Problem podwójnego spadku (*Double descent*)

Stanisław Kaźmierczak

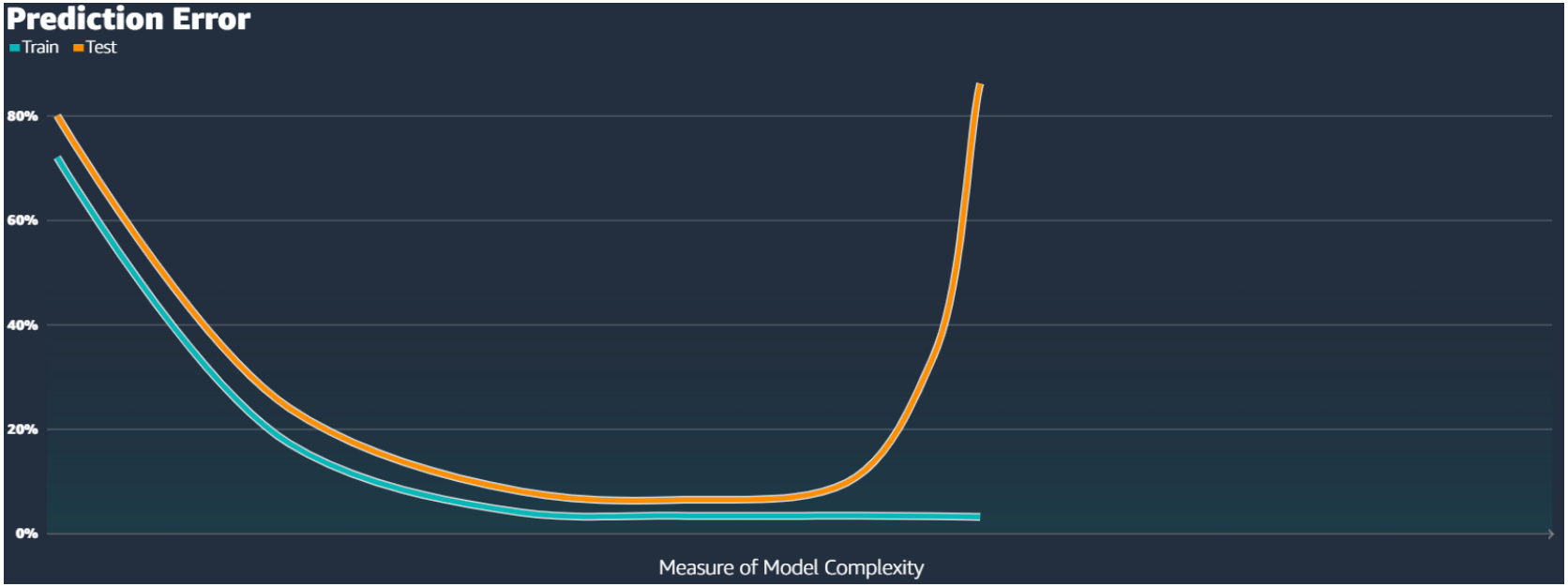
Agenda

1. Podwójny spadek
2. Efektywna złożoność modelu
3. Model-wise double descent
4. Epoch-wise double descent
5. Niemonotoniczność względem liczby obserwacji
6. Podsumowanie

Wprowadzenie

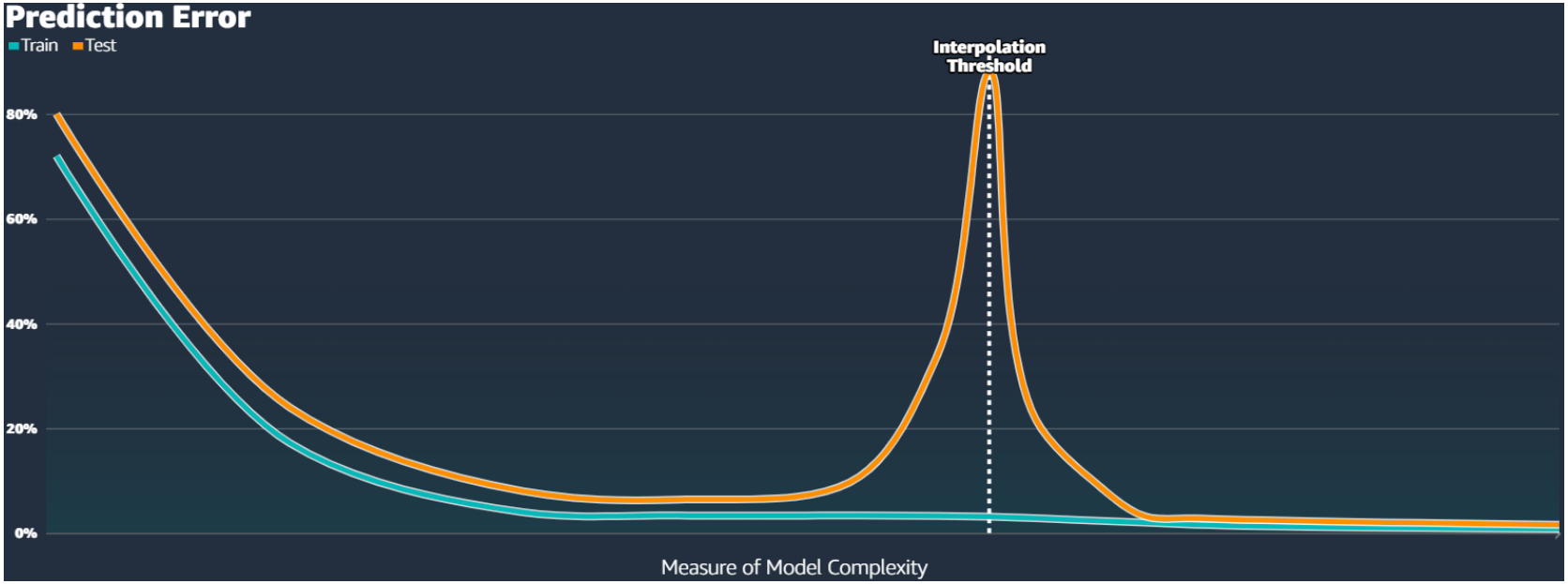
- Zjawisko ripostujące klasyczne rozumienie *bias-variance tradeoff*.
  - Standardowo oczekujemy, że najlepszy model będzie stanowił pewien kompromis pomiędzy biasem (niedouczenie) oraz wariancją (przeuczenie).
- Obserwowana jest sytuacja, w której złożone, (bardzo) przeuczone modele osiągają wysoką jakość na zbiorze testowym.
- W konsekwencji wielu badaczy jak i praktyków kwestionuje znaczenie tradycyjnego rozumienia kompromisu między biasem i wariancją.

