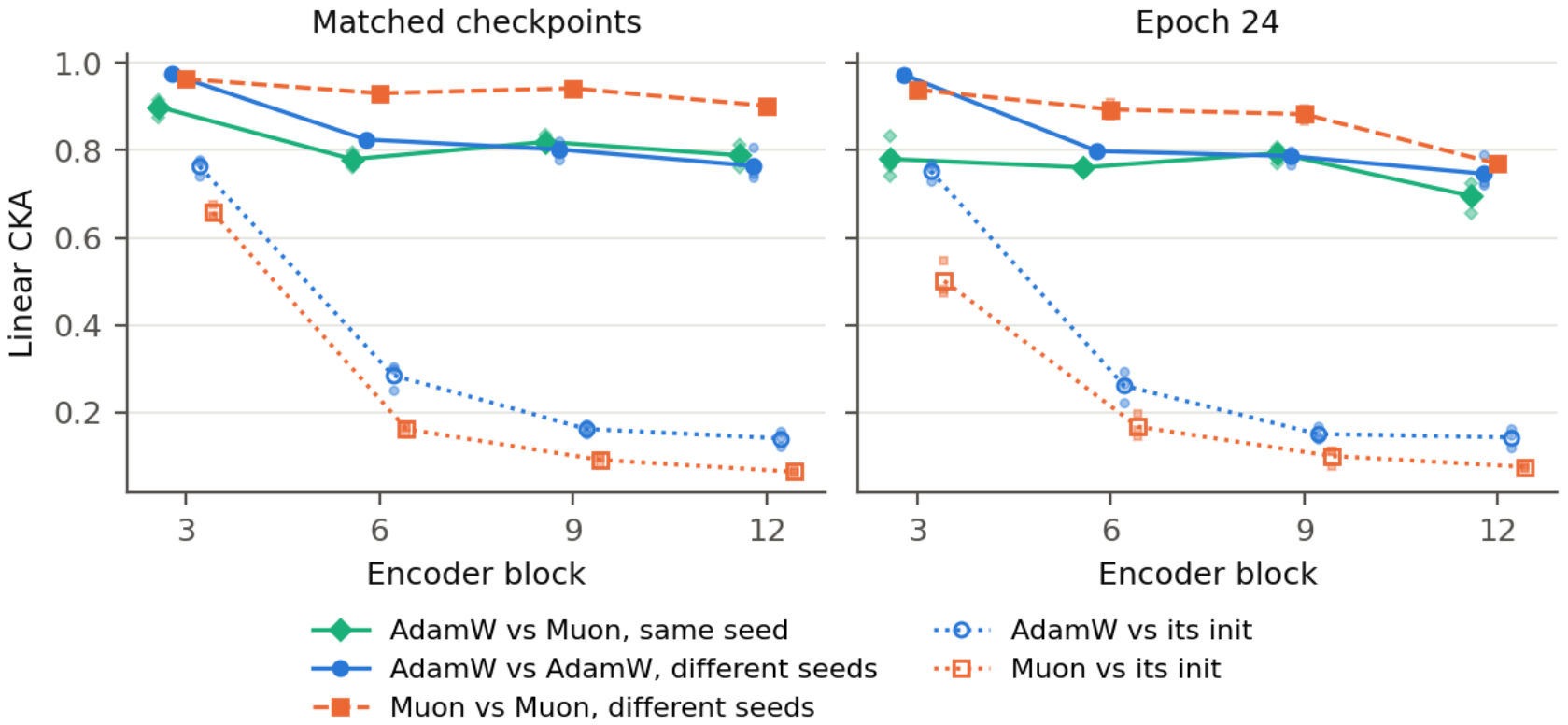
Laskentatehokkuus



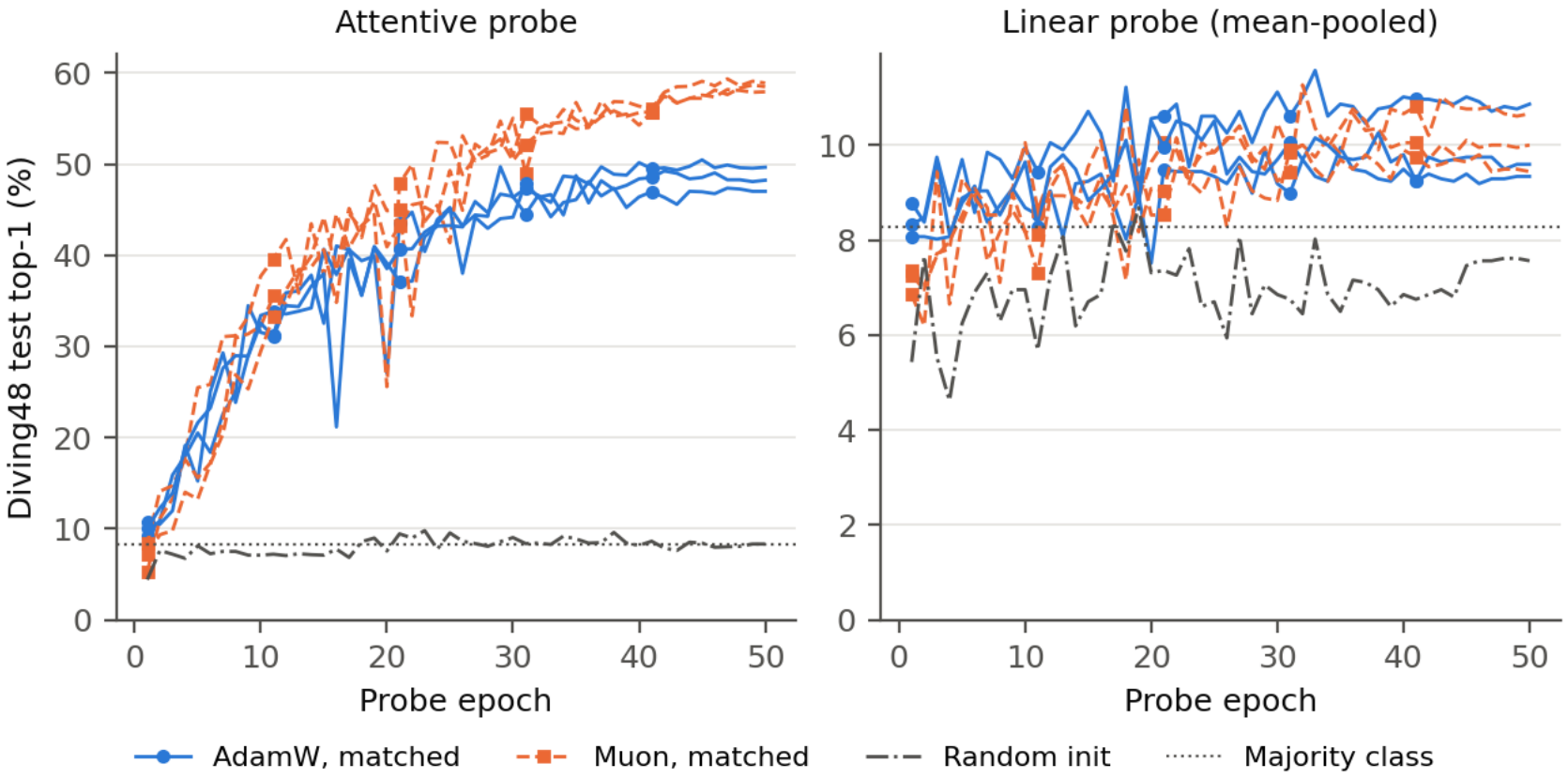
Piirteiden yhtenevyydet

$$\mathrm{CKA}(X, Y) = \frac{\|Y^{\mathrm{T}}X\|_F^2}{\|X^{\mathrm{T}}X\|_F \|Y^{\mathrm{T}}Y\|_F}$$

$X, Y \in \mathbb{R}^{n \times d}$: n viittaa aineiston kokoon ja d piirteiden ulottuvuuteen



Siirrettävyys toisiin tehtäviin



Tutkimuksen rajoitteet

- Yleistä skaalautuvuutta esimerkiksi mallin parametrien lukumäärän tai aineiston koon suhteen ei tutkittu
- Ei optimointimenetelmien hyperparametrien hienosäätöä
  - Tulokset ovat vain suuntaa antavia, koska hienosäädöllä voidaan vaikuttaa oppimisen tehokkuuteen
- Suorituskyvyn mittaaminen lineaarisella luokittimella ei paljasta kaikkea opittujen piirteiden laadusta ja käytettävyydestä
  - Lineaarinen luokitin saa syötteeksi piirteiden keskiarvon, jolloin paikallisesti ja ajallisesti tärkeätä informaatiota katoaa

Johtopäätökset

- Muon toimii 25 miljoonan parametrin videomallien optimointiin
  - Muonilla SSv2-aineiston luokitustarkkuus nousi samaan kuin Adamilla keskimäärin noin 2,4 kertaa pienemmällä laskennalla
- Muonilla opitut piirteet olivat vähemmän riippuvaisia käytetystä satunnaissiemenestä kuin AdamW:llä opitut piirteet
- Muonilla ja Adamilla optimoitujen mallien piirteiden siirrettävyydessä ei ollut eroa, kun tehtävänä oli Diving48 aineiston luokittaminen lineaarisella luokittimella
  - Monimutkaisempaan neuroverkkoon perustuvalla luokittimella saatiin siirto-oppimisessa parempia tuloksia Muonilla kuin AdamW:llä

Lähteet

- Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, et al. “V-JEPA 2.1: Unlocking Dense Features in Video Self-Supervised Learning.” *European Conference on Computer Vision, 2026*. https://doi.org/10.1007/978-3-032-37369-4_36.
- Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, et al. “The ‘Something Something’ Video Database for Learning and Evaluating Visual Common Sense.” *IEEE International Conference on Computer Vision*, 2017. https://doi.org/10.1109/ICCV.2017.622.
- Olga Russakovsky, Jia Deng, Hao Su, et al. “ImageNet Large Scale Visual Recognition Challenge.” *International Journal of Computer Vision,* 2015. https://doi.org/10.1007/s11263-015-0816-y.
- Yingwei Li, Yi Li, and Nuno Vasconcelos. “RESOUND: Towards Action Recognition Without Representation Bias.” European Conference on Computer Vision*, 2018*. https://doi.org/10.1007/978-3-030-01231-1_32.
- Jingyuan Liu, Jianlin Su, Xingcheng Yao, et al. “Muon Is Scalable for LLM Training.” arXiv:2502.16982. Preprint, arXiv, 2025. https://doi.org/10.48550/arXiv.2502.16982.
- Keller Jordan, Yuchen Jin, Vlado Boza, et al. “Muon: An Optimizer for Hidden Layers in Neural Networks.” *Keller Jordan’s Blog*, December 8, 2024. https://kellerjordan.github.io/posts/muon/.
- Ilya Loshchilov, and Frank Hutter. “Decoupled Weight Decay Regularization.” *International Conference on Learning Representations*, 2019. https://openreview.net/forum?id=Bkg6RiCqY7.
- Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. “Similarity of Neural Network Representations Revisited.” *Proceedings of the 36th International Conference on Machine Learning*, 2019. https://proceedings.mlr.press/v97/kornblith19a.html.