Klasyczny *bias-variance tradeoff*



[1]

Podwójny spadek



[1]

- *Classical regime* – na lewo od progu interpolacji
- *Interpolation regime* – na prawo od progu interpolacji
  - W zakresie tym model w sposób (blisko) perfekcyjny interpoluje wszystkie obserwacje uczące.
  - Zasadniczą rzeczą, która się zmienia jest zachowanie modelu pomiędzy obserwacjami uczącymi.
- Osiągane jest nowe minimum funkcji błędu dla zbioru testowego.


Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., & Sutskever, I. (2021). Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment, 2021*(12), 124003.

Efektywna złożoność modelu

Niech:

- $\mathcal{T}$ – procedura ucząca
- $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ – zaetykietowany zbiór uczący
- $\mathcal{T}(S)$ – predykcje
- $\mathcal{D}$ – rozkład danych

Efektywna złożoność modelu (*Effective model complexity*) definiowana jest jako maksymalna liczba obserwacji *n*, dla których $\mathcal{T}$ osiąga błąd na zbiorze treningowym *bliski* 0.

Definition 1 (Effective Model Complexity) *The* Effective Model Complexity *(EMC) of a training procedure $\mathcal{T}$, with respect to distribution $\mathcal{D}$ and parameter $\epsilon > 0$, is defined as:*

$$\mathrm{EMC}_{\mathcal{D},\epsilon}(\mathcal{T}) := \max\{n \mid \mathbb{E}_{S\sim\mathcal{D}^n}[\mathrm{Error}_S(\mathcal{T}(S))] \leq \epsilon\}$$

where $\mathrm{Error}_S(M)$ is the mean error of model M on train samples S.

Hipoteza

Under-paremeterized regime. *If* $\mathrm{EMC}_{\mathcal{D},\epsilon}(\mathcal{T})$ *is sufficiently smaller than* n*, any perturbation of* $\mathcal{T}$ *that increases its effective complexity will decrease the test error.*

Over-parameterized regime. *If* $\mathrm{EMC}_{\mathcal{D},\epsilon}(\mathcal{T})$ *is sufficiently larger than* n*, any perturbation of* $\mathcal{T}$ *that increases its effective complexity will decrease the test error.*

Critically parameterized regime. *If* $\mathrm{EMC}_{\mathcal{D},\epsilon}(\mathcal{T}) \approx n$*, then a perturbation of* $\mathcal{T}$ *that increases its effective complexity might decrease* **or increase** *the test error.*

Hipoteza jest nieformalna:

- Autorzy nie mają pryncypialnego sposobu na wybór wartości parametru ϵ
  - Na potrzeby eksperymentów przyjęli $\epsilon = 0.1$
- Brak formalnej specyfikacji *sufficiently smaller* i *sufficiently larger*.
  - Autorzy przyznają, że nie umieją wyczerpująco wyjaśnić od czego zależy szerokość przedziału.

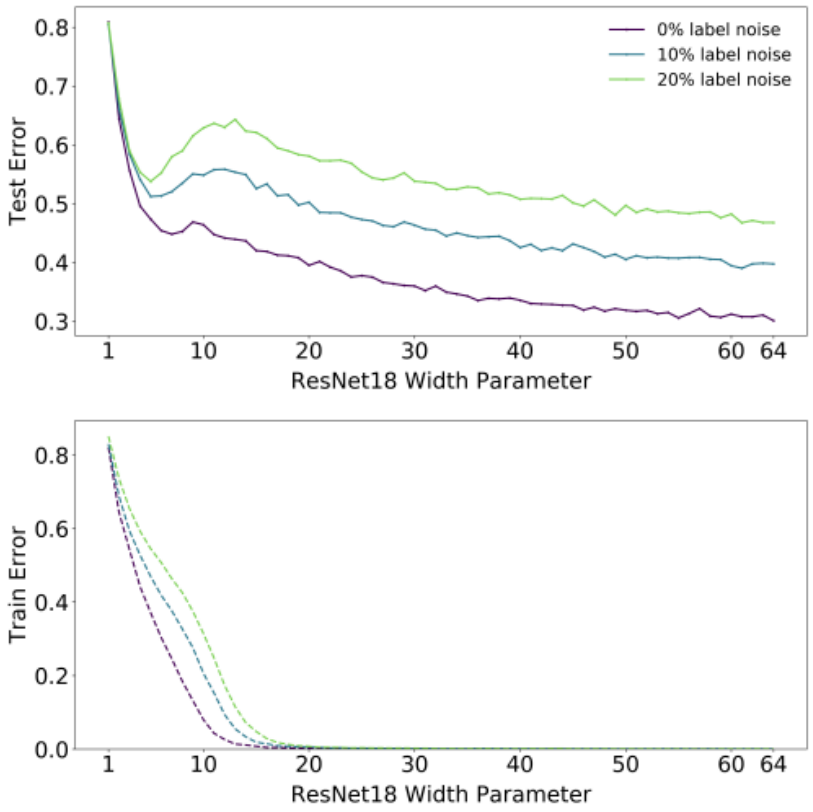
Różne aspekty złożoności modelu

- Złożoność modelu może być rozumiana na kilka sposobów:
  - Wielkość modelu (liczba jego parametrów)
  - Długość treningu
  - Rozmiar zbioru uczącego
- Autorzy badają ponadto wpływ zaszumienia etykiet na analizowane zjawisko.

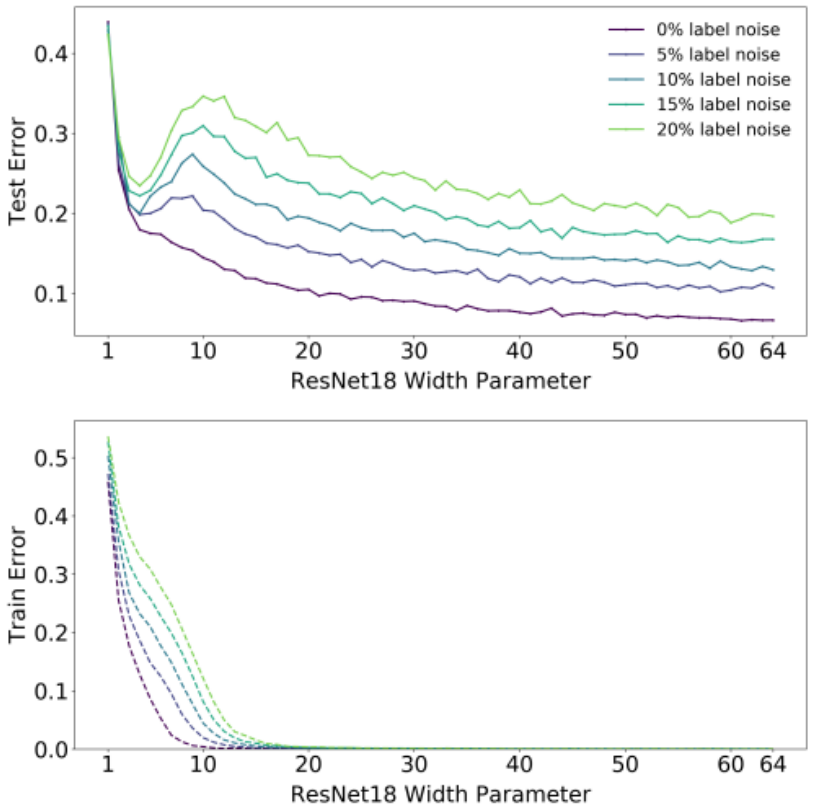
Konfiguracja eksperymentów

- Trzy rodziny architektur:
  - ResNet
  - Standardowe CNN
  - Transformery
- Każda z rodzin testowana była w wielu różnych parametryzacjach:
  - Dla ResNet i CNN analizowano różną szerokość (liczba filtrów) warstw konwolucyjnych.
  - Dla transformerów zmieniano wielkość warstwy zanurzonej (*embedding dimension*) oraz proporcjonalnie szerokość warstw gęstych.

Model-wise double descent – wyniki (1)



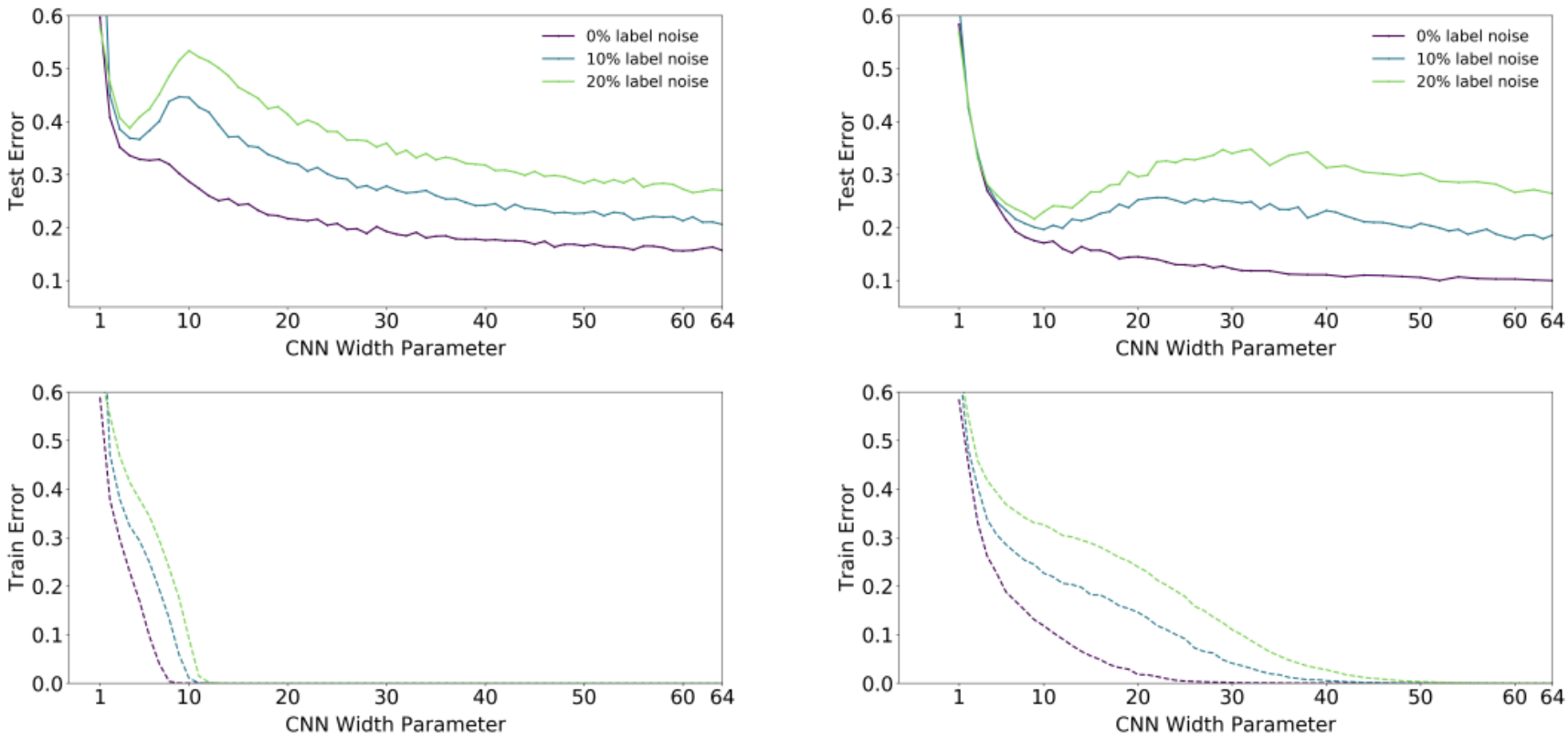
(a) **CIFAR-100.** There is a peak in test error even with no label noise.



(b) **CIFAR-10.** There is a "plateau" in test error around the interpolation point with no label noise, which develops into a peak for added label noise.

Figure 4: **Model-wise double descent for ResNet18s.** Trained on CIFAR-100 and CIFAR-10, with varying label noise. Optimized using Adam with LR 0.0001 for 4K epochs, and data-augmentation.

Model-wise double descent – wyniki (2)



(a) Without data augmentation.

(b) With data augmentation.

Figure 5: **Effect of Data Augmentation.** 5-layer CNNs on CIFAR10, with and without data-augmentation. Data-augmentation shifts the interpolation threshold to the right, shifting the test error peak accordingly. Optimized using SGD for 500K steps. See Figure 27 for larger models.

Model-wise double descent – wyniki (3)

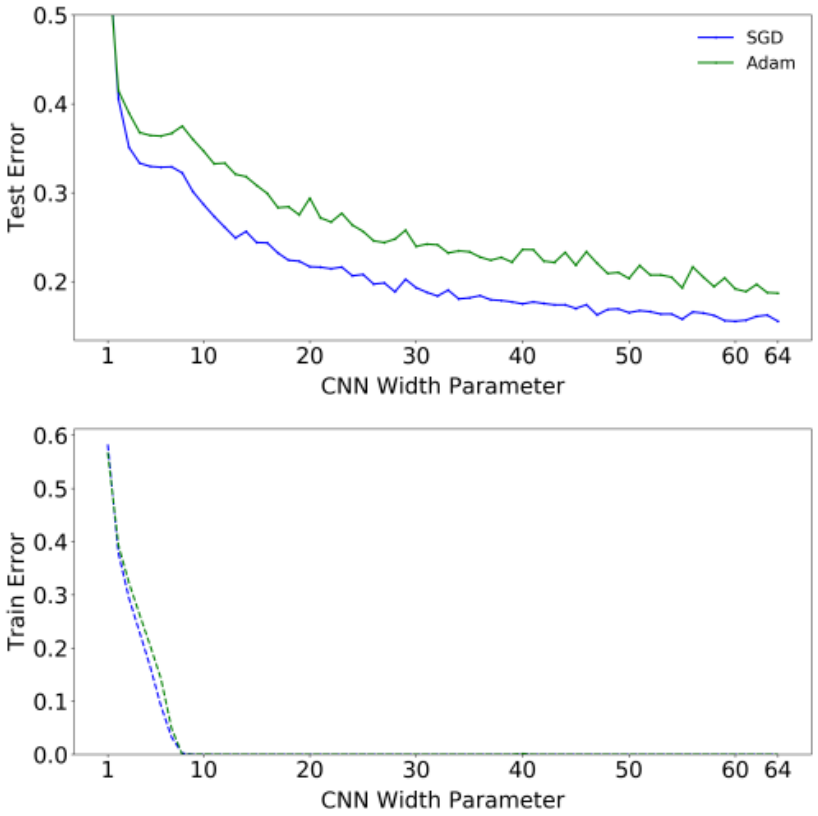


Figure 6: **SGD vs. Adam.** 5-Layer CNNs on CIFAR-10 with no label noise, and no data augmentation. Optimized using SGD for 500K gradient steps, and Adam for 4K epochs.

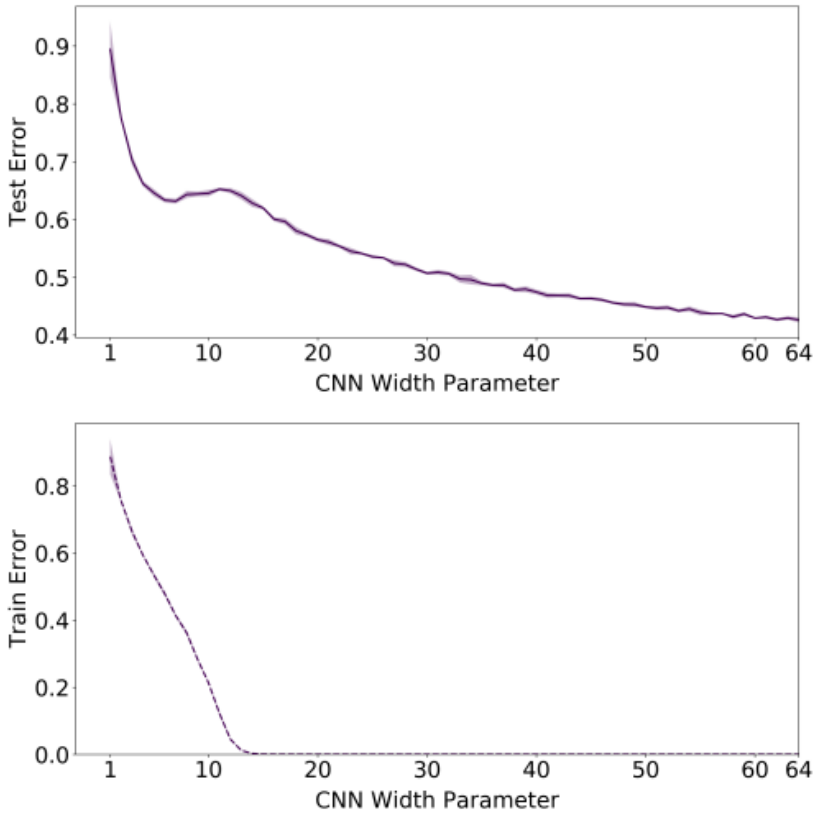


Figure 7: **Noiseless settings.** 5-layer CNNs on CIFAR-100 with no label noise; note the peak in test error. Trained with SGD and no data augmentation. See Figure 20 for the early-stopping behavior of these models.

Model-wise double descent – wyniki (4)

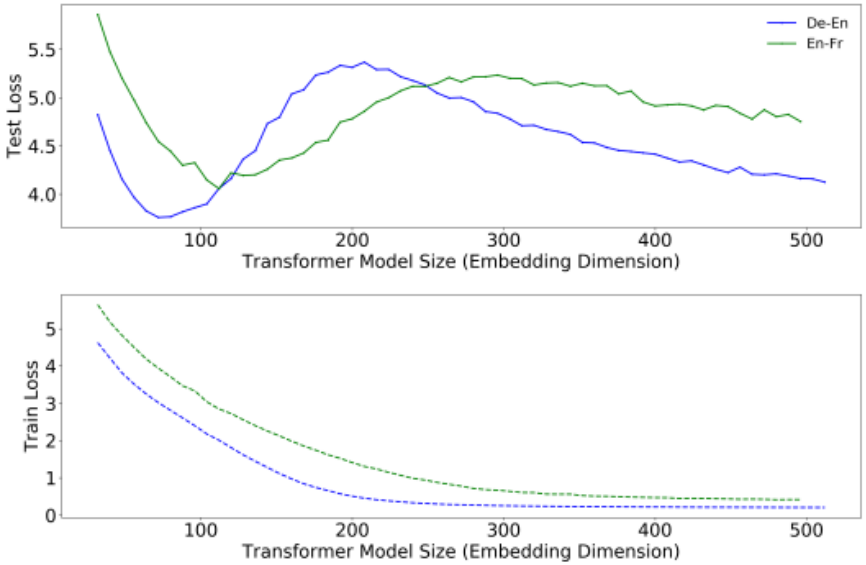


Figure 8: **Transformers on language translation tasks:** Multi-head-attention encoder-decoder Transformer model trained for 80k gradient steps with labeled smoothed cross-entropy loss on IWSLT‘14 German-to-English (160K sentences) and WMT‘14 English-to-French (subsampled to 200K sentences) dataset. Test loss is measured as per-token perplexity.

Model-wise double descent – intuicja (1)

- Istnieje tylko jeden model o złożoności progu interpolacji, który dokładnie dopasowuje dane treningowe.
- Model ten jest bardzo wrażliwy na szum w danych.
- Model (ze względu na swą złożoność) jest ledwo w stanie dopasować się do danych, dlatego nauczy się również szumu.
- Największy zysk z ensemblingu następuje właśnie w okolicach progu interpolacji.

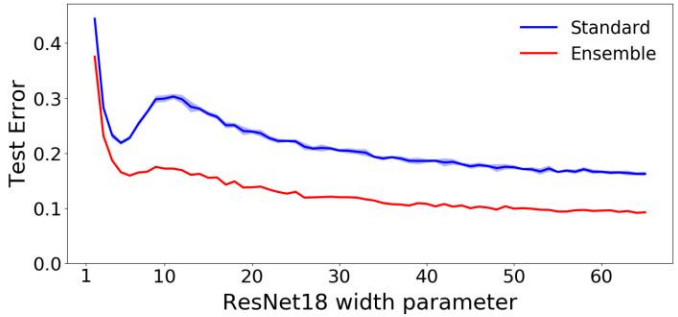


Figure 28: **Effect of Ensembling (ResNets, 15% label noise).** Test error of an ensemble of 5 models, compared to the base models. The ensembled classifier is determined by plurality vote over the 5 base models. Note that emsembling helps most around the critical regime. All models are ResNet18s trained on CIFAR-10 with 15% label noise, using Adam for 4K epochs (same setting as Figure 1). Test error is measured against the original (not noisy) test set, and each model in the ensemble is trained using a train set with independently-sampled 15% label noise.

Model-wise double descent – intuicja (2)

However for over-parameterized models, there are many interpolating models that fit the train set, and SGD is able to find one that "memorizes" (or "absorbs") the noise while still performing well on the distribution.

Model-wise double descent – wnioski (1)

- Zjawisko występuje dla różnych zbiorów danych, poziomów zaszumień, architektur, optymalizatorów, wielkości zbioru treningowego, procedury treningu (augmentacja danych, regularyzacja)
  - Stanowi to empiryczny dowód, że zjawisko nie jest przypadkiem.
- Skok błędu występuje w okolicach progu interpolacji
- Skok błędu na zbiorze testowym jest bardziej widoczny dla danych z zaszumionymi etykietami.
  - Praktyczna implikacja: dla czystych danych ciężej wychwycić skok.

Model-wise double descent – wnioski (2)

- Wszystkie operacje, które zwiększają próg interpolacji (zaszumianie etykiet, augmentacja danych, zwiększenie liczby obserwacji treningowych), przesuwają skok błędu na zbiorze testowym w kierunku większych modeli.
- Pełne zrozumienie zjawiska dla wariantu *model-wise* pozostaje otwartym pytaniem.

Epoch-wise double descent – wyniki (1)

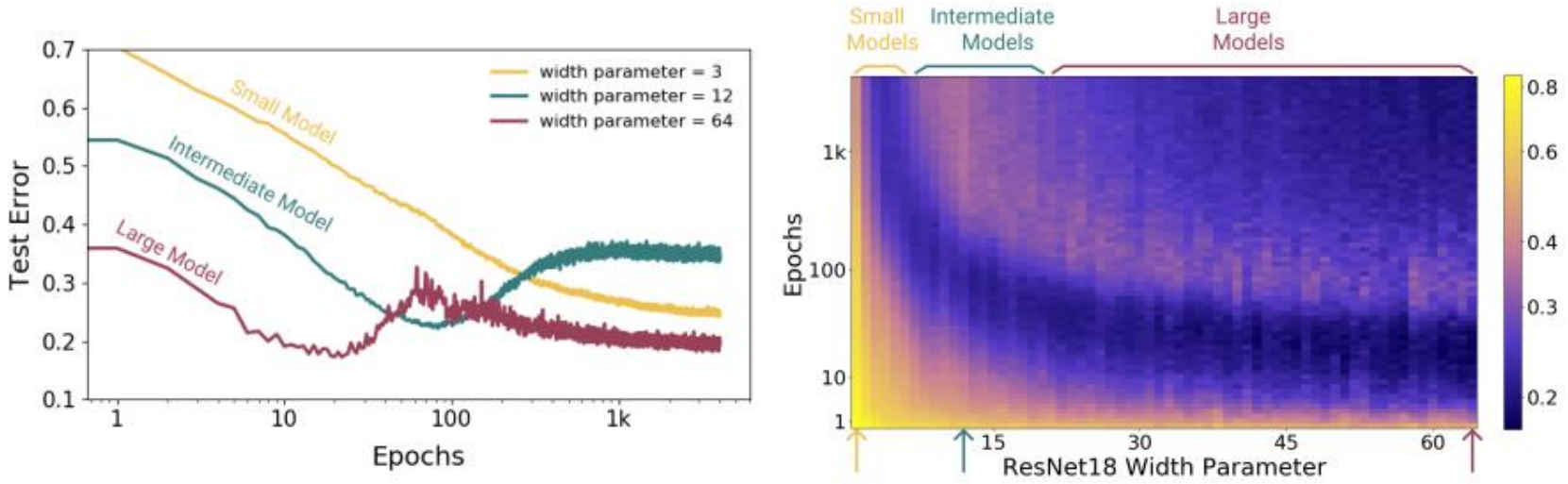
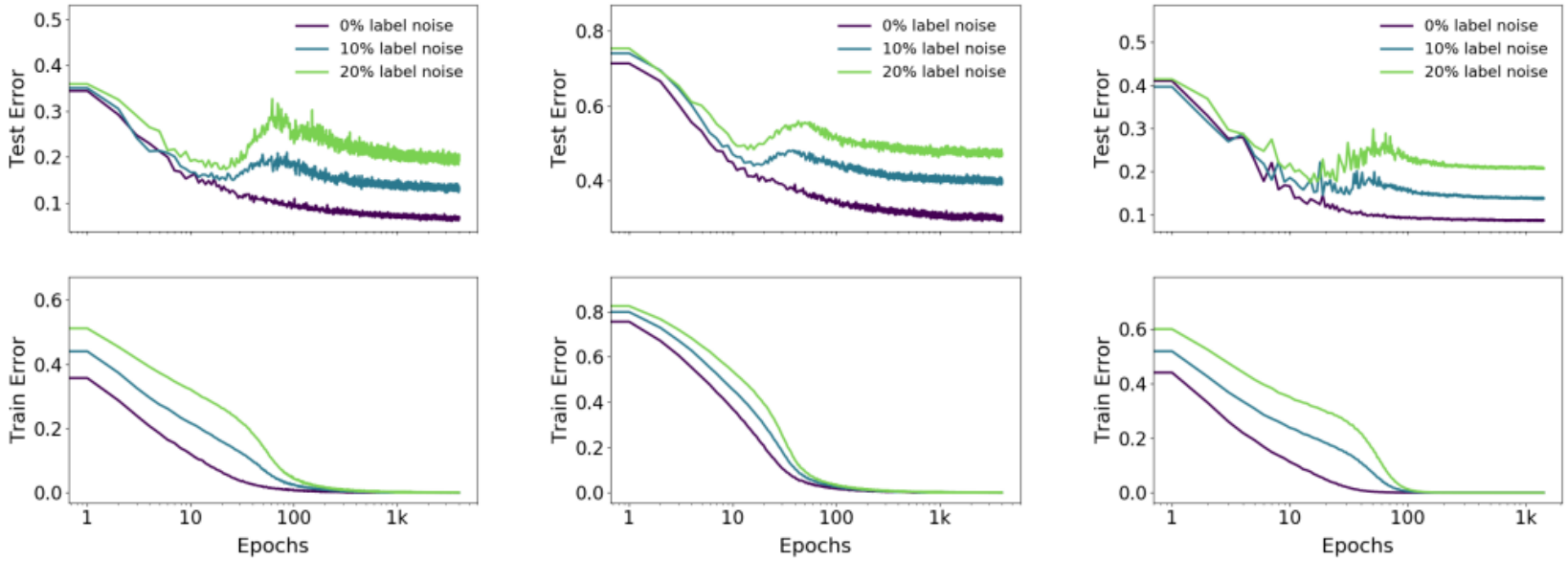


Figure 9: **Left:** Training dynamics for models in three regimes. Models are ResNet18s on CIFAR10 with 20% label noise, trained using Adam with learning rate 0.0001, and data augmentation. **Right:** Test error over (Model size × Epochs). Three slices of this plot are shown on the left.

Epoch-wise double descent – wyniki (2)



(a) ResNet18 on CIFAR10. (b) ResNet18 on CIFAR100. (c) 5-layer CNN on CIFAR 10.

Figure 10: **Epoch-wise double descent** for ResNet18 and CNN (width=128). ResNets trained using Adam with learning rate 0.0001, and CNNs trained with SGD with inverse-squareroot learning rate.

Epoch-wise double descent – wnioski

- Większość wniosków dla spojrzenia *model-wise* jest również poprawnych dla *epoch-wise*.
- W szczególności:
  - Zjawisko występuje dla różnych konfiguracji zbiorów i architektur, z różnym poziomem zaszumienia etykiet.
  - Skok błędu na zbiorze testowym jest szczególnie widoczny dla danych zaszumionych.

Sample non-monotonicity – wyniki (1)

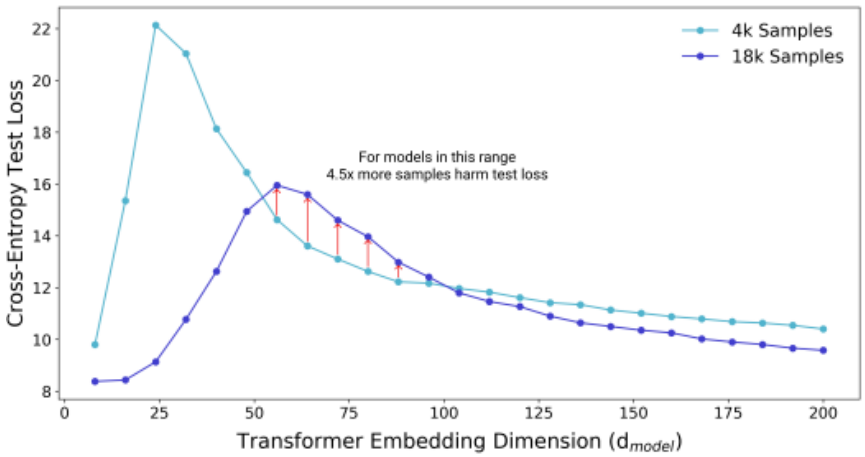
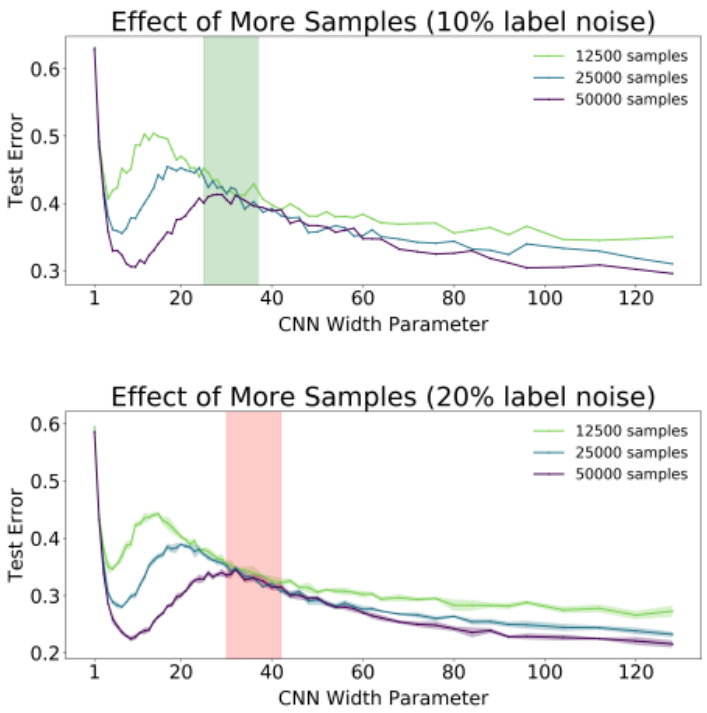
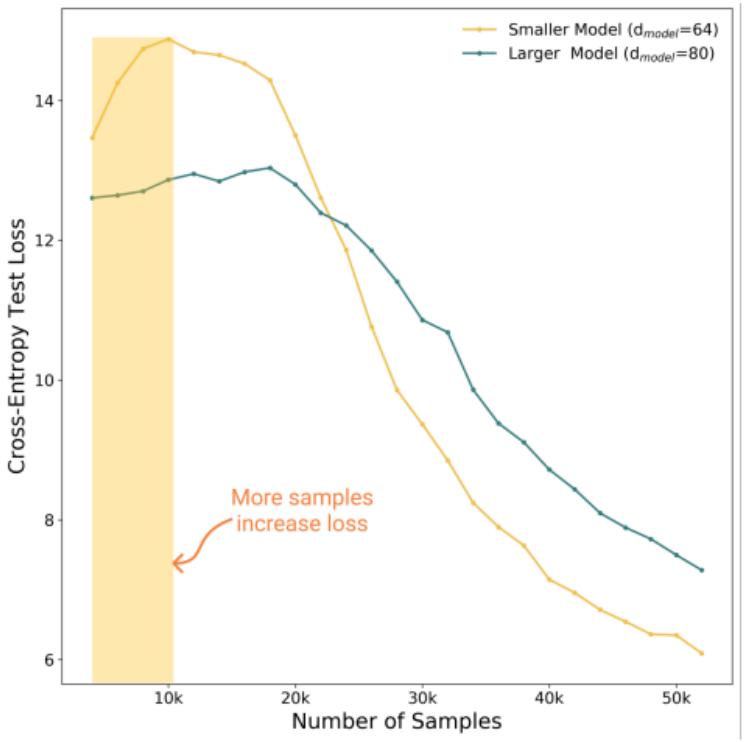


Figure 3: Test loss (per-token perplexity) as a function of Transformer model size (embedding dimension d_{model}) on language translation (IWSLT'14 German-to-English). The curve for 18k samples is generally lower than the one for 4k samples, but also shifted to the right, since fitting 18k samples requires a larger model. Thus, for some models, the performance for 18k samples is *worse* than for 4k samples.

Sample non-monotonicity – wyniki (2)



(a) Model-wise double descent for 5-layer CNNs on CIFAR-10, for varying dataset sizes. **Top:** There is a range of model sizes (shaded green) where training on 2× more samples does not improve test error. **Bottom:** There is a range of model sizes (shaded red) where training on 4× more samples does not improve test error.



(b) **Sample-wise non-monotonicity.** Test loss (per-word perplexity) as a function of number of train samples, for two transformer models trained to completion on IWSLT'14. For both model sizes, there is a regime where more samples hurt performance. Compare to Figure 3, of model-wise double-descent in the identical setting.

Figure 11: Sample-wise non-monotonicity.

Sample non-monotonicity – wyniki (3)

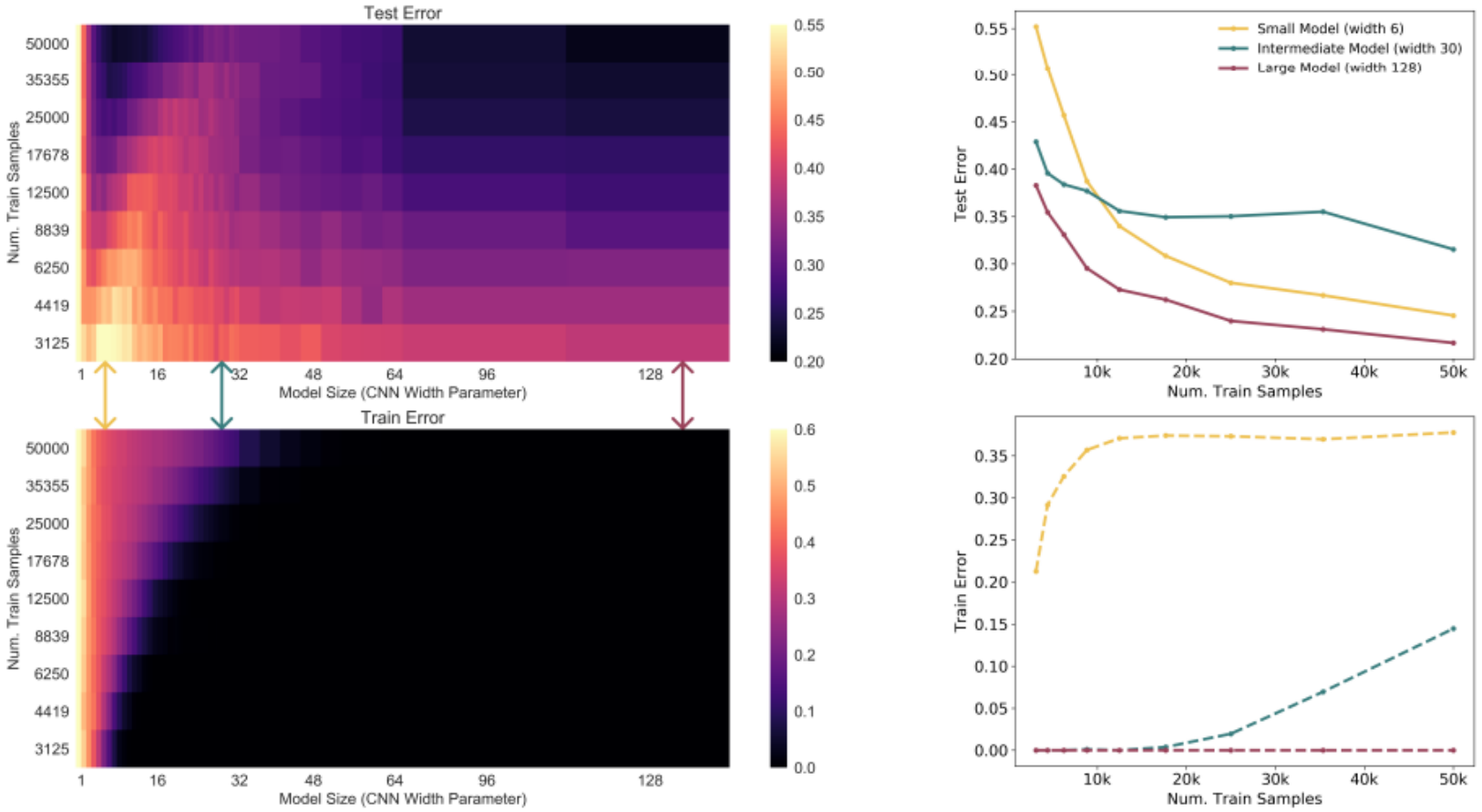


Figure 12: **Left:** Test Error as a function of model size and number of train samples, for 5-layer CNNs on CIFAR-10 + 20% noise. Note the ridge of high test error again lies along the interpolation threshold. **Right:** Three slices of the left plot, showing the effect of more data for models of different sizes. Note that, when training to completion, more data helps for small and large models, but does not help for near-critically-parameterized models (green).

Podsumowanie

- Atypowe zachowanie, gdy EMC jest porównywalne z liczbą obserwacji treningowych.
  - Odpowiada to sytuacji, gdzie błąd na zbiorze treningowym jest bliski 0.
- Eksperymentalnie pokazano, że zjawisko to zachodzi dla różnych zbiorów danych, architektur i procedur uczenia (regularyzaja/augmentacja)
- Praktyczna uwaga: gdy model osiąga błąd na zbiorze treningowym bliski 0, niewielkie zmiany w nim lub procedurze uczenia mogą prowadzić do nieprzewidzianych zachowań, np. istotne pogorszenie jakości na zbiorze testowym.
- Mechanizm wczesnego zatrzymania (*early stopping*) z reguły nie dopuszcza do wystąpienia zjawiska
  - EMC jest mniejsze niż liczba próbek uczących.
- Czy zatem powinniśmy starać się zawsze osiągnąć błąd treningowy bliski 0 i zobaczyć co dzieje się dalej?

Źródła

1. https://mlu-explain.github.io/double-descent/


Q&